Chapter 8
Bayesian Risk Analysis

Claudia Czado and Eike Christian Brechmann

Uncertainty in the behavior of quantities of interest causes risk. Therefore statistics is used to estimate these quantities and assess their variability. Classical statistical inference does not allow to incorporate expert knowledge or to assess the influence of modeling assumptions on the resulting estimates. This is however possible when following a Bayesian approach which therefore has gained increasing attention in recent years. The advantage over a classical approach is that the uncertainty in quantities of interest can be quantified through the posterior distribution. We first introduce the Bayesian approach and illustrate its use in simple examples, including linear regression models. For more complex statistical models Markov Chain Monte Carlo methods are needed to obtain an approximate sample from the posterior distribution. Due to the increase in computing power over the last years such methods become more and more attractive for solving complex problems which are intractable using classical statistics, for instance spam e-mail filtering or the analysis of gene expression data. We illustrate why these methods work and introduce two most commonly used algorithms: the Gibbs sampler and Metropolis Hastings algorithms. Both methods are derived and applied to statistical models useful in risk analysis. In particular a Gibbs sampler is developed for a change point detection in yearly counts of events and for a regression model with time dependence, while a Metropolis Hastings algorithm is derived for modeling claim frequencies in an insurance context.

Keywords Bayesian inference · Markov Chain Monte Carlo samplers · Bayesian risk · Prior · Posterior

Mathematics Subject Classification (2010) 62F15 · 62F86 · 62J05

C. Czado (✉) · E.C. Brechmann
Applied Statistics, Center for Mathematical Sciences, Technische Universität München,
Boltzmannstr. 3, 85748 Garching bei München, Germany
e-mail: cczado@ma.tum.de

C. Klüppelberg et al. (eds.), *Risk – A Multidisciplinary Introduction*,
DOI 10.1007/978-3-319-04486-6_8,

The Facts

- Risk is regarded as induced by the uncertainty in the behavior of quantities of interest. Therefore this random behavior has to be modeled using probability models and characteristics such as expected value and variance to be estimated.
- An introduction to Bayesian statistics is given, which—in contrast to classical statistics—can accommodate prior knowledge about the risk parameters under consideration, in particular using Bayes' famous theorem. Especially expert knowledge can be incorporated.
- Bayesian inference is based on the posterior distribution of the risk parameters which summarizes the knowledge about the risk quantity after the data is observed. Common Markov Chain Monte Carlo methods for deriving the Bayesian posterior distribution are discussed, namely the Gibbs sampler and Metropolis Hastings algorithms.
- Concepts are illustrated by examples from insurance, health care, mining and agriculture involving the risk quantities number of claims, complication rate of new medical treatment, number of coal-mining disasters and crop yield, respectively.

1 The Bayesian Approach

In this chapter we are interested in the study of quantities which are subject to uncertainty. In this context we understand risk as a process which is induced by uncertainty or randomness in the behavior of these quantities. To be more precise we will consider among other the following risk quantities: yearly crop rates, number of complications following a new medical treatment and the annual number of claims for a car insurance company. For the statistical risk analyst these quantities are random variables for which a probability distribution has to be chosen which depends on unknown population parameters and fits the observed data well. These population parameters determine the expectation and variance of the risk quantity. Classical—usually called *frequentist*—statistics uses solely the observed data to estimate the unknown population parameters. This is a sensible approach, however, the randomness in the observations and the limited number of observations available can lead to errors in subsequent inference. We assume that the reader has basic knowledge in probability and statistics; for convenience a glossary is provided in Appendix. Three illustrative examples are presented after this first short introduction.

In the simplest possible setting, we assume that observations come from a population whose members follow a specific probability distribution which depends on a single parameter θ. Given that we know this particular underlying distribution, we are interested in estimating θ based on the observed data. We denote such an estimate by $\hat{\theta}$. For example, if θ is the expectation of the distribution, we can estimate it by the average of all observations, that is

$$\hat{\theta} = \bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1.1}$$

where n is the number of observations with values $x_1, \ldots, x_n$.

In practice, the estimate $\hat{\theta}$ will however pretty much never equal the true parameter θ, that is, in general $\hat{\theta} \neq \theta$. Moreover, we might obtain an estimated value $\hat{\theta}$ which is unbelievable because it maybe lies outside a range where we expected the parameter to be in. If we however still believe that our probability model for the observed data is correct, we are in the dilemma that we have to decide between our belief in the data model and our prior belief in the parameter.

