Chapter 5
Bayesian Statistics

We already briefly introduced the Bayesian approach to statistical inference in Chap. 3 and in this chapter we will dive deeper into this methodology. Whole books have been written about the different techniques in Bayesian statistics, which is a huge and very well developed field that we could not hope to cover in a single chapter. For this reason we will focus on major principles and will provide a list of references for deeper exploration.

A comprehensive history of Bayes reasoning can be found in McGrayne (2011), in which one can find many interesting ways it has been applied in the last centuries. The field was named after the reverend Thomas Bayes (1701–1761), who solved the "inverse probability" problem. His solution is nowadays known as Bayes' rule. This work was published posthumously in his name by Richard Price in 1763. It was also developed independently by Laplace (1774) but then was somewhat forgotten. When the field of Bayesian reasoning emerged in the 1950s, it was given his name as the process was largely based on manipulating conditional distributions.

5.1 Bayesian Principles

The fundamental principle in Bayesian reasoning is that uncertainty about a parameter θ is expressed in terms of a probability. This implies that the parameter θ is considered a random variable, with some prior probability distribution $f_\theta(\vartheta)$, where $\vartheta \in \Theta$ and Θ is the parameter space containing all possible (or reasonable) values for the parameter θ. If we assume a distributional model for the random variables $Y_1, \dots Y_n$, this gives us, after observing $y_1, \dots, y_n$, the posterior distribution

$$f_\theta(\vartheta | y_1, \dots y_n) = \frac{f(y_1, \dots, y_n; \vartheta) f_\theta(\vartheta)}{f(y_1, \dots y_n)}.$$

The denominator

$$f(y) = f(y_1, \ldots, y_n) = \int_{\Theta} f(y_1, \ldots, y_n; \vartheta) f_\theta(\vartheta) d\vartheta$$

can be seen from multiple perspectives. It can be seen as the marginal density for the observations after integrating out the uncertainty about the parameter. It can also be seen as the probability of the data based on our prior understanding of likely values for θ. A further way to view $f(y)$, as it does not depend on θ, is as the normalisation constant of the posterior distribution. The posterior distribution itself is proportional to the product of the prior and the likelihood function, that is

$$f_\theta(\vartheta | y_1, \ldots, y_n) \propto L(\vartheta; y_1, \ldots, y_n) f_\theta(\vartheta),$$

where $\propto$ stands for "is proportional to". We will see that in general the marginal distribution $f(y)$ is difficult to derive analytically and numerical methods are absolutely necessary to make Bayesian statistics work in practice. An exception can be found in **conjugate priors**.

Definition 5.1 Assume that $Y \sim f(y; \theta)$ follows a given probability model coming from a family of distributions $\mathcal{F}_y = \{f(y; \theta), \theta \in \Theta\}$ and that the prior distribution $f_\theta(\vartheta)$ comes from a family of distributions $\mathcal{F}_\theta = \{f_\theta(\vartheta; \gamma), \gamma \in \Gamma\}$. $\mathcal{F}_\theta$ is **conjugate** to $\mathcal{F}_y$, if for the posterior distribution it holds that $f_\theta(\vartheta | y) \in \mathcal{F}_\theta$.

This definition requires some explanation. First, the reader should note that the distribution of the parameter θ also depends on some further parameters, defined as γ above. These parameters are called **hyperparameters** and need to be specified in advance. Moreover, we can see that the families $\mathcal{F}_y$ and $\mathcal{F}_\theta$ are linked, meaning that the conjugate prior depends on our model of Y. To better understand this definition and the role of conjugate priors in Bayesian reasoning, let us look at a few examples.

Example 11 Assume $Y_i \sim N(\mu, \sigma^2)$ *i.i.d.* and that for the moment σ^2 is known. For the mean μ we postulate the prior

$$\mu \sim N(\gamma, \tau^2),$$

where both hyperparameters γ and τ^2 are assumed to be known. Let $y = (y_1, \ldots, y_n)$ be the data. We find that the posterior of μ is proportional to the prior distribution multiplied by the likelihood, i.e.

$$\begin{aligned} f_\mu(\mu | y) &\propto f_y(y; \mu, \sigma^2) f_\mu(\mu; \gamma, \tau^2) \\ &\propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2} \frac{(y_i - \mu)^2}{\sigma^2}\right) \exp\left(-\frac{1}{2} \frac{(\mu - \gamma)^2}{\tau^2}\right) \end{aligned}$$

$$\propto \exp\left(-\frac{1}{2}\frac{\sum_{i=1}^{n} y_i^2 - 2\sum_{i=1}^{n} y_i\mu + n\mu^2}{\sigma^2} - \frac{1}{2}\frac{\mu^2 - 2\mu\gamma + \gamma^2}{\tau^2}\right)$$

$$\propto \exp\left(-\frac{1}{2}\left\{\mu^2\left[\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right] - 2\mu\left[\frac{n\bar{y}}{\sigma^2} + \frac{\gamma}{\tau^2}\right]\right\}\right). \tag{5.1.1}$$

Note that the normal distribution is a member of the exponential family of distributions, as shown in Sect. 2.1.5. In fact, we can derive from (5.1.1) that $f(\mu|y)$ has the form of a normal distribution with parameters

$$\mu|y \sim N\left(\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\left(\frac{n\bar{y}}{\sigma^2} + \frac{\gamma}{\tau^2}\right), \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right).$$

Note that terms or factors that do not depend on μ can be ignored on the right-hand side of (5.1.1). With this, we can conclude that, when estimating the mean of a normal distribution, the normal prior is also a conjugate prior. Furthermore, the posterior mean is a weighted average of the sample mean $\bar{y}$ and the prior mean γ.

If we look at the parameters of the posterior, we can see an asymptotic property of Bayesian statistics. With appropriate prior parameters, i.e. γ bounded and τ^2 bounded away from zero, and increasing sample size n the posterior distribution of μ given y converges to a normal $N(\bar{y}, \sigma^2/n)$ distribution. That is,

$$\mu|y \qquad \overrightarrow{n \to \infty} \qquad N\left(\bar{y}, \frac{\sigma^2}{n}\right).$$

Hence, with increasing sample size we find that the posterior distribution of μ becomes centred around $\bar{y}$ with variance $\frac{\sigma^2}{n}$. Note that this is exactly the "opposite" probability formulation as for the Maximum Likelihood estimate $\hat{\mu} = \bar{y}$. For the ML estimate we found asymptotic normality of $\bar{y}$ around θ, while for Bayes inference we find posterior normality of θ around $\bar{y}$. ▷

Example 12 Let us again look at the normal distribution, but this time focus on the variance σ^2 and, for simplicity, fix the mean μ to a known value. We assume that σ^2 comes from an inverse gamma distribution with hyperparameters α and β, such that

$$f_{\sigma^2}(\sigma^2; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}\exp\big(-\frac{\beta}{\sigma^2}\big),$$

where $\alpha > 0, \beta > 0$ and $\Gamma(.)$ is the Gamma function (which is the continuous extension of the factorial operation, such that $\Gamma(k) = (k-1)!$ for $k \in \mathbb{N}$). We denote the prior model with $\sigma^2 \sim IG(\alpha, \beta)$. The posterior is then obtained with

$$\begin{aligned} f_{\sigma^2}(\sigma^2|y) &\propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2}\frac{\sum_{i=1}^n (y_i-\mu)^2}{\sigma^2}\right)(\sigma^2)^{-\alpha-1}\exp\left(-\frac{\beta}{\sigma^2}\right) \\ &\propto (\sigma^2)^{-(\alpha+n/2)-1}\exp\left(-\frac{(\beta+\frac{1}{2}\sum_{i=1}^n (y_i-\mu)^2)}{\sigma^2}\right), \end{aligned}$$

such that $\sigma^2|y \sim IG(\alpha + n/2, \beta + \frac{1}{2}\sum_{i=1}^n (y_i-\mu)^2)$. ▹

We have already seen a further example of a conjugate prior in Sect. 3.4.2, with the beta distribution for the parameter π of a Binomial distribution. The last example relates to the Poisson distribution.

Example 13 Let $Y \sim \text{Poisson}(\lambda)$ and assume for λ a Gamma prior of the form

$$f_\lambda(\lambda|\alpha,\beta) = \frac{\lambda^{\alpha-1}\ \exp(-\lambda\beta)}{\beta^{-\alpha}\Gamma(\alpha)}.$$

Then the posterior of λ given y is again a Gamma distribution with parameters $(\alpha + y)$ and $(\beta + 1)$ because

$$\begin{aligned} f_\lambda(\lambda|y;\alpha,\beta) &\propto \lambda^y \exp(-\lambda)\lambda^{\alpha-1}\exp(-\lambda\beta) \\ &\propto \lambda^{\alpha+y-1}\exp(-(\beta+1)\lambda). \end{aligned}$$

We continue with this example and visualise the effect of the sample size. Let us take a flat prior with $\alpha = 1$ and $\beta = 0$ and calculate the posterior as given above. Note that the resulting prior is not a proper distribution, but this turns out not to be a problem, since we are interested in the posterior. Note that for Y_i sampled *i.i.d.* we get

$$f_\lambda(\lambda|y_1,\ldots,y_n;\alpha,\beta) \propto \lambda^{\alpha+n\bar{y}-1}\exp(-(\beta+n)\lambda).$$

We plot this density in Fig. 5.1 for three different sample sizes, namely $n = 10$, $n = 100$ and $n = 1000$, where $\bar{y} = 2$ in all three cases. The increasing amount of information becomes obvious, as with increasing sample size the posterior becomes more centred around $\bar{y}$.

▹

We can conclude that with conjugate priors Bayesian reasoning becomes rather simple. Unfortunately, conjugate priors are often not a reasonable way to model and quantify our prior knowledge. Moreover, beyond the classical cases discussed above it can be difficult, or even impossible, to find a conjugate prior for a more

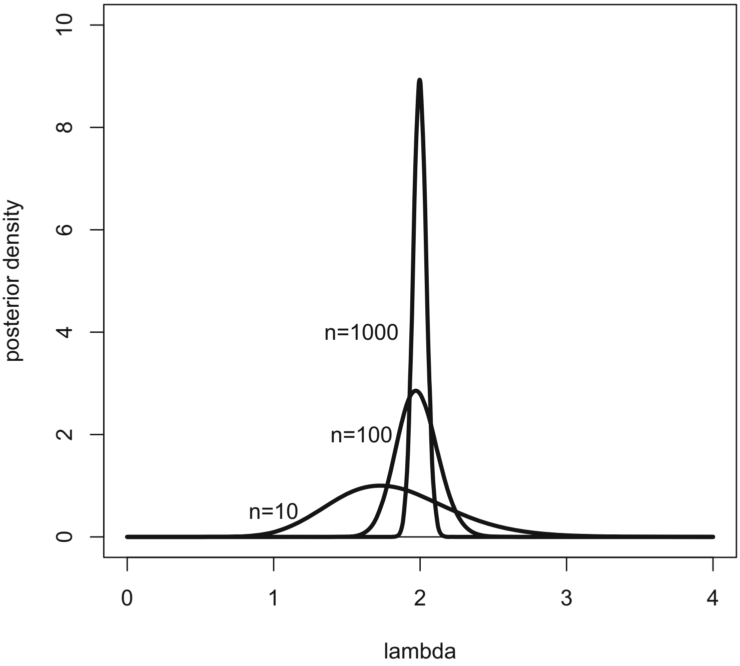

Fig. 5.1 Posterior distribution for Poisson distributed variables with Gamma prior and different sample sizes

complicated model, particularly if θ is multidimensional. Therefore, it is essential that we discuss different prior structures and numerical routines for calculating posterior distributions under more general conditions. We will begin with the choice of the prior.