Bayesian statistics solves this problem by combining prior expert knowledge with information obtained from the observations. From now on, let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)' \in \Theta$ be the unknown parameter of interest belonging to the parameter space Θ, where usually $\Theta \subset \mathbb{R}^k$. Then we a priori assign a probability to each parameter value $\boldsymbol{\theta}$ according to the prior expert knowledge available, that is, we treat the population parameter as random variable and not as a fixed unknown quantity. Statistically speaking, this means that we choose an appropriate *prior distribution* with density or probability function $p(\boldsymbol{\theta})$, which summarizes the knowledge about the parameter of interest. We now observe a random sample $\boldsymbol{x} = (x_1, \ldots, x_n)'$, which are realizations of random variables $\boldsymbol{X} = (X_1, \ldots, X_n)'$ with true probability density $f(\cdot|\boldsymbol{\theta})$. For example x_i is the observed crop yield in plot i of the random crop yield X_i. Considering $f(\boldsymbol{x}|\boldsymbol{\theta})$ as a function of the parameter $\boldsymbol{\theta}$ for given observations $\boldsymbol{x}$ yields the *likelihood* denoted as

$$\ell(\boldsymbol{\theta}|\boldsymbol{x}) := f(\boldsymbol{x}|\boldsymbol{\theta}), \tag{1.2}$$

which summarizes the available information in the data about the parameter.

Note that in frequentist statistics, parameters are often estimated by so-called *maximum likelihood estimation* which means finding the parameter values $\hat{\boldsymbol{\theta}}$ that maximize (1.2), that is, finding the value of $\boldsymbol{\theta}$ which makes the observations "most likely". For example, the quantity in (1.1) is the maximum likelihood estimate of the expectation μ of a normal distribution (see Illustration 1.1 below).

In Bayesian statistics, we however would like to incorporate prior knowledge about the parameter $\boldsymbol{\theta}$, that is the prior distribution, into the estimation procedure. Since the observations $\boldsymbol{x}$ contain information about $\boldsymbol{\theta}$, we update our knowledge about $\boldsymbol{\theta}$ by considering the conditional distribution of $\boldsymbol{\theta}$ given observations $\boldsymbol{x}$. This distribution is called the *posterior distribution* and can be calculated by *Bayes' theorem* as

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{f(\boldsymbol{x})}, \tag{1.3}$$

where

$$f(\boldsymbol{x}) = \int_{\Theta} f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{1.4}$$

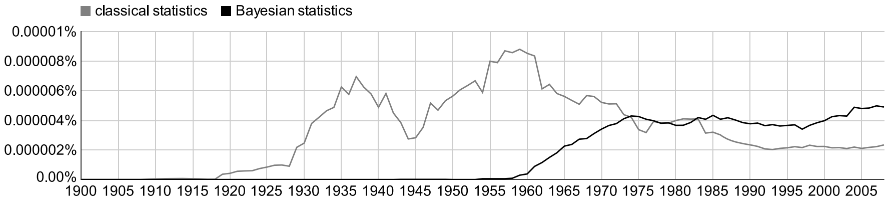


Fig. 1 Ngram of "classical statistics" (*gray*) and "Bayesian statistics" (*black*) created using Google Books Ngram Viewer available at http://books.google.com/ngrams

is the unconditional density function of the observations $\boldsymbol{x}$, called the *marginal distribution*. It does not depend on $\boldsymbol{\theta}$, in other words, it is only a normalizing constant with respect to $\boldsymbol{\theta}$ that ensures that the posterior distribution is a proper density expression integrating to 1. Hence it holds that

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto \ell(\boldsymbol{\theta}|\boldsymbol{x})p(\boldsymbol{\theta}), \tag{1.5}$$

that is, the posterior is proportional to the product of the likelihood and the prior. The computation of the posterior distribution however often is rather intricate so that so-called Markov Chain Monte Carlo methods are needed as discussed in Sect. 2.

A standard reference on Bayesian inference is the book by Berger [8], more recent references are Lee [5], Gelman et al. [15], Bolstad [1] and Hoff [20]. To illustrate the increasing importance of Bayesian methods in statistics, Fig. 1 shows how often the terms "classical statistics" and "Bayesian statistics" have occurred in books since 1900.

Three illustrative examples for different types of data (continuous, binary, count) are given below. These represent common types of risk quantities.

Illustration 1.1 (Crop Yields) Too small crop yields constitute a major risk to farmers. A reliable estimate of the expected crop yield and its variability therefore is needed for careful business planning. For this purpose, an agronomist studies the behavior of the random annual crop yields $X_1, \dots, X_n$ of n acres of the same size and with similar soil and growth conditions. From her experience and discussions with farmers she assumes that the crop yields are normally distributed with common mean θ and (known) variance σ^2 and independent of each other, that is $X_i \sim N(\theta, \sigma^2)$, $i = 1, \dots, n$. Then the likelihood (1.2) is

$$\ell(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x_i - \theta)^2\right] \propto \exp\left[-\frac{n}{2\sigma^2}(\overline{x} - \theta)^2\right],$$

where $\overline{x}$ is the empirical mean as defined in (1.1). It is a unimodal function in θ with mode given by $\overline{x}$.

From previous years the agronomist has some prior knowledge about the likely values of the expected crop yield θ and therefore specifies a prior distribution as normal with known mean μ and known variance τ^2. Having observed the crop yields

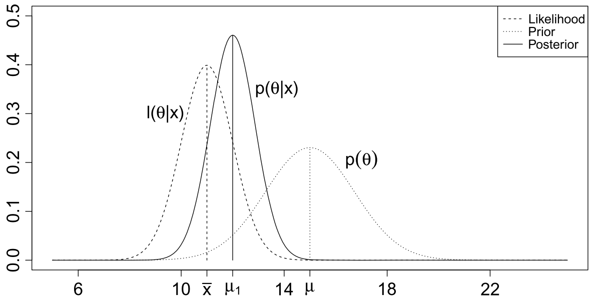


Fig. 2 Likelihood, prior and posterior densities for $n = 5$, observation variance $\sigma^2 = 5$, prior mean $\mu = 15$, prior variance $\tau^2 = 3$ and observed mean $\overline{x} = 11$

$x_1, \dots, x_n$, she therefore calculates the posterior density (1.3) using (1.5) as

$$p(\theta|\boldsymbol{x}) \propto \exp\left[-\frac{n}{2\sigma^2}(\overline{x}-\theta)^2\right]\exp\left[-\frac{1}{2\tau^2}(\theta-\mu)^2\right] \propto \exp\left[-\frac{1}{2}\frac{(\theta-\mu_1)^2}{\tau_1^2}\right], \tag{1.6}$$

where

$$\tau_1^2 = \frac{1}{n\sigma^{-2}+\tau^{-2}} \quad \text{and} \quad \mu_1 = \tau_1^2\left(\frac{\overline{x}}{n^{-1}\sigma^2}+\frac{\mu}{\tau^2}\right). \tag{1.7}$$

From (1.6) it follows that the posterior distribution is again normal but now with mean μ_1 and variance τ_1^2. To illustrate these concepts further, let us assume the agronomist expects an average yield of 15 per acre, that is, she sets the prior mean $\mu = 15$. She is however uncertain about her guess and therefore allows for a large uncertainty by choosing the prior variance τ^2 to be 3. After harvesting $n = 5$ acres, the observed average yield was $\overline{x} = 11$ per acre. The seed manufacturer claims that the variability under normal growing conditions is $\sigma^2 = 5$ per acre. Therefore the posterior distribution has posterior moments $\tau_1^2 = 0.75$ and $\mu_1 = 2$. This is illustrated in Fig. 2.

The expression of the posterior expectation μ_1 in (1.7) can conveniently be rewritten as

$$\mu_1 = w\overline{x} + (1-w)\mu, \tag{1.8}$$

where $w := w(\sigma^2, \tau^2, n) := \frac{\tau^2}{\tau^2+\sigma^2/n}$ is a weight varying from 0 to 1. Expression (1.8) shows that the posterior mean is the weighted average of the empirical mean $\overline{x}$ and the prior mean μ. As the uncertainty in the prior knowledge, reflected by the prior variance τ^2, increases, the weight $(1-w)$ for the prior mean decreases and the posterior mean is more heavily pulled towards the empirical mean. Moreover, the belief in the observed data as measured by the weight w also increases when the number of observations n, the number of acres under consideration, is increased. In the example it is $w = 0.75$. This means that there is already a quite strong belief in the data.