Example 14 We extend our analysis from Chapter 3 on evaluating opinion polls for election preferences (see Bauer et al. 2018). In a multi-party system, survey participants are asked for their preferences between various political parties. The answers of the participants can be modelled by a multinomial distribution, which is a generalisation of the binomial distribution. Having the choice between K parties, let $Y_1, \ldots, Y_K$ be the number of participants that decided in favour of party $k = 1, \ldots, K$. The total number of participants is denoted by n. Then, the probability distribution is given by

$$P(Y_1 = y_1, \ldots, Y_K = y_K) = \frac{n!}{y_1! \ldots y_K!} \prod_{k=1}^{K} \theta_k^{y_k},$$

with $\sum_{k=1}^{K} y_k = n$. The parameters θ_k are the probabilities for choosing each party k. Assuming a simple random sample, θ_k is the (unknown) proportion of voters for each party k in the population. This is exactly the parameter of interest and we can approach inference on the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ from a Bayesian perspective.

To this end, we first need to specify a prior distribution for $\boldsymbol{\theta}$. The conjugate prior for the multinomial distribution is the Dirichlet distribution, which is given by

$$f(\theta_1, \ldots, \theta_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_i - 1} \text{ for all } 0 \leq \theta_k \leq 1 \text{ and } \sum_{k=1}^{K} \theta_k = 1. \tag{5.1.2}$$

Here, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ is the parameter vector and $B(\boldsymbol{\alpha})$ the normalisation constant, such that (5.1.2) is a well defined density. A common choice for the prior distribution is $\boldsymbol{\alpha} = (\frac{1}{2}, \ldots, \frac{1}{2})$. The calculation of the posterior distribution is straightforward in this case, because

$$\begin{aligned} f_{post}(\theta|y) &= P(Y = y|\theta) \cdot f_\theta(\theta) \\ &\propto \prod_{k=1}^{K} \theta_j^{y_k} \cdot \prod_{k=1}^{K} \theta_k^{-\frac{1}{2}} = \prod_{k=1}^{K} \theta_k^{y_k - 1/2}. \end{aligned}$$

Therefore, the posterior is again a Dirichlet distribution with parameter $\boldsymbol{\alpha} = (y_1 + \frac{1}{2}, \ldots, y_K + \frac{1}{2})$:

$$(\theta|y) \sim Dirichlet(y_1 + \frac{1}{2}, \ldots, y_K + \frac{1}{2}). \tag{5.1.3}$$

The resulting expectation is given by

$$E(\boldsymbol{\theta}|y) = \left(\frac{y_1 + \frac{1}{2}}{\sum_{k=1}^{K} (y_k + \frac{1}{2})}, \ldots, \frac{y_K + \frac{1}{2}}{\sum_{k=1}^{K} (y_k + \frac{1}{2})} \right).$$

Furthermore, we can calculate probabilities of relevant events. For example, assume that we are interested whether Party 1 has the most votes or whether Party 2 has more than 5% of the votes, which is the limit for entering parliament in Germany. For concrete calculations of these probabilities, a Monte Carlo approach can be applied, where a sample of sufficient size (e.g. 10^4) is drawn from the posterior distribution. Given the sample we can now calculate

$$P(\theta_1 = max(\theta_1, \ldots, \theta_k)|y) = \frac{\#(\text{samples with } \theta_1 = max(\theta_1, \ldots, \theta_K))}{10^4},$$

$$P(\theta_2 \geq 0.05|y) = \frac{\#(\text{samples with } \theta_2 \geq 0.05)}{10^4}.$$

With this method it is possible to calculate posterior probabilities for all relevant events concerning the multivariate distribution of $(\theta_1, \ldots, \theta_K)$, which is one of the key advantages of the Bayesian approach, see Bauer et al. (2018).

5.2 Selecting a Prior Distribution

The necessity of choosing a prior distribution is the Achilles heel of Bayesian statistics. It can be problematic if the result of a statistical evaluation depends too much on the choice of the prior distribution. To perform an objective analysis based on the data, the prior information inserted into the model needs to be as small as possible. Therefore, one tries to apply a **non-informative** prior. This will be explored in the coming sections.

5.2.1 Jeffrey's Prior

Let us assume $f_\theta(.)$ as the prior distribution of our parameter θ, of which we have no prior knowledge. Because we lack any prior knowledge, we decide to set $f_\theta(.)$ to be constant. This is commonly called a **flat prior**. If we now transform the parameter θ to $\gamma = g(\theta)$, where $g(.)$ is an invertible transformation function, then the prior for γ is given by

$$f_\gamma(\gamma) = f_\theta(g^{-1}(\gamma)) \left| \frac{\partial g^{-1}(\gamma)}{\partial \gamma} \right| .$$

We assumed that no prior knowledge exists for θ, and hence that $f_\theta(.)$ is constant. However, we can clearly see that $f_\gamma(.)$ is not constant in expressing knowledge about the *transformed* parameter γ. Hence, if we assume no prior knowledge of θ, we implicitly have a non-uniform prior for γ. This may or may not be reasonable, depending upon the situation.

Example 15 Let us run a simple Binomial experiment with $Y \sim B(n, \pi)$. We assume a non-informative prior on π, which is given by $\pi \sim Beta(1, 1)$, such that $f_\pi(\pi) = 1$ for $\pi \in [0, 1]$. Hence, the prior is constant. Instead of π we now reparameterise the model with the log odds $\theta = \log(\pi/(1 - \pi))$. Then $\pi = g^{-1}(\theta) = \exp(\theta)/(1 + \exp(\theta))$ and the resulting prior for θ is given by

$$f_\theta(\theta) = f_\pi(g^{-1}(\theta)) \left| \frac{\partial g^{-1}(\theta)}{\partial \theta} \right| = \frac{\exp(\theta)}{(1 + \exp(\theta))^2}$$

which is clearly not constant. This is visualised in the top row of Fig. 5.2. ▷

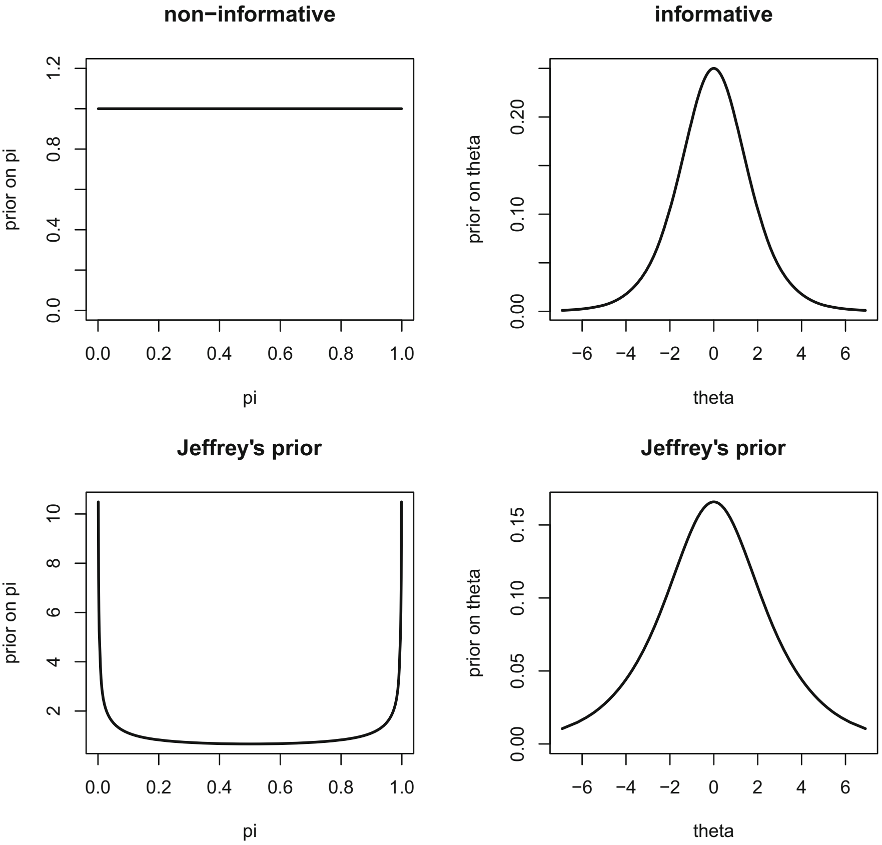

Fig. 5.2 Different priors for the binomial distribution. Top row: flat prior for π (left) and resulting informative prior for the log odds ratio θ. Bottom row: Jeffrey's prior for π (left) and θ (right)

Transformations of a parameter induce a structure on the prior of the transformed parameter. If we consider this unreasonable, we might want to formulate a prior distribution which remains unchanged, even under transformation. Such a transformation-invariant prior is called the Jeffery's prior. Transformation-invariance means that if we make use of Jeffrey's prior for one parameterisation, then a transformation of the parameter gives the Jeffrey's prior in this alternative parameterisation.

Definition 5.2 (Jeffrey's Prior) For Fisher-regular distributions $f(y;\theta)$ Jeffrey's prior is proportional to

$$f_\theta(\theta) \propto \sqrt{I_\theta(\theta)},$$

where $I_\theta(\theta)$ is the Fisher information.

To see that Jeffrey's prior is transformation invariant, we can make use of our previous result in Eq. (4.2.9) about the response of the Fisher matrix to parameter transformation. If $\gamma = g(\theta)$, then the Fisher information for γ is given by

$$I_\gamma(\gamma) = \frac{\partial g^{-1}(\gamma)}{\partial \gamma} I_\theta(g^{-1}(\gamma)) \frac{\partial g^{-1}(\gamma)}{\partial \gamma}.$$

If $f_\theta(\theta) \propto \sqrt{I_\theta(\theta)}$, then

$$f_\gamma(\gamma) \propto f_\theta(g^{-1}(\gamma)) \Big| \frac{\partial g^{-1}(\gamma)}{\partial \gamma} \Big| \propto \sqrt{\frac{\partial g^{-1}(\gamma)}{\partial \gamma} I_\theta(g^{-1}(\gamma)) \frac{\partial g^{-1}(\gamma)}{\partial \gamma}} = \sqrt{I_\gamma(\gamma)}$$

which is the Jeffrey's prior for the transformed parameter γ. For the binomial case, see the bottom row of Fig. 5.2.

An interesting property of the Jeffrey's prior is that it maximises the information gain from the data. This means that the distributional change from prior to posterior is as large as possible. In order to continue, we first need to clarify a few details. Firstly, we need to quantify the information difference between the prior and the posterior distribution. In Sect. 3.2.5, we introduced the Kullback–Leibler divergence to quantify the difference between two distributions (i.e. densities). This leads us to the following divergence measure between the prior and the posterior:

$$KL(f_\theta(.|y), f_\theta(.)) = \int_\Theta \log\left(\frac{f_\theta(\vartheta|y)}{f_\theta(\vartheta)}\right) f_\theta(\vartheta|y)\partial\vartheta.$$

The intention is now to choose the prior $f_\theta(.)$, such that the Kullback–Leibler divergence is maximised. Note that the Kullback–Leibler divergence depends on the data y, but the prior distribution does not. Hence, we also need to integrate out the data, which leaves us with the expected information (see Berger et al. 2009)

$$I(f_\theta) = \int KL(f_\theta(.|y), f_\theta(.)) f(y) dy,$$

where $f_y(y) = \int f(y; \vartheta) f_\theta(\vartheta) d\vartheta$. Clarke and Barron (1994) showed that, under appropriate regularly conditions, the information is maximised when $f_\theta(.)$ is chosen as the Jeffrey's prior. Thus, Jeffrey's prior has the additional property that the difference between the prior and posterior distribution is maximal and the maximum possible information is drawn from the data. Even though this supports the use of Jeffrey's prior, one should be aware that Jeffrey's prior still imposes prior knowledge on the parameter, as seen in Fig. 5.2. In fact, looking at Jeffrey's prior for the parameter π of a binomial distribution gives the impression that, before seeing any data, we favour values around 0 and 1 and a priori find values around 0.5 less plausible. If, however, we really have no prior knowledge about π, why should we choose this clearly informative prior?

For this reason, the use of Jeffrey's prior is contested and it is, in fact, not common to use non-informative priors, even if no prior information on the parameter is available. As Bernardo and Smith (1994) state: "Put bluntly: data cannot even speak entirely for themselves; every prior specification has *some* informative posterior (...) implications. (...) There is no objective prior that represents ignorance". We certainly do not want to engage in an ideological discussion of this issue here, but nevertheless emphasise that the choice of the prior may always be criticised and may have an influence on the outcome of Bayesian analyses. Nevertheless, Bayesian reasoning is generally useful and for increasing sample size, the influence of the prior vanishes. Therefore, a reasonable prior still needs to be chosen to make the Bayesian machinery run, which is why we will now discuss common alternatives.

5.2.2 Empirical Bayes

In this section, we discuss the derivation of priors from data with the "empirical Bayes" approach. Let us begin by noting that, as parameter θ is considered a random variable, we may integrate it out to get the distribution of the data, i.e.

$$f(y) = \int f(y; \vartheta) f_\theta(\vartheta) d\vartheta.$$

We can also assume that the prior distribution $f_\theta(.)$ depends on further hyperparameters, i.e. $f_\theta(\theta; \gamma)$ with γ as the (possibly multidimensional) hyperparameter. By extension, it should be clear that the marginal distribution $f(y)$ also depends on γ, because

$$f(y; \gamma) = \int f(y; \vartheta) f_\theta(\vartheta; \gamma) d\vartheta.$$

This is exactly the type of model that we dealt with in previous chapters and what we have learnt thus far would suggest choosing γ such that the likelihood $L(\gamma) = f(y; \gamma)$ is maximal. This approach is called **empirical Bayes**. In principle, however, empirical Bayes completely contradicts the Bayesian thinking. The prior is supposed to express the prior knowledge of the parameter *before* seeing data. As such, its hyperparameters should technically not be chosen *based on* the data. Counterintuitive as it may sound, empirical Bayes routines have become quite fashionable in some areas of statistics, such as penalised smoothing, and we will come back to the approach in Chap. 7. For now, however, we must stress that empirical Bayes technically contradicts the fundamentals of Bayesian reasoning.

Example 16 To clarify the relation between empirical Bayes and classical likelihood theory, let us look at a simple example. Let us take $Y \sim \text{Binomial}(n, \pi)$ and $\pi \sim \text{Beta}(\alpha, \beta)$. Then, integrating out π, we get

$$\begin{aligned} f(y|(\alpha, \beta)) &= \binom{n}{y} \frac{1}{\text{beta}(\alpha, \beta)} \int \pi^{y+\alpha-1}(1-\pi)^{n-y+\beta-1} d\pi \\ &= \binom{n}{y} \frac{\text{beta}(\alpha+y, \beta+n-y)}{\text{beta}(\alpha, \beta)}, \end{aligned}$$

where beta(,) is the beta function. The resulting distribution is known as Beta-Binomial model. Hence by integrating out π we obtain a "new" distribution which we can now use to estimate our hyperparameters α and β with Maximum Likelihood theory. ▷

5.2.3 *Hierarchical Prior*

A common strategy for the specification of the prior is to shift the problem of quantifying prior knowledge (or prior uncertainty) to a "higher level". This becomes an option if the form or the family of the prior distribution is fixed, but the prior distribution again depends on some hyperparameter(s) γ, which needs to be specified. Hence, the prior distribution takes the form $f_\theta(\vartheta; \gamma)$. The choice of the prior now simplifies to the question how to choose the hyperparameter γ. We can again take this from a Bayesian perspective and specify our knowledge of γ with a hyper-prior, i.e. γ itself has a distribution $f_\gamma(\gamma)$. In this case, we are interested in the posterior distribution of γ, which in turn will lead to a posterior distribution on θ, the parameter of interest. A model following this pattern takes the form

$$y|\theta \sim f_y(y; \theta); \quad \theta|\gamma \sim f_\theta(\theta; \gamma); \quad \gamma \sim f_\gamma(\gamma).$$

Clearly, the hyper-prior $f_\gamma(\gamma)$ may again depend on unknown parameters (i.e. hyper-hyperparameters), which would again need to be specified. That is, we have shifted the problem of quantifying our prior uncertainty to the "next level" and it may even seem that specifying our prior is now even more complicated than before! In fact, this is not necessarily the case. The intuition behind putting an extra layer of distributions on top of the hyperparameter is that it may be easier to select a distribution for the hyperparameters instead of fixing them to a specific value. This approach is commonly known as **hierarchical Bayes** and has proven to be quite powerful and applicable to a wide field of problems. We will see, however, that it will no longer be possible to calculate the posterior analytically.

Example 17 Assume $Y \sim B(n, \pi)$ and let $\pi \sim \text{Beta}(\alpha, \beta)$ be the conjugate prior. Instead of specifying the hyperparameters α, and β, we assume that

$$(\alpha, \beta) \sim f_{\alpha,\beta}((\alpha, \beta)|\varsigma),$$

where ς is a, possibly multidimensional, hyper-hyperparameter specifying the prior distribution of the Beta distribution parameters. Robert and Casella (2010), for instance, propose the hyper-prior

$$f_{\alpha,\beta}((\alpha, \beta); \varsigma) \propto \text{beta}(\alpha, \beta)^{\varsigma_0} \varsigma_1^{\alpha} \varsigma_2^{\beta}, \tag{5.2.1}$$

where $\varsigma = (\varsigma_0, \varsigma_1, \varsigma_2)$ are the hyper-hyperparameters. Clearly, the hyper-prior distribution does not have a common form, but is conjugate with respect to the Beta distribution. We are interested in the parameter π and therefore intend to find the posterior distribution $f_\pi(\pi|y)$. Note that,

$$f(\pi, (\alpha, \beta)|y) \propto f(y; \pi) f_\pi(\pi; (\alpha, \beta)) f_{(\alpha,\beta)}((\alpha, \beta); \varsigma),$$

such that

$$f(\pi|y) \propto f(y; \pi) \underbrace{\int f_\pi(\pi; (\alpha, \beta)) f_{(\alpha,\beta)}((\alpha, \beta); \varsigma) d\alpha d\beta,}_{:= f(\pi|\varsigma)}$$

which requires numerical integration. We can see that this extra layer in principle just defines a more complex prior distribution $f(\pi; \varsigma)$, which now depends on hyper-hyperparameters ς instead of hyperparameters α and β. Figure 5.3 (top row) shows the hyper-prior for α and β on the left, with $\varsigma_0 = 0$, $\varsigma_1 = \varsigma_2 = 0.9$ and the resulting prior for π on the right. This prior for π is clearly more complex than one coming from a simple Beta distribution. We get a similar shape if we replace the rather artificial prior (5.2.1) with a more pragmatic prior, assuming that $\log(\alpha)$ and $\log(\beta)$ are independently normally distributed, i.e. that α and β are log-normal. This is shown in Fig. 5.3 (bottom row) where we set the mean value of $\log(\alpha)$ and $\log(\beta)$ to 1. ▷

We can conclude that with hierarchical priors we get a more flexible parameter distribution, at the cost of extra computation. This naturally leads us to the next section, where we discuss numerical approaches for calculating the posterior distribution for arbitrarily chosen priors.