Illustration 1.2 (Complication Rate in Medical Studies) In a medical study, the researcher is interested in the rate of complications θ of n subjects. Clearly, the risk of the researcher is that this rate θ is higher than a small but admissible limit rate. At the end of the study, for each subject it is known whether he or she developed a complication or not. The event of complication occurrence can be modeled by a binary random variable X_i which is either 1 if the patient $i \in \{1, \ldots, n\}$ develops a complication or 0 otherwise. Because the researcher developed a completely new treatment, no prior knowledge about the success probability θ of the Bernoulli distribution representing the complication probability is available. Hence, she simply assumes equal likelihood for each parameter value θ, in other words, a prior density $p(\theta) = 1$ corresponding to the uniform distribution. For observations $x_1, \ldots, x_n$ the posterior distribution (1.3) for θ therefore simplifies to the likelihood (1.2):

$$p(\theta|\boldsymbol{x}) \propto \ell(\theta|\boldsymbol{x})p(\theta) = \ell(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}.$$

If, however, prior information based on studies of similar treatments is available, the researcher can specify a more informative prior distribution. For a parameter in the range of 0 to 1, the Beta distribution with parameters $\alpha > 0$ and $\beta > 0$ is a reasonable and quite flexible choice. Its density is given by

$$p(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}, \tag{1.9}$$

with normalizing constant $B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta$. Furthermore, its mean and variance are $E(\theta) = \alpha/(\alpha+\beta)$ and $\mathrm{Var}(\theta) = \alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$, respectively, and for $\alpha = \beta = 1$ the Beta distribution corresponds to the uniform distribution on $[0, 1]$. For example, if the researcher expects a 20 % complication rate with 0.1 standard error, then she solves $E(\theta) = 0.2$ and $\mathrm{Var}(\theta) = 0.1^2$ for α and β and obtains $\alpha = 3$ and $\beta = 12$.

It can be shown that the posterior distribution is again Beta with parameters $\alpha_1 = \sum_{i=1}^n x_i + \alpha$ and $\beta_1 = n - \sum_{i=1}^n x_i + \beta$. The posterior mean then can be written similarly to (1.8) as a weighted average of the sample mean and the prior mean:

$$\frac{\alpha_1}{\alpha_1 + \beta_1} = \frac{\sum_{i=1}^n x_i + \alpha}{n + \alpha + \beta} = w\overline{x} + (1-w)\frac{\alpha}{\alpha+\beta},$$

where $w := w(\alpha, \beta, n) := \frac{n}{n+\alpha+\beta}$. As before, belief in the observed data increases as the number of subjects n increases.

Illustration 1.3 (Claim Numbers in Car Insurance) In car insurance, a good estimate of the expected number of claims is essential for adequate policy pricing. An insurance company here faces a two-way risk. Overestimation of the expected number of claims means too high premiums and therefore a loss of clients. Expecting too few claims however poses the risk of large losses in the portfolio. Assuming that an insurance company has a portfolio of n homogeneous policy holders, a common choice for the distribution of the number of claims $X_i, i = 1, \dots, n$, is the Poisson distribution with mean and variance parameter θ and probability mass function

$$f(x_i|\theta) = \frac{\theta^{x_i}}{x_i!}e^{-\theta} \quad \text{for } x_i \in \{0, 1, 2, \dots\}. \tag{1.10}$$

Even if the portfolio consists of rather homogeneous policy holders, there is significant uncertainty regarding the expected number of claims θ because it also depends on unobservable quantities such as risk affinity or exogenous risks like extreme weather events.

The insurance company decides to choose a Gamma prior distribution with parameters $\alpha > 0$ and $\beta > 0$, mean α/β, and density

$$p(\theta) = \frac{1}{\Gamma(\alpha)}\beta^\alpha \theta^{\alpha-1} e^{-\beta\theta}, \tag{1.11}$$

where $\Gamma(\alpha)$ is the Gamma function $\Gamma(\alpha) = \int_0^\infty \theta^{\alpha-1}e^{-\theta}d\theta$.

The posterior distribution (1.3) based on observations $x_1, \dots, x_n$ from a previous year for example, is then obtained as follows:

$$p(\theta|\boldsymbol{x}) \propto \left[\prod_{i=1}^n \frac{\theta^{x_i}}{x_i!}e^{-\theta}\right] \frac{1}{\Gamma(\alpha)}\beta^\alpha \theta^{\alpha-1}e^{-\beta\theta} \propto \theta^{\alpha_1-1}e^{-\beta_1\theta},$$

which is again a Gamma distribution with parameters $\alpha_1 = \sum_{i=1}^n x_i + \alpha$ and $\beta_1 = n + \beta$. As before, the posterior mean can be decomposed into a weighted average of the empirical and prior mean. Such a convenient decomposition is however not always possible.

This mixture of Poisson and Gamma densities has another interesting interpretation: if the insurance company is interested in the claim number probabilities given an unknown parameter θ, Bayes' theorem can be "inverted" to compute the marginal density as $f(x_i) = f(x_i|\theta)p(\theta)/p(\theta|x_i)$ which results in a negative binomial distribution with the same mean as the Poisson distribution but with a higher variance due to the uncertainty in the unknown parameter.