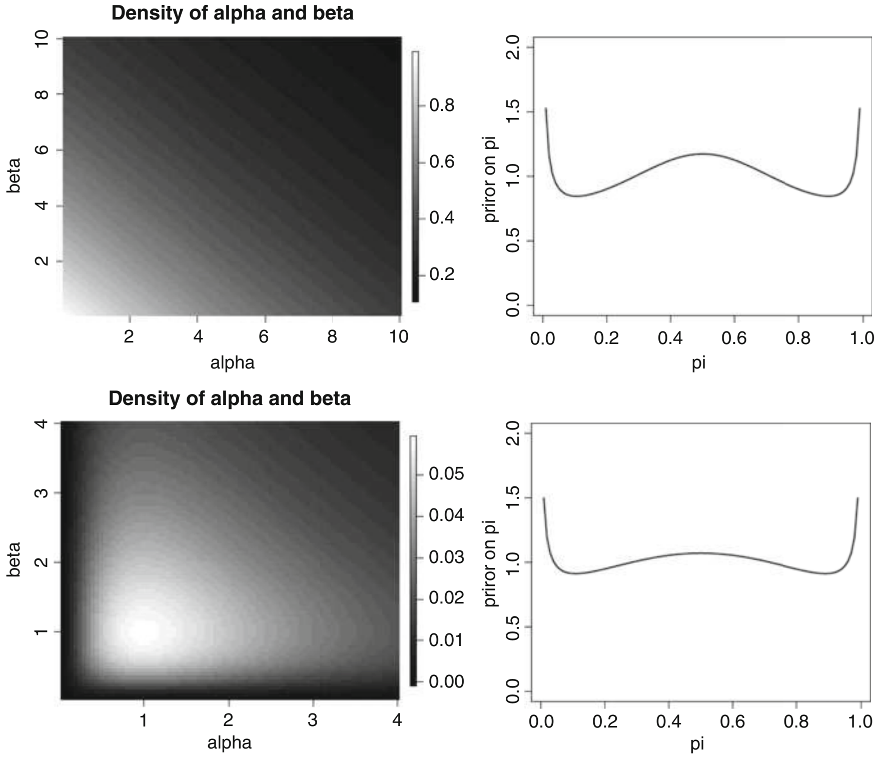

Fig. 5.3 Prior distribution for α and β of a beta distribution (left column) and resulting prior structure for π (right column) by integrating out α and β. Top row is for conjugate priors, bottom row is for independent normal priors for $\log(\alpha)$ and $\log(\beta)$

5.3 Integration Methods for the Posterior

5.3.1 Numerical Integration

The distribution that we are ultimately interested in is the posterior distribution of the parameter θ given data y, that is

$$f_\theta(\theta|y) = \frac{f(y;\theta) f_\theta(\theta)}{\int f(y;\vartheta) f_\theta(\vartheta) d\vartheta}. \tag{5.3.1}$$

In this setting we ignore (for the moment) that the prior $f_\theta(.)$ might depend on hyperparameters which themselves might have their own distributions. Instead, we assume that $f_\theta(.)$ is known and can be evaluated, that is, for every value of θ we can (easily) calculate $f_\theta(\theta)$. In this case, the unknown quantity in (5.3.1) is the denominator, which, for an arbitrary prior distribution $f_\theta(.)$, requires numerical integration. This could simply be carried out by standard numerical integration algorithms, such as rectangular, trapezoid or Simpson approximation. For example, with trapezoid approximation we use a grid $a = \theta_0 < \theta_1 < \ldots < \theta_k = b$. When Θ is an interval, i.e. $\Theta = [a, b]$, a and b are the boundary values of the parameter space Θ. If Θ is unbounded, a and b are chosen such that the integral components beyond a and b are of ignorable size, i.e. $\int_{-\infty}^{a} f(y; \vartheta) f_\theta(\vartheta) d\vartheta \approx 0$ and $\int_b^\infty f(y; \vartheta) f_\theta(\vartheta) d\vartheta \approx 0$. Trapezoid approximation of the denominator in (5.3.1) gives us

$$\int_\Theta f(y; \vartheta) f_\theta(\vartheta) d\vartheta \approx \sum_{k=1}^{K} \frac{f(y; \theta_k) f_\theta(\theta_k) + f(y; \theta_{k-1}) f_\theta(\theta_{k-1})}{2} (\theta_k - \theta_{k-1}).$$

Other routines for numerical integration, like Simpson approximation or rectangular approximation, are described in Stoer and Bulirsch (2002). While numerical integration is perfectly reasonable for one-dimensional integrals, it becomes problematic when applied to higher-dimensional problems. Even though we have thus far made the simplifying assumption that θ is univariate, we are in trouble when faced with the common task of approximating the integral of multivariate (often high dimensional) parameter vectors θ. There are, however, a number of alternative techniques for the approximation of this integral.

5.3.2 Laplace Approximation

Note that we typically draw n independent samples from $f(y|\theta)$. Hence, the data at hand are given by $y_1, \ldots, y_n$ and the posterior distribution takes the form

$$f(\theta|y_1, \ldots, y_n) = \frac{\prod_{i=1}^n f(y_i|\theta) f_\theta(\theta)}{\int \prod_{i=1}^n f(y_i|\theta) f_\theta(\theta) d\theta}.$$

Looking again at the denominator, we can rewrite the integral component as

$$\int \exp\left\{\sum_{i=1}^{n} \log f(y_i|\theta) + \log f_\theta(\theta)\right\} d\theta = \int \exp\left\{l_{(n)}(\theta; y_1, \ldots, y_n) + \log f_\theta(\theta)\right\} d\theta, \tag{5.3.2}$$

where $l_{(n)}(.)$ is the log-likelihood as defined in Chap. 4. Note that $l_{(n)}(.)$ is increasing in n, as discussed in Sect. 4.2. Hence, with increasing sample size n, the first component in the exp(.) in (5.3.2) becomes dominant. The idea is now to approximate the inner part of the exp(.) term with a second order Taylor approximation. We therefore twice differentiate the inner component of the exp(.) and get

$$\begin{aligned}
s_{P,(n)}(\theta) &:= \frac{\partial l_{(n)}(\theta; y_1, \dots, y_n)}{\partial \theta} + \frac{\partial \log f_\theta(\theta)}{\partial \theta} \\
&= s_{(n)}(\theta; y_1, \dots, y_n) + \frac{\partial \log f_\theta(\theta)}{\partial \theta}, \\
J_{P,(n)}(\theta) &:= -\frac{\partial^2 l_{(n)}(\theta; y_1, \dots, y_n)}{\partial \theta \partial \theta} - \frac{\partial^2 \log f_\theta(\theta)}{\partial \theta \partial \theta},
\end{aligned} \tag{5.3.3}$$

where the subscript n, similar to that of the score and Fisher information in Chap. 4, makes clear the influence of the sample size. The additional index P indicates that the prior distribution is also being considered in the calculations. Analogously, we define with $l_{P,(n)}(\theta)$ the component in the exp(.) in (5.3.2). Now we let $\hat{\theta}_P$ be the posterior mode estimate such that $s_{P,(n)}(\hat{\theta}_P) = 0$. With second order Taylor approximation we get

$$\int \exp(l_{P,(n)}(\vartheta))d\vartheta \approx \int \exp\left(l_{P,(n)}(\hat{\theta}_P) - \frac{1}{2} J_{P,(n)}(\hat{\theta}_P)(\vartheta - \hat{\theta}_P)^2 \right) d\vartheta.$$

A formal proof shows that with increasing sample size n the approximation error vanishes, which makes use of similar arguments as those applied to proving asymptotic normality of the ML estimate in Sect. 4.2. We refer the curious reader to Severini (2000) for a detailed discussion. Note that the integral now simplifies to

$$\exp(l_{P,(n)}(\hat{\theta}_P)) \int \exp\left(-\frac{1}{2} J_{P,(n)}(\hat{\theta}_P)(\vartheta - \hat{\theta}_P)^2 \right) d\vartheta.$$

The function in the integral mirrors the form of a normal distribution, such that the integral can in fact be calculated analytically and yields the inverse of the normalisation constant of a normal distribution. That is, one obtains the result

$$\int f(y; \vartheta) f_\theta(\vartheta) d\vartheta \approx f(y; \hat{\theta}_P) f_\theta(\hat{\theta}_P) \sqrt{2\pi} (J_{P,(n)}(\hat{\theta}_P))^{-1/2}. \tag{5.3.4}$$

While this approximation can be quite poor for small samples, it proves to be reasonably reliable if n is large.

Example 18 Let us once again take the Beta(α, β) prior for the parameter π of a Binomial distribution $B(n, \pi)$. Let $\alpha = \beta = 1$, which gives a constant prior and $\hat{\theta}_P = \bar{y} = y/n$. The second order derivative is $J_{P,(n)}(\hat{\theta}_{Pm}) = n/\{\bar{y}(1 - \bar{y})\}$ and, consequently, the Laplace approximation of the posterior is given by

$$f_\pi(\pi | y) = \frac{\pi^y (1 - \pi)^{n-y}}{\bar{y}^y (1 - \bar{y})^{n-y}} \frac{1}{\sqrt{2\pi} \sqrt{\bar{y}(1 - \bar{y})/n}}.$$

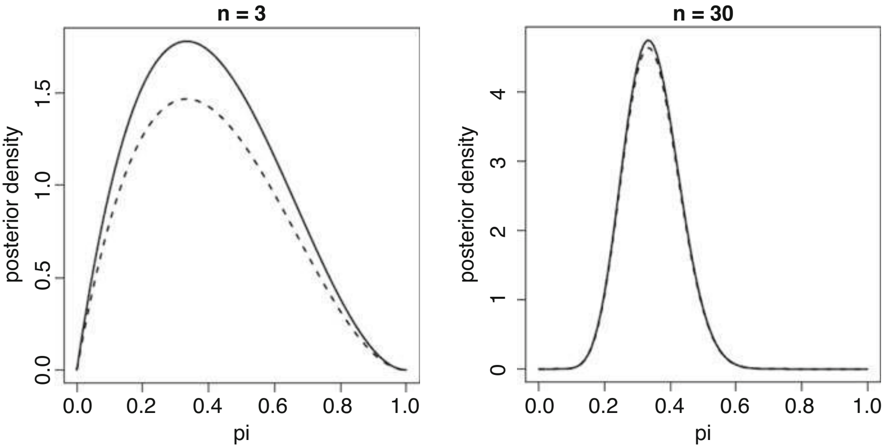

Fig. 5.4 Laplace approximation (dashed line) of posterior density (solid line) for Beta-Binomial model

In Fig. 5.4, we show the true posterior (solid line) and the Laplace approximated version (dashed line) for a sample size of $n = 3$ and $n = 30$, where $\bar{y} = \frac{1}{3}$ in both cases. The true posterior is marked by the solid line and its approximation by the dashed line. We can see that, even with a moderate sample size, the true posterior distribution is approximated nicely. ▷

The Laplace approximation is numerically simple and also works in higher dimensions. If θ is multidimensional, the approximation in (5.3.4) becomes

$$\int f(y;\vartheta) f_\theta(\vartheta) d\vartheta \approx f(y;\hat{\theta}_P) f_\theta(\hat{\theta}_P)(2\pi)^{p/2} \left| J_{P,\theta}(\hat{\theta}_P) \right|^{-1/2},$$

where exponent p is the dimension of the parameter, the vertical bars $|\cdot|$ denote the determinant and $J_{P,(n)}(\cdot)$ is the p-dimensional version of the Fisher matrix, including the prior given in (5.3.3). Laplace approximation is today occasionally used in a more advanced form in applied Bayesian analysis, see e.g. Rue et al. (2009) or www.r-inla.org. This circumvents the more numerically demanding routines which we will discuss later in this chapter.

5.3.3 *Monte Carlo Approximation*

An alternative approach is inspired by comprehending the integral in the denominator of the posterior (5.3.1) as an expected value

$$E_\theta(f(y;\theta)) = \int f(y;\vartheta) f_\theta(\vartheta) d\vartheta. \qquad (5.3.5)$$

Hence, we are interested in the expected value of $f(y;\theta)$ with θ randomly distributed according to the prior $f_\theta(\theta)$. The natural statistical approach to estimate this expectation would be to draw a random sample $\theta_1^*, \ldots, \theta_N^*$ from $f_\theta(\theta)$ and estimate the unknown expected value (5.3.5) with the corresponding arithmetic mean derived from the sample, i.e.

$$\hat{E}_\theta(f(y|\theta)) = \frac{1}{N}\sum_{j=1}^{N} f(y|\theta_j^*).$$

Clearly, as the number of samples $N \to \infty$ we get $\hat{E}_\theta(f(y;\theta)) \to E_\theta(f(y;\theta))$. Given that we can draw the sample computationally, the size of N is only limited by the numerical effort we are willing to make. Moreover, and this can be essential, we need to be able to actually draw a sample from the prior $f_\theta(\theta)$. If we are able to calculate the cumulative distribution function $F_\theta(\theta) = \int_{-\infty}^{\theta} f_\theta(\vartheta)d\vartheta$, that is, if we have $F_\theta(\vartheta)$ in numerical form, sampling is easy. In this case we just draw U_j^* from a uniform distribution on [0, 1] and get random draws from $f_\theta(.)$ with

$$\theta_j^* = F_\theta^{-1}(u_j^*).$$

Note that θ_j^* is then distributed according to $F_\theta(.)$, because

$$P(\theta_j^* \leq \vartheta) = P(F_\theta^{-1}(U_j^*) \leq \vartheta) = P(U_j \leq F_\theta(\vartheta)) = F_\theta(\vartheta).$$

If, on the other hand, $f_\theta(\theta)$ is not a standard distribution, e.g. in the case of our hierarchical prior with hyper-hyperparameters, there is often no ad hoc routine available, with which to draw a sample from $f_\theta(\theta)$. Unfortunately, in this case, the cumulative distribution function is not available in analytic form and requires numerical integration to be calculated.

An alternative sampling routine, which does not require that we sample from $f_\theta(.)$ directly, is **rejection sampling**. First we need a distribution $f_\theta^*(.)$ from which we can easily sample, that is for which the cumulative distribution function $F_\theta^*(.)$ can be easily calculated. This distribution must satisfy the envelope (or also called umbrella) property, which means that there must exist a constant $a < \infty$ such that

$$f_\theta(\vartheta) \leq a f_\theta^*(\vartheta) \quad \text{for all} \quad \vartheta \in \Theta.$$

We can now sample from $f_\theta(.)$ as follows:

1. Draw U^* from a uniform distribution on [0, 1].
2. Draw θ^* from $f_\theta^*(.)$, independent from U^*.
3. If $U^* \leq f_\theta(\theta^*)/(a f_\theta^*(\theta^*))$ then accept θ^*. Otherwise reject θ^* and jump back to Step 1.

Let us demonstrate that this strategy really gives random samples from $f_\theta(.)$, even though θ^* in Step 2 is drawn from $f^*_\theta(.)$. Let us therefore look at the distribution conditional upon acceptance in Step 3 in the sampling algorithm above, i.e.

$$P(\theta^* \leq \vartheta | \theta^* \text{ is accepted}) = \frac{P(\theta^* \leq \vartheta, \theta^* \text{ is accepted})}{P(\theta^* \text{ is accepted})}.$$

The numerator is given by

$$\begin{aligned} P\left(\theta^* \leq \vartheta, U^* \leq \frac{f_\theta(\theta^*)}{a f^*_\theta(\theta^*)}\right) &= \int_{-\infty}^{\vartheta} P\left(U^* \leq \frac{f_\theta(\theta^*)}{a f^*_\theta(\theta^*)}\right) f^*_\theta(\theta^*) d\theta^* \\ &= \int_{-\infty}^{\vartheta} \frac{f_\theta(\theta^*)}{a} d\theta^* = \frac{F_\theta(\vartheta)}{a}, \end{aligned}$$

where $F_\theta(.)$ is the cumulative distribution function corresponding to $f_\theta(.)$. The denominator results in the same form but now with ϑ set to ∞, i.e.

$$P(\theta^* \text{ is accepted}) = P(\theta^* \leq \infty, \theta^* \text{ is accepted}) = \frac{F_\theta(\infty)}{a} = \frac{1}{a}.$$

This directly proves that the rejection sampling algorithm produces a sample from $f_\theta(.)$ without actually drawing from $f_\theta(.)$. This sounds like a formidable idea, as we can sample directly from any distribution $f^*(.)$ and apply the routine above to obtain samples from our distribution of interest $f_\theta(\theta)$. However, the rejection step is a big obstacle in practice. If the proposal distribution $f^*_\theta(.)$ is chosen badly, then the proposal θ^* is only accepted with a (very) low probability, which could be very close to zero. In this case, we very rarely pass the acceptance step. This happens if a is large and $f^*_\theta(.)$ is far away from the target distribution $f_\theta(.)$. As $a \geq \max_\vartheta f_\theta(\vartheta)/f^*_\theta(\vartheta)$ the intention is to have $f^*_\theta(\vartheta)$ be as close as possible to $f_\theta(\vartheta)$ for $\vartheta \in \Theta$, i.e. we want to choose an umbrella distribution that shares the contours of our existing distribution, otherwise we will need a higher a to satisfy the umbrella property. More complex strategies, such as adaptive rejection sampling proposed by Gilks and Wild (1992), are able to elegantly address the problem of selecting an umbrella distribution. In this case, one adaptively constructs a piecewise umbrella distribution which covers $f_\theta(.)$.

An alternative strategy is **importance sampling**, see Kloek and Dijk (1978). In this case, we again draw a sample from $f^*_\theta(.)$ instead of $f_\theta(.)$ leading to independent draws $\theta^*_1, \ldots, \theta^*_N \sim f^*_\theta(.)$. We take this sample and estimate the mean value (5.3.5) with

$$\frac{1}{N} \sum_{i=1}^{N} \frac{f(y|\theta^*_i) f_\theta(\theta^*_i)}{f^*_\theta(\theta^*_i)}. \tag{5.3.6}$$

Note that this is a consistent estimate, because by taking the expectation with respect to sample $\theta_1^*, \dots, \theta_N^*$ we get

$$E^* \left(\frac{1}{N} \sum_{i=1}^{N} \frac{f(y|\theta_i^*) f_\theta(\theta_i^*)}{f_{\theta^*}(\theta_i^*)} \right) = \int \frac{f(y|\theta^*) f_\theta(\theta^*)}{f_{\theta^*}(\theta^*)} f_{\theta^*}(\theta^*) d\theta_i^*$$

$$= \int f(y|\theta^*) f_\theta(\theta^*) d\theta^* = f(y).$$

This sampling trick again looks surprisingly simple, as we just draw samples from $f_\theta^*(.)$, whatever the distribution is and plug the samples into (5.3.6) to get an estimate for (5.3.5). One should note, however, that the terms in the sum of (5.3.6) can have a tremendous variability, which occurs if the ratio $f_\theta(\theta^*)/f_\theta^*(\theta^*)$ is far away from 1. Hence, if the proposal density $f_\theta^*(\theta)$ is far away from the target density $f_\theta(\theta)$, we need a very (!) large sample N to get a reliable estimate. Consequently, we are faced with the same problem that occurred above in rejection sampling, that is, for it to work we need $f_\theta^*(.)$ to be close to $f_\theta(.)$.

5.4 Markov Chain Monte Carlo (MCMC)

Thus far we have been trying to approximate the normalisation constant of the posterior. While this approach looks like a plausible strategy, all approximation methods turn out to be problematic for multivariate θ. Alternatively, we can pursue a different strategy and try to simulate directly from the posterior. Assume, that we were able to draw θ_j^* directly from the posterior, i.e.

$$\theta_j^* \sim f_\theta(\theta|y).$$

In this case, we could get an $i.i.d.$ sample $\theta_j^*, j = 1, \dots, N$. This sample yields consistent estimates of the empirical distribution function and other parameters of the posterior. Hence, the idea of simulating θ from the posterior makes perfect sense, although the process is generally more complicated because we are not able to directly draw independent replicates θ_j^*. Instead we will draw a Markov chain $\theta_1^*, \theta_2^*, \dots$ where θ_j^* and θ_{j+1}^* will be correlated. The distribution of the constructed Markov chain will converge (under regularity conditions that are usually valid in this context) to a stationary distribution which is, in fact, the posterior distribution $f_\theta(.|y)$. In this case using the ergodicity of Markov chains, we have for any function $h(\theta^*)$

$$\frac{1}{N} \sum_{j=1}^{N} h(\theta_j^*) \to \int h(\vartheta) f_\theta(\vartheta|y) d\vartheta.$$

Given that we apply a **simulation based approach**, often referred to simply as Monte Carlo in statistics, we obtain a *M*arkov *C*hain *M*onte *C*arlo approach, which is abbreviated as **MCMC**.

We begin by introducing the **Metropolis-Hasting algorithm**, dating back to Metropolis et al. (1953) and Hastings (1970). The breakthrough in statistics took place when Gelfand and Smith (1990) applied the algorithm to Bayesian statistics. In the following, the reader should bear in mind that we still do *not* know the posterior probability

$$f_\theta(\theta|y) = \frac{f(y;\theta)f_\theta(\theta)}{f(y)},$$

as we do not know its denominator. However, we actually have quite a lot of information at our disposal. We know the shape of the distribution, because the posterior is proportional to the product of the likelihood and the prior, i.e.

$$f_\theta(\theta|y) \propto f(y;\theta)f_\theta(\theta).$$

In fact, if we compare the density of two values θ and $\tilde{\theta}$, say, we get the ratio

$$\frac{f_\theta(\theta|y)}{f_\theta(\tilde{\theta}|y)} = \frac{f(y;\theta)f_\theta(\theta)}{f(y)} \Big/ \frac{f(y)}{f(y;\tilde{\theta})f_\theta(\tilde{\theta})} = \frac{f(y;\theta)f_\theta(\theta)}{f(y;\tilde{\theta})f_\theta(\tilde{\theta})}.$$

Hence, the unknown quantity $f(y)$ cancels out. It is exactly this property that is exploited by the Metropolis-Hastings algorithm. Let $\theta^*_{(1)}, \theta^*_{(2)}, \ldots, \theta^*_{(t)}$ be a Markov chain constructed as follows:

1. Given a current value $\theta^*_{(t)}$ and a proposal distribution $q(\theta|\theta^*_{(t)})$ draw θ^*, i.e.

$$\theta^* \sim q(.|\theta^*_{(t)}).$$

2. Accept θ^* as new step in the Markov Chain with probability

$$\alpha(\theta^*_{(t)}, \theta^*) = \min\left\{1, \frac{f_\theta(\theta^*|y)q(\theta^*_{(t)}|\theta^*)}{f_\theta(\theta^*_{(t)}|y)q(\theta^*|\theta^*_{(t)})}\right\}.$$

If θ^* is not accepted set $\theta^*_{(t+1)} = \theta^*_{(t)}$, otherwise set $\theta^*_{(t+1)} = \theta^*$.