1.1 From Non-informativeness to Conjugacy

Illustrations 1.1 and 1.2 also demonstrate a general problem of Bayesian statistics, namely the question: how do we choose an appropriate prior distribution? In certain applications, this choice might be evident but in general this is a non-trivial question and should be as objective as possible in order to not influence the results in an unwanted way. If for example in Illustration 1.1 the uncertainty in the prior knowledge τ^2 is very large, that is, the prior knowledge is rather vague, the prior will be close to $p(\boldsymbol{\theta}) \propto 1$ like the first prior choice in Illustration 1.2. Such a prior is called *non-informative* because it assigns equal likelihood to each possible parameter value.

One however has to be careful if the parameter space Θ is unbounded. In that case we have $\int_\Theta p(\boldsymbol{\theta})d\boldsymbol{\theta} = \infty$, and $p(\boldsymbol{\theta})$ is an improper prior.

Hence, such non-informative priors has to be dealt with care to ensure that the resulting posterior is proper. In Illustration 1.1, as $\tau^2 \to \infty$ corresponding to a non-informative prior, the posterior density is a normal density with mean $\overline{x}$ and variance $\frac{\sigma^2}{n}$, which is a proper distribution.

Another issue of non-informative priors is that they are not invariant under reparametrization of the model. For example a uniform prior on the success probability $\theta \in (0, 1)$ (see Illustration 1.2) does not result in a uniform prior on the so-called *odds* parameter given by $\theta/(1-\theta)$. An alternative approach for defining non-informative priors which has this invariance property was developed by Jeffreys [21]. Jeffreys prior is given as

$$p(\boldsymbol{\theta}) \propto \left|I(\boldsymbol{\theta})\right|^{\frac{1}{2}},$$

where

$$I(\boldsymbol{\theta}) = E\left[-\frac{\partial^2 \ln f(\boldsymbol{X}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\middle|\boldsymbol{\theta}\right] \tag{1.12}$$

is the expected *Fisher information matrix* about $\boldsymbol{\theta}$, which is a measure for the information about the parameter contained in the sample. In general, Jeffrey's approach leads to prior densities in the form of $p(\boldsymbol{\theta}) \propto 1$ for *location parameters* $\boldsymbol{\theta}$ and $p(\sigma) \propto \sigma^{-1}$ for *scale parameters* σ. For example the mean μ of a normal distribution is a location parameter and the standard error σ is a scale parameter.

On the other hand, the choice of an informative prior is always preferable if there is some kind of a priori knowledge about the parameter of interest. However, it will not be possible to get an analytically closed form expression of the posterior in complex situations, since the normalizing constant $f(\boldsymbol{x})$ defined in (1.4) of the posterior distribution requires a possibly high-dimensional integration. Posterior calculations are however simple if one considers *conjugate prior distributions*. A class of prior distributions $\mathcal{P}$ is conjugate to a class of observational models $\mathcal{F}$ if for every prior p out of $\mathcal{P}$ and for any observational distribution f from $\mathcal{F}$, the posterior distribution $p(\cdot|\boldsymbol{x})$ remains in the class of the prior distribution $\mathcal{P}$.

Example 1.4 (Conjugate Prior Distributions) The class of normal priors for the mean (Illustration 1.1) is conjugate for the observational model of normal distributions with known variance, while the class of Beta priors (Illustration 1.2) is conjugate for the observational model of Bernoulli distributions. Finally Illustration 1.3 also shows that the class of Gamma priors is conjugate for Poisson distributions.

1.2 Bayesian Inference

In Bayesian statistics all information about the parameter $\boldsymbol{\theta}$ is contained in the posterior distribution, while in classical statistics the information about $\boldsymbol{\theta}$ is captured by point and interval estimates. However, for the Bayesian, these quantities can be straightforwardly derived as well.

The main location measures are the *posterior mean*, as discussed in Illustrations 1.1–1.3, the *posterior median* and the *posterior mode*, where the last quantity is closest to the maximum likelihood principle from frequentist statistics, that is, the parameter $\boldsymbol{\theta}$ is most likely to be observed as judging from the available information contained in the observations. In maximum likelihood (ML) estimation we choose $\hat{\boldsymbol{\theta}}_{ML} = \text{argmax}_{\boldsymbol{\theta}\in\Theta}\, \ell(\boldsymbol{x}|\boldsymbol{\theta})$, while the posterior mode (PM) is augmented by the prior and given by $\hat{\boldsymbol{\theta}}_{PM} = \text{argmax}_{\boldsymbol{\theta}\in\Theta}\, \ell(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Note that, for example, for normal distributions the mean, mode and median coincide, while this is in general not the case, such as for the Gamma distribution.