The proposal density $q(\theta^*|\theta^*_{(t)})$ needs to be adequately chosen as discussed below. If the proposal is symmetric, we have $q(\theta^*|\theta^*_{(t)}) = q(\theta^*_{(t)}|\theta^*)$ and the acceptance probability simplifies to

$$\alpha(\theta^*_{(t)}, \theta^*) = \min\left\{1, \frac{f_\theta(\theta^*|y)}{f_\theta(\theta^*_{(t)}|y)}\right\}.$$

In this case, the algorithm is also simply called the Metropolis algorithm. We see that if the posterior density for θ^* is higher than for $\theta^*_{(t)}$, we always accept the proposal θ^* as new step in the Markov Chain. Otherwise we accept the proposal θ^* with a probability less than one, dependent on the ratio of the posterior densities. Note that we only know the posterior $f_\theta(.|y)$ up to its normalisation constant, which however cancels out in the calculation of $\alpha(\theta^*_{(t)}, \theta^*)$, such that we can actually quite easily calculate the acceptance probability.

Property 5.1 The sequence of random numbers $\theta^*_{(j)}$, $j = 1, 2, \ldots$ drawn from the Markov chain with Metropolis (Hastings) as above has $f_\theta(.|y)$ as stationary distribution for $N \to \infty$.

Proof A formal proof on convergence requires deeper knowledge about Markov chains. Clearly we need some requirements to get the stationary distribution from the Markov chain. For the details we refer to classical literature like Grimmett and Stirzaker (2001). A technically rigorous discussion on requirements and properties of the Metropolis-Hastings algorithm can also be found in Robert and Casella (2004). However, for the purposes of explanation we remain on a heuristic level, instead motivating the central components of the proof. First of all, we need the proposal density to cover the full parameter space Θ, that is, all values in Θ are possible. For instance, if $\Theta = \mathbb{R}$, a normal distribution as proposal guarantees this condition. Let $K(\theta^*_{(t)}, \theta^*)$ be the transition probability (usually called the Kernel) that the Markov Chain proceeds from $\theta^*_{(t)}$ to θ^*. For the proposed Metropolis-Hastings algorithm this is given by

$$K(\theta^*_{(t)}, \theta^*) = q(\theta^*|\theta^*_{(t)})\alpha(\theta^*_{(t)}, \theta^*) + \delta_{\theta^*_{(t)}}(\theta^*)(1 - \alpha(\theta^*_{(t)})),$$

where $\delta_{\theta^*_{(t)}}(.)$ here stands for a Dirac measure taking value 1 if $\theta^* = \theta^*_{(t)}$ and 0 otherwise, and $\alpha(\theta^*_{(t)}) = \int \alpha(\theta^*_{(t)}, \theta)q(\theta|\theta^*_{(t)})d\theta$ is the overall acceptance probability. Note that we have a mixed update step, in that if θ^* is accepted it can take an arbitrary value given by a density, while if θ^* is not accepted it remains at $\theta^*_{(t)}$, giving a point mass at this parameter value. In other words, looking at the cumulative distribution function we get

$$P(\theta^* \leq \vartheta|\theta^*_{(t)}) = \int_{-\infty}^{\vartheta} q(\theta|\theta^*_{(t)})\alpha(\theta^*_{(t)}, \theta)d\theta + 1_{\{\vartheta \geq \theta^*_{(t)}\}}(1 - \alpha(\theta^*_{(t)})),$$

such that a jump of size $(1 - \alpha(\theta^*_{(t)}))$ occurs at the current state $\theta^*_{(t)}$. We show first that the Markov Chain is reversible, i.e.

$$K(\theta^*, \theta^*_{(t)})f_\theta(\theta^*_{(t)}|y) = K(\theta^*_{(t)}, \theta^*)f_\theta(\theta^*|y),$$

where $f_\theta(\theta|y)$ is the posterior distribution. Note that

$$\begin{aligned} q(\theta^*|\theta^*_{(t)})\alpha(\theta^*_{(t)},\theta^*)f_\theta(\theta^*_{(t)}|y) &= \min\left\{f_\theta(\theta^*|y)q(\theta^*_{(t)}|\theta^*),\ f_\theta(\theta^*_{(t)}|y)q(\theta^*|\theta^*_{(t)})\right\} \\ &= \min\left\{1,\ \frac{f_\theta(\theta^*_{(t)}|y)q(\theta^*|\theta^*_{(t)})}{f_\theta(\theta^*|y)q(\theta^*_{(t)}|\theta^*)}\right\} q(\theta^*_{(t)}|\theta^*)f(\theta^*|y) \\ &= q(\theta^*_{(t)}|\theta^*)\alpha(\theta^*,\theta^*_{(t)})f_\theta(\theta^*|y). \end{aligned}$$

This implies for the point mass probability at $\theta^* = \theta^*_{(t)}$ that

$$\delta_{\theta^*_{(t)}}(\theta^*)(1-\alpha(\theta^*_{(t)}))f_\theta(\theta^*_{(t)}|y) = \delta_{\theta^*}(\theta^*_{(t)})(1-\alpha(\theta^*))f_\theta(\theta^*|y).$$

Consequently, the Metropolis-Hastings Markov Chain is reversible. This in fact also proves that $f_\theta(\theta^*|y)$ is the stationary distribution of the Markov chain, which can be shown as follows. Assume that $\theta^*_{(t)}$ is drawn from the posterior distribution $f_\theta(\theta|y)$. Then $\theta^*_{(t+1)}$ is also drawn from $f_\theta(\theta|y)$, because

$$\begin{aligned} P(\theta^*_{(t+1)} \le \vartheta) &= \int_{-\infty}^{\vartheta}\left[\int_{-\infty}^{\infty} K(\theta^*_{(t)},\theta^*_{(t+1)})f_\theta(\theta^*_{(t)}|y)d\theta^*_{(t)}\right]d\theta^*_{(t+1)} \\ &= \int_{-\infty}^{\vartheta}\left[\int_{-\infty}^{\infty} K(\theta^*_{(t+1)},\theta^*_{(t)})f_\theta(\theta^*_{(t+1)}|y)d\theta^*_{(t+1)}\right]d\theta^*_{(t)} = P(\theta^*_{(t)} \le \vartheta). \end{aligned}$$

□

The Metropolis-Hastings algorithm allows us to sample from the posterior, even though we do not know its normalisation constant. To do this in practice, we start a Markov Chain and propose new values θ^* given the proposal distribution $q(\theta^*|\theta^*_{(t)})$. The proposal distribution should thereby fulfil two properties. First of all, it should allow us to sample the entire parameter space Θ. Secondly, it should provide a reasonable acceptance rate $\alpha(\theta^*_{(t)},\theta^*)$, implying that we should remain somewhat close to $\theta^*_{(t)}$ to keep the density of $f_\theta(\theta^*|y)$ and $f_\theta(\theta^*_{(t)}|y)$ of similar scale. These two goals clearly contradict each other and need to be balanced. Let us demonstrate the use of the Metropolis-Hastings algorithm in a small example. In Fig. 5.5 we plot a tri-modal posterior distribution $f_\theta(\theta|y)$ for a two-dimensional parameter and attempt to sample from it.

For the proposal distribution, we use the normal distribution

$$q(\theta^*|\theta^*_{(t)}) = N(\theta^*_{(t)},\sigma^2 I_2)$$

with the standard deviation set to $\sigma = 0.1$ for the first run and $\sigma = 1$ for the second run of the algorithm. We start the algorithm at $\theta^*_{(0)} = (1, 1)$ and plot the first 1000

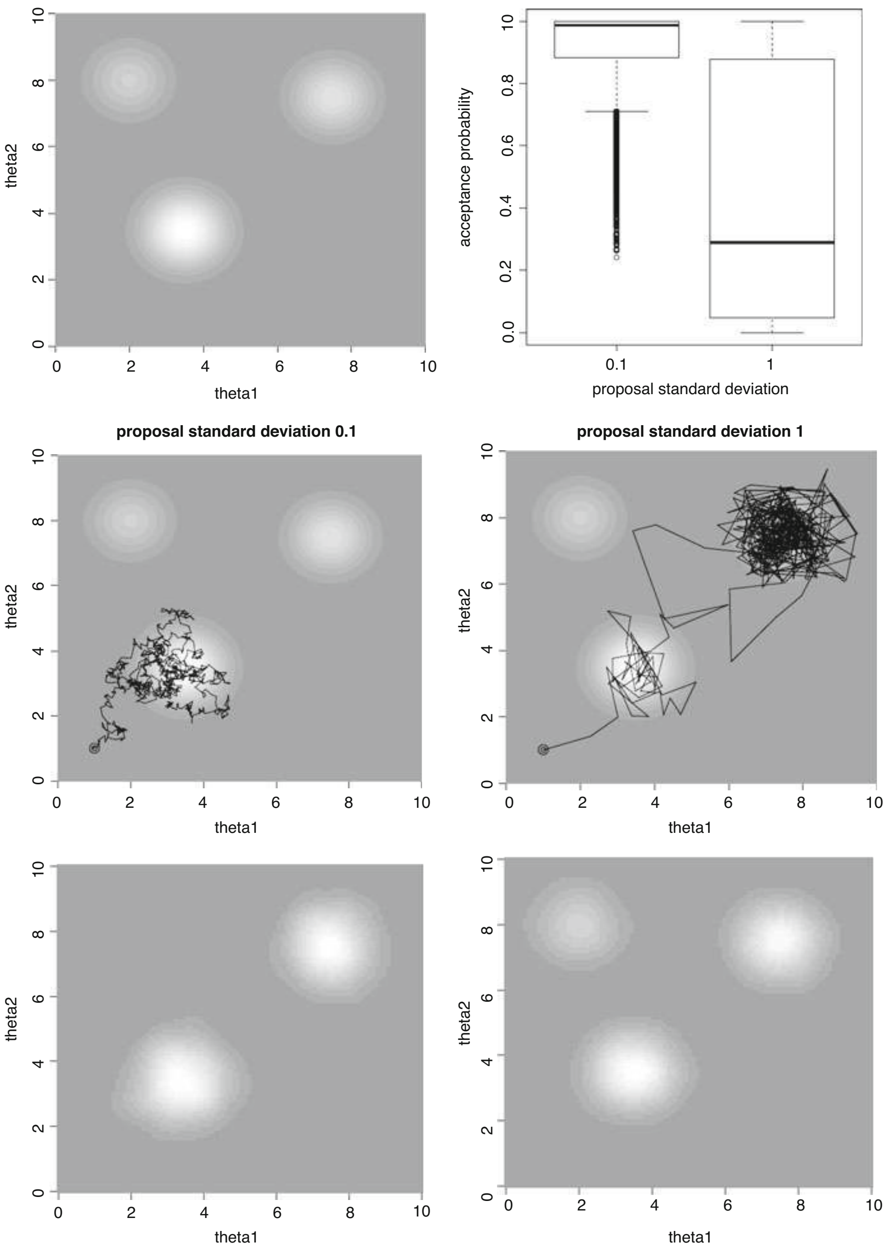

Fig. 5.5 Image plot of multi-mode posterior density (top left) and the resulting Markov Chain-Metropolis-Hastings estimates with proposal density $\sigma = 0.1$ (bottom left) and $\sigma = 1$ (bottom right). First 1000 steps of the algorithm (middle row) and acceptance probability (top right)

steps for both standard deviations. This is shown in the second row of Fig. 5.5, where the left plot corresponds to $\sigma = 0.1$ and the right plot to $\sigma = 1$.

We can see that smaller steps in the parameter space occur for a smaller proposal standard deviation, which resulted in the entire parameter space not being explored. In contrast, for the large standard deviation $\sigma = 1$, the steps are larger, which allows the algorithm to jump from the first high density area at mode (3.5, 3.5) to the second mode at (7.5, 7.5). The price for this exploration of the parameter space is that the acceptance rate for $\sigma = 1$ is much lower than that of $\sigma = 0.1$. This is shown in the top right of Fig. 5.5, where we plot the acceptance rate for the two proposals given 50,000 iterations. While for $\sigma = 0.1$ we have a median acceptance probability of close to 1, it is about 0.3 for $\sigma = 1$. While a high acceptance probability appears beneficial, it may lead to an incomplete exploration of the parameter space and posterior distribution. This is demonstrated in the bottom row of Fig. 5.5, where we plot the posterior density of 50,000 MCMC samples for the two proposal densities. While for $\sigma = 1$ the three modes are found and reproduced, for $\sigma = 0.1$ the top left mode is omitted, simply because the proposal density had too little range.

In practice, one also needs to pay attention to the performance of the Markov Chain with respect to convergence, autocorrelation and dependence on the starting value. The first aspect of managing these properties is to run the Markov Chain for a number of iterations until it reaches its stationary distribution. This is called the “burn-in” phase and visualised in Fig. 5.6 for a univariate parameter.

Figure 5.6 was calculated with a normally distributed posterior, with mean 10 and standard deviation 1. The proposal is also normal with $\theta^* \sim N(\theta^*_{(t)}; \sigma)$ with $\sigma = 0.1$. We start two Markov chains, one with starting value (black line), the second with 4 (grey line). Both series reach the mean value after about 500 steps. We take the first 1000 steps as burn-in, meaning that all simulated values in this burn-in phase will be ignored.

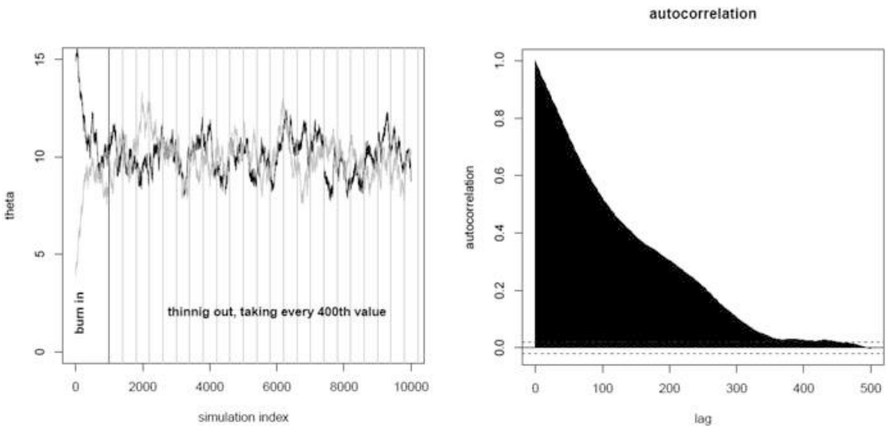

Fig. 5.6 Trajectory of Markov chain with two starting points and indicated burn-in phase (left plot). Autocorrelation of Markov chain series (right plot)

The next factor to be aware of is autocorrelation. Looking at the Markov chain after the burn-in phase, we see that successive steps are correlated over time, i.e. there is autocorrelation present. This is due to the construction of the series and a natural consequence of the Metropolis-Hastings approach and the Markov Chain model used. The autocorrelation is visualised on the right of Fig. 5.6, where we plot the empirical correlation between θ_t^* and θ_{t+lag}^*, which is strong but diminishes with larger lags. In fact, for lags of size 400 there is no empirical correlation observed. Given that $\theta_{(t)}^*$ are not uncorrelated, we may not take all simulated values $\theta_{(t)}^*$ but perhaps every 400th. The procedure is called "thinning out" and is visualised in Fig. 5.6 in the left-hand plot. Thinning out guarantees uncorrelated samples and thus an $i.i.d.$ sample from the posterior distribution.

The Metropolis-Hastings approach works well as long as the acceptance rate is reasonably high. However, if the parameter θ is high dimensional, it may occur that the acceptance probability $\alpha(\theta_{(t)}^*, \theta^*)$ is close to zero. This takes place because the density of each point in a multivariate distribution becomes smaller as dimensionality increases. Assume, for instance, that θ is p dimensional, i.e. $\theta = (\theta_1, \dots, \theta_p)$ and let, for simplicity, the posterior distribution be given by $f_\theta(\theta|y) = f_{\theta_1}(\theta_1|y) \cdot \dots \cdot f_{\theta_p}(\theta_p|y)$. With a symmetric proposal density

$$\alpha(\theta_{(t)}^*, \theta^*) = min\left\{1, \frac{f_{\theta_1}(\theta_1^*|y)}{f_{\theta_1}(\theta_{1(t)}^{(t)}|y)} \cdot \dots \cdot \frac{f_{\theta_p}(\theta_p^*|y)}{f_{\theta_p}(\theta_{p(t)}^*|y)}\right\}.$$

If for every p $f_{\theta_p}(\theta_p^*|y)/f_{\theta_p}(\theta_{p(t)}^*|y) = 0.8$, that is, for every parameter component separately we have 80% acceptance rate, the overall acceptance rate is 0.8^p, which for $p = 10$ is already only 10%. Clearly, the higher the dimension of θ, the lower the acceptance rate. This in turn makes the Metropolis-Hastings algorithm as proposed infeasible if p is large.

This can be circumvented by a process called **Gibbs sampling**, which again makes use of the conditional distribution. Note that for every $k \in \{1, \dots, p\}$ we have

$$f_{\theta_k|y,\theta_{-k}}(\theta_k|y, \theta_{-k}) \propto f(y|\theta) f_\theta(\theta), \tag{5.4.1}$$

where θ_{-k} denotes the parameter vector with (5.4.1) component k excluded, that is $\theta_{-k} = (\theta_1, \dots, \theta_{k-1}, \theta_{k+1}, \dots, \theta_p)$. This proportionality can now be used for sampling. The idea is that we sample only a single component of the entire parameter vector in each step, namely θ_k, and not the entire parameter vector.

To explain this more formally, we let $\theta = (\theta_1, \dots, \theta_p)$ be a p dimensional parameter. Assume further that $f_{\theta_k|y,\theta_{-k}}$ is known and can be sampled from. Let $\theta_{(t)}^*$ be the current value and, for $t = 0$, the starting value.

1. Draw $\theta_1^* \sim f_{\theta_1|y,\theta_{-1}}(\theta_1^*|y, \theta_{2(t)}, \ldots, \theta_{p(t)})$ and set $\theta^*_{1(t+1)} = \theta_1^*$
2. Draw $\theta_2^* \sim f_{\theta_2|y,\theta_{-2}}(\theta_2^*|y, \theta^*_{1(t+1)}, \theta^*_{3(t)}, \ldots, \theta^*_{p(t)})$ and set $\theta^*_{2(t+1)} = \theta_2^*$

$\vdots$

p. Draw $\theta_p^* \sim f_{\theta_p|y,\theta_{-p}}(\theta_p^*|y, \theta^*_{1(t+1)}, \ldots, \theta^*_{p-1(t+1)})$ and set $\theta^*_{p(t+1)} = \theta_p^*$
p+1. Jump back to 1

Property 5.2 The sequence of random variables $\theta^*_{(j)}$, $j = 1, 2, \ldots$ drawn from the above Gibbs sampling scheme converges to the posterior distribution $f_\theta(.|y)$ as a stationary distribution.

Proof Let us briefly sketch why the Gibbs algorithm produces random samples from the posterior distribution. For simplicity of notation we drop the observation y in the conditioning argument and assume that we want to prove that $\theta^*_{(t)}$ drawn using Gibbs sampling is in fact drawn from $f_{\theta|y}(\theta^*)$. Note that this also applies for individual components of the parameter vector, e.g. for $\theta^*_{1(t)}$. We want to show that $\theta^*_{1(t)}$ is drawn from $f_{\theta_1|y}(\theta_1)$. Also, for simplicity, let $p = 2$. Then the density after the first step in the Markov chain is given by

$$f_{\theta_{1(1)}|\theta_{1(0)}}(\theta^*_{1(1)}|\theta^*_{1(0)}) = \int f_{\theta_1|\theta_2}(\theta^*_{1(1)}|\theta^*_2) f_{\theta_2|\theta_1}(\theta^*_2|\theta_{1(0)}) d\theta^*_2,$$

where the two densities mirror the two proposal distributions in the first Gibbs sampling loop. Similarly we get

$$f_{\theta_{1(t)}|\theta_{1(0)}}(\theta^*_{1(t)}|\theta_{1(0)}) = \int f_{\theta_{1(t)}|\theta_{1(t-1)}}(\theta^*_{1(t)}|\theta_{1(t-1)}) f_{\theta_{1(t-1)}|\theta_{1(0)}}(\theta_{1(t-1)}|\theta_{1(0)}) d\theta_{1(t-1)}.$$

The density $f_{\theta_{1(t)}|\theta_{1(t-1)}}(.)$ gives the one-step transition. If $t \rightarrow \infty$ it follows under appropriate regularity conditions that $f_{\theta_{1(t)}|\theta_{1(0)}}(\theta^*_{1(t)}|\theta_{1(0)})$ converges to the stationary distribution $f_{\theta_1}(\theta_1^*)$, which is in fact the distribution we aim to sample from. □

Clearly, Gibbs sampling and the Metropolis-Hastings algorithm can be combined, which is necessary if the conditional density is only known up to a constant. Such hybrid approaches are rather common and we exemplify the procedure using the same example from above, but now every update is made univariately using a Gibbs sampling scheme in combination with Metropolis-Hastings. As a proposal density we take a univariate normal distribution, i.e. $\theta_k^* \sim N(\theta^*_{k(t)}, \sigma)$. The top row of Fig. 5.7 shows the first 1000 updates, when the proposal standard deviation is either set to $\sigma = 1$ (left-hand side) or $\sigma = 2$ (right-hand side).