The main dispersion measures are the *variance*, *standard deviation* (square root of the variance), *precision* (inverse of the variance) and *interquartile range* (difference between 75 %- and 25 %-quantiles) of the posterior distribution. Corresponding to the Fisher information defined in (1.12), one also often considers the *posterior curvature at the mode* which is the matrix of second derivatives of the posterior density in log form at the mode. If $\boldsymbol{\theta}$ is a vector, marginal densities can also be assessed.

In addition to these Bayesian point estimates $100(1-\alpha)$ % *credible intervals* provide interval estimates for $\boldsymbol{\theta}$ and are given for a scalar parameter θ by an interval $I(\boldsymbol{x})$, depending on the observations $\boldsymbol{x}$, such that

$$\int_{I(\boldsymbol{x})} p(\theta|\boldsymbol{x})d\theta = 1-\alpha.$$

In contrast to the confidence interval in classical statistics, the credible interval allows the interpretation that the parameter θ is contained with probability $1-\alpha$ in $I(\boldsymbol{x})$, since θ is now modeled as a random quantity.

Example 1.5 (Inference of the Normal Distribution) In Illustration 1.1 we have seen that the posterior distribution is given by the normal distribution with mean μ_1 and variance τ_1^2. Therefore the posterior mean, mode and median are μ_1, while the posterior variance is τ_1^2 and the posterior precision is τ_1^{-2}, which is also the posterior curvature at the mode.

A $100(1-\alpha)$ % credible interval $[\theta_l(\boldsymbol{x}), \theta_u(\boldsymbol{x})]$ for θ is given by appropriate quantiles of the posterior distribution: $\theta_l(\boldsymbol{x}) = \mu_1 - \tau_1\Phi^{-1}(1-\frac{\alpha}{2})$ and $\theta_u(\boldsymbol{x}) = \mu_1 + \tau_1\Phi^{-1}(1-\frac{\alpha}{2})$, where Φ^{-1} is the inverse of the standard normal distribution function. This is also the shortest possible credible interval. Note that the corresponding classical $100(1-\alpha)$ % and confidence interval for θ is given by $\bar{x} \pm \frac{s}{\sqrt{n}}\Phi^{-1}(1-\frac{\alpha}{2})$ where $s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ is the sample variance.

Returning to the specific example of Illustration 1.1, a corresponding 95 % credible interval for the mean yield is [10.303, 13.697], while a 95 % confidence interval is [9.040, 12.960] when assuming a sample variance of $s^2 = 5$. From the Bayesian theory the agronomist can say that the mean yield is between 10.303 and 13.697 with 95 % probability. The frequentist approach gives that the random interval $\bar{x} \pm \frac{s}{\sqrt{n}}\Phi^{-1}(1-\frac{\alpha}{2})$ covers the mean in 95 % of times. For the specific observations this interval is given by [9.040, 12.960].

1.3 Conjugacy and Regression Models

Before closing this section we consider the problem of modeling the influence of potential *explanatory variables* on a risk quantity called response. The simplest such model is the *linear regression model* for the *response* vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$:

$$Y_i \sim N\big(x_{i1}\beta_1 + \cdots + x_{id}\beta_d, \sigma^2\big) \quad \text{independent for } i = 1, \ldots, n, \tag{1.13}$$

where $x_{i1}, \ldots, x_{id}$ are known values of d explanatory variables for the ith observation and $\beta_1, \ldots, \beta_d$ are unknown *regression coefficients*. We can rewrite this model in matrix form as follows:

$$\boldsymbol{Y} \sim N_n\big(X\boldsymbol{\beta}, \sigma^2 I_n\big), \tag{1.14}$$

where $N_n(\boldsymbol{\mu}, \Sigma)$ denotes the n-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Further we define

$$X := \begin{pmatrix} x_{11} & \ldots & x_{1d} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{nd} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} := \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}.$$

The matrix X is called the *design matrix* and we assume that its columns are not linearly dependent.

Applications of such models can be found in virtually all areas of scientific research. For example, in Illustration 1.1 the agronomist may also try to model the crop yields with respect to a set of explanatory variables such as rainfall or sunshine duration. An experienced agronomist may have some prior expert knowledge about the effect of these variables and therefore can choose appropriate prior distributions for the regression coefficients. Similarly, based on her experience she may also be able to specify a prior for the variance parameter of the model parameters.

In model (1.14) it is more convenient to formulate priors in terms of $\boldsymbol{\beta}$ and the precision $\phi := \sigma^{-2}$. A typical choice is the Normal-Gamma, NG($\boldsymbol{b}_0, B_0, n_0, S_0$), prior, which is, for known constants n_0 and S_0, known vector $\boldsymbol{b}_0$ and known matrix B_0, defined in a hierarchical way as

$$\boldsymbol{\beta}|\phi \sim N_d\left(\boldsymbol{b}_0, \frac{B_0}{\phi}\right) \quad \text{and} \quad \phi \sim \text{Gamma}\left(\frac{n_0}{2}, \frac{n_0 S_0}{2}\right). \tag{1.15}$$

Equivalently we can assume $\boldsymbol{\beta}|\sigma^2 \sim N_d(\boldsymbol{b}_0, \sigma^2 B_0)$ and $\sigma^2 \sim$ Inverse Gamma($\frac{n_0}{2}$, $\frac{n_0 S_0}{2}$). Here the Inverse Gamma distribution is derived as follows: if $X \sim$ Gamma(α, β) then $1/X \sim$ Inverse Gamma(α, β). Under this setup the following theorem holds:

Theorem 1.6 (Conjugacy in Regression) *For the linear model given in* (1.14) *with observed response* $\mathbf{y}$ *and prior distribution given by* (1.15) *the posterior distribution* *of* $(\boldsymbol{\beta}, \phi)$ *is given by an* $NG(\boldsymbol{b}_1, B_1, n_1, S_1)$ *distribution with*

$$\boldsymbol{b}_1 = B_1\big(B_0^{-1}\boldsymbol{b}_0 + X'\boldsymbol{y}\big), \qquad B_1 = \big(B_0^{-1} + X'X\big)^{-1},$$

$$n_1 = n_0 + n, \qquad S_1 = \frac{1}{n_1}\big[n_0 S_0 + (\boldsymbol{y} - X\boldsymbol{b}_1)'y + (\boldsymbol{b}_0 - \boldsymbol{b}_1)'B_0^{-1}\boldsymbol{b}_0\big].$$

See Gamerman and Lopes [4, Sect. 2.3.2] for a proof.

2 MCMC—Markov Chain Monte Carlo

In Sect. 1.1 we studied the choice of prior distributions. In particular, we discussed non-informative priors and conjugate families which allow for an easy derivation of the posterior distribution (1.3). This is however not the case in general. *Markov Chain Monte Carlo* (MCMC) methods are used to approximate the posterior in more complex situations. Although being very computer intensive, the increasing availability of computing power nowadays makes the use of MCMC methods increasingly attractive. In particular, MCMC methods may be used to solve complex problems which cannot be treated using classical statistics. Examples of such problems are spam e-mail filtering and the analysis of gene expression data, just to name a few.

MCMC methods are based on the two well-known concepts of *Markov Chains* and *Monte Carlo* techniques. Both concepts will be explained first, before we then introduce the two most commonly used algorithms, namely the Gibbs sampler and Metropolis Hastings algorithms. Recent comprehensive references on MCMC methods include Gamerman and Lopes [4] and Marin and Robert [22].

2.1 ∗∗MC—Monte Carlo

To understand MCMC methods, we begin with the second "MC" which refers to "Monte Carlo" and which is due to the often used *Monte Carlo integration* techniques. In general Monte Carlo methods repeatedly sample from a probability distribution to determine analytically difficult quantities. For example, let us assume that $t(\cdot)$ is a function and we are interested in computing the integral

$$I = \int_0^1 t(\theta)d\theta, \tag{2.1}$$

of which no closed form solution is known. This is, for example, often the case for the marginal density function f defined in (1.4) which is part of the posterior distribution defined in (1.3). For such problems we use the following numerical approximation. First let $\theta \in (0, 1)$ be a random variable with density p. Then the expectation of the random variable $t(\theta)$ is $E(t(\theta)) = \int_0^1 t(\theta)p(\theta)d\theta$. If we can sample from p, an estimate of $E(t(\theta))$ is the sample mean. In particular, let θ be uniform on $(0, 1)$ and $\theta_1, \dots, \theta_n$ a corresponding independent and identically distributed (i.i.d.) random sample. Then (2.1) can be estimated by

$$\hat{I} := \frac{1}{n} \sum_{i=1}^{n} t(\theta_i). \tag{2.2}$$

By the strong law of large numbers (see Durrett [12]) $\hat{I}$ converges to $I = E(t(\theta))$ with probability 1, since $p(\theta) = 1$ for all $\theta \in (0, 1)$.