It can be observed that the Markov Chain moves horizontally and vertically, mirroring the fact that we simulate each parameter separately. Diagonal steps occur only if both Gibbs steps are accepted at the same time. The proposals are, however, only horizontal and vertical. The resulting density estimates for the entire sample of 50,000 simulations are shown in the bottom row of Fig. 5.7.

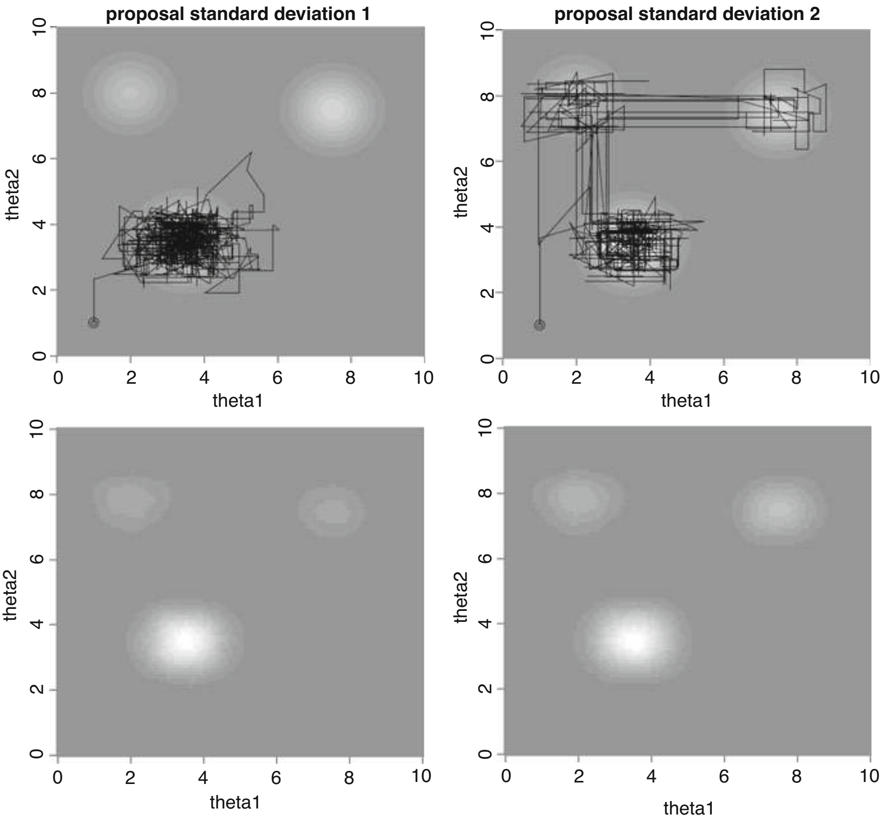

Fig. 5.7 Gibbs sampling combined with Metropolis-Hastings. The proposal density is normally distributed with $\sigma = 1$ (left side column) and $\sigma = 2$ (right side column) and centred on the previous value. The top row shows the first 1000 steps, the bottom row shows the kernel density estimate of 50,000 draws

The field of MCMC algorithms is very wide and numerous alternatives and modifications have been proposed. We refer the interested reader to Congdon (2003) or Held and Sabanés Bové (2014) for further material and references.

5.5 Variational Bayes

We now turn our attention to a recent method for the approximation of the posterior distribution, that makes use of **variational Bayes** principles. The approach intends to replace the true posterior distribution $f_\theta(.|y)$ with an approximation $q_\theta(.)$. The approximate distribution $q_\theta(.)$ belongs to a given class of distribution $\mathcal{Q}$ and is chosen such that the Kullback–Leibler divergence between the true (unknown) posterior and the approximation $\mathcal{Q}$ is minimised. We therefore look at

$$KL(f_\theta(.|y), q_\theta(.)) = \int \log \frac{q_\theta(\vartheta)}{f_\theta(\vartheta|y)} q_\theta(\vartheta) d\vartheta.$$

The idea is now to choose $q_\theta \in \mathcal{Q}$ such that it minimises the Kullback–Leibler divergence, i.e. we aim to find

$$\hat{q}_\theta = \arg\min_{q_\theta \in \mathcal{Q}} KL(f_\theta(\cdot|y), q_\theta(\cdot)).$$

Note that the Kullback–Leibler divergence decomposes to

$$KL(f_\theta(.|y), q_\theta(.)) = \int \log q_\theta(\vartheta) q_\theta(\vartheta) d\vartheta - \int \log f_\theta(\vartheta|y) q_\theta(\vartheta) d\vartheta,$$

where the first component is also called the entropy. Clearly taking $q_\theta(.) = f_\theta(.|y)$ minimises the Kullback–Leibler divergence. But instead of taking the exact posterior, the idea is to take $q_\theta(.)$ as some simple and easy to calculate distribution. For instance, we may choose $q_\theta(.)$ to be a normal distribution, i.e. $q_\theta(\theta) = \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp(-\frac{1}{2}(\theta - \mu_\theta)^2/\sigma_\theta^2)$ where we now choose μ_θ and σ_θ^2 such that the Kullback–Leibler divergence is minimised. This is comparable with a Laplace approximation discussed in Sect. 5.3.2. However, the variational method has a broader scope and is suitable even if the parameter vector θ is multi- or even high dimensional. Assume therefore that $\theta = (\theta_1, \ldots, \theta_p)^T$. A possible simplification occurs if we restrict the distribution $q_\theta(.)$ to independence, such that

$$q_\theta(\theta) = \prod_{k=1}^{p} q_k(\theta_k).$$

The component-wise independence could also be replaced by block-wise independence, that is, we could group the components of θ to blocks and assume independence between the blocks. For simplicity, we present only the idea for univariate independent components here. Note that with independence we get the following Kullback–Leibler divergence

$$\begin{aligned} KL(f_\theta(.|y), q_\theta(.)) &= \int \prod_{i=1}^{p} \log \frac{\prod_{j=1}^{p} q_j(\theta_j)}{f_\theta(\theta|y)} q_i(\theta_i) d\theta_i \\ &= \int q_k(\theta_k) \Big\{ \log q_k(\theta_k) - \underbrace{\int \prod_{i=1, i\neq k}^{p} \log(f_\theta(\theta|y)) q_i(\theta) d\theta_i}_{=: E_k(\log f_\theta(\theta|y)) =: \log \tilde{f}_k(\theta_k|y)} \Big\} d\theta_k \end{aligned}$$

$$+\int q_k(\theta_k)\left\{\int \prod_{i=1,i\neq k}^{p} \sum_{j=1,j\neq k}^{p} \log q_j(\theta) q_i(\theta) d\theta_i\right\} d\theta_k.$$

Note that only the first component depends on the true posterior. Moreover, the component $E_k(\log f_\theta(\theta|y))$ depends only on θ_k, as all of the other parameters have been integrated out. This allows us to denote this component as $\log \tilde{f}_k(\theta_k|y)$. It should also be clear that, $\tilde{f}_k(\theta_k|y)$ also depends on all marginal densities $q_j(\theta_j)$ for $j \neq k$, which is suppressed in the notation. Considering the first term of the sum above, we can comprehend this again as Kullback–Leibler divergence $KL(\tilde{f}_k(\cdot|y), q_k(\cdot))$, which is minimised if we choose

$$q_k(\theta_k) \propto \tilde{f}_k(\theta_k|y).$$

Note that $\tilde{f}_k(\theta_k|y)$ is not necessarily a density, because the integral over all values of θ_k is not necessarily equal to one. One may not even know the posterior $f_k(\theta_k|y)$ exactly, but only up to a normalisation constant. We therefore set

$$q_k(\theta) = \frac{\tilde{f}_k(\theta_k|y)}{\int \tilde{f}_k(\theta_k|y) d\theta_k},$$

where the integral is univariate only and may be calculated with numerical integration or even analytically, if possible. The idea is now to update iteratively one component after the other. Let $q_{\theta(t)} = \prod_{k=1}^{p} q_{k(t)}(\theta_k)$ be the current approximate. We then update the k-th component by

$$q_{k(t+1)}(\theta_k) = \frac{\tilde{f}_{k(t)}(\theta_k|y)}{\int \tilde{f}_{k(t)}(\vartheta_k|y) d\vartheta_k},$$

with obvious notation for $\tilde{f}_{k(t)}$. Variational Bayes has advantages in high dimensional parametric models, where the separate components are standard in structure. In this case, it can speed up the computational when compared with MCMC approaches. This happens in particular if one works with conjugated priors. We refer to Fox and Roberts (2012) for more details.

5.6 Exercises

Exercise 1

Let $Y_1, \ldots, Y_n$ be an *i.i.d.* sample from an exponential distribution $Exp(\lambda)$. As prior distribution for λ, we assume a Gamma distribution, $Ga(\alpha, \beta)$.

1. Derive the posterior distribution for λ and its mean and variance.
2. Calculate the posterior mode. For what choice of α and β are the posterior mode estimate and Maximum Likelihood estimate of λ numerically identical?

Exercise 2
Having derived the posterior distribution for a parameter θ given a sample $y = (y_1 \ldots, y_n)$, the posterior predictive distribution for a new observation y_{n+1} is defined as

$$f(y_{n+1}|y) = \int_{-\infty}^{+\infty} f(\theta|y) f(y_{n+1}|\theta) d\theta \ ,$$

where $f(\theta|y)$ is the posterior distribution and $f(y_{n+1}|\theta)$ is the likelihood contribution of the new observation.

Derive the posterior predictive distribution for a new observation for the case of an $i.i.d.$ sample y of a normal distribution $N(\mu, \sigma^2)$ with known variance σ^2 and flat constant prior 1 for μ.

Exercise 3 (Use R Statistical Software)
The file `ch6exerc3.csv` contains annual numbers Y_i, $i = 1, \ldots, 112$ of accidents due to disasters in British coal mines from 1850 to 1962 (the data are contained in the `R` package `GeDS`).

A change point model is applied to the data, which means that the parameters of the model change at a given point in time. To be specific, the model takes the following form:

$$X_i = \begin{cases} Poisson(\lambda_1) \ i = 1, \ldots, \theta, \\ Poisson(\lambda_2) \ i = \theta + 1, \ldots, 112, \end{cases}$$

where we assume as priors: $\lambda_i|\alpha \sim \Gamma(3, \alpha)$ $i.i.d.$, $i = 1, 2$ and $\alpha \sim \Gamma(10, 10)$. We assume that θ is known.

1. Derive the univariate full conditionals and describe a Gibbs sampler to get draws from the posterior distribution $p(\lambda_1, \lambda_2, \alpha|x, \theta)$. *Hint:* all three full conditionals are Gamma distributions.
2. Implement the Gibbs sampler in `R` and let it run for different values of θ. Use a heuristic criterion to decide for which θ a breakpoint exists.