In Bayesian statistics the posterior expectation $E(t(\boldsymbol{\theta})|\boldsymbol{x})$ can be estimated by the sample mean (2.2) when $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$ is a sample from the posterior distribution $p(\cdot|\boldsymbol{x})$. As long as the posterior distribution and sampling algorithms are available, there are no problems and the first "MC" referring to "Markov chain" is not needed.

As mentioned above, it is unfortunately not the case that an analytical form of the posterior density $p(\cdot|\boldsymbol{x})$ is always available. The idea of MCMC methods therefore is to construct a Markov chain with limiting distribution $p(\cdot|\boldsymbol{x})$. If the Markov chain is run for a sufficiently long time, it can be assumed that the stationary state is reached and therefore the realizations of the chain represent a sample from $p(\cdot|\boldsymbol{x})$. In the following section we therefore give a brief overview of Markov chain theory. Readers familiar with it can skip Sect. 2.2 and continue reading with Sects. 2.3 and 2.4 which discuss the two most common MCMC methods.

2.2 *MC∗∗—Markov Chains*

We give a short introduction to Markov chains and state major results. A more detailed treatment can be found in Meyn and Tweedie [24], Nummelin [25], Resnick [26] and Guttorp [18]. The set of random variables $\{\boldsymbol{\theta}^{(t)} : t \in T\}$ is said to be a stochastic process taking values in the state space S for time points t in the index set T. In our discussion we will only consider discrete time stochastic processes with T being the set of natural numbers $\mathbb{N} = \{1, 2, \dots\}$. The state space S can generally be a subset of the d-dimensional set of real numbers, $\mathbb{R}^d$, but in the following we will concentrate on a discrete state space S. Details on continuous state space Markov chains can be found in Meyn and Tweedie [24].

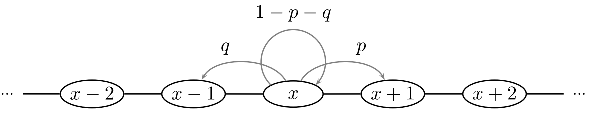
A *Markov chain* is a process, such that given the present state, past and future states are independent:

$$\begin{aligned} P\big(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{x}_{n+1} | \boldsymbol{\theta}^{(n)} = \boldsymbol{x}_n, \boldsymbol{\theta}^{(n-1)} = \boldsymbol{x}_{n-1}, \dots, \boldsymbol{\theta}^{(0)} = \boldsymbol{x}_0\big) \\ = P\big(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{x}_{n+1} | \boldsymbol{\theta}^{(n)} = \boldsymbol{x}_n\big) \end{aligned} \tag{2.3}$$

for all $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{n+1} \in S$. If the probabilities in (2.3) do not depend on n, we say that the Markov chain is *homogenous*. In this case we define the *transition probability* $P(\boldsymbol{x}, \boldsymbol{y})$ of moving from state $\boldsymbol{x}$ to state $\boldsymbol{y}$ as:

$$P(\boldsymbol{x}, \boldsymbol{y}) := P\big(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{y} | \boldsymbol{\theta}^{(n)} = \boldsymbol{x}\big).$$

Fig. 3 Probabilities of molecule movement



In general, for $A \subset S$, $P(\boldsymbol{x}, A) := \sum_{\boldsymbol{y} \in A} P(\boldsymbol{x}, \boldsymbol{y})$ is called the *transition kernel.*

Illustration 2.1 (Molecule Movement) Consider a molecule traveling in a liquid or a gas which moves independently left and right with successive displacements from its current position governed by a probability function f over the integers, that is $S = \mathbb{Z}$. Such a process is called a *random walk*. Let $\theta^{(n)}$ represent the position of the molecule at time n. Therefore we have

$$\theta^{(n)} = \theta^{(n-1)} + w_n = \theta^{(0)} + w_1 + \cdots + w_n,$$

where $w_i \sim f$ independently and for all $i \geq 1$. For the initial position $\theta^{(0)}$ we assume an initial distribution $\pi^{(0)}$.

The case where the probabilities of right, left or stay move are given by p, q and $1 - p - q$, respectively, is represented by assuming $f(1) = p$, $f(-1) = q$ and $f(0) = 1 - p - q$. This implies that

$$P(x, y) = P\big(\theta^{(n+1)} = y|\theta^{(n)} = x\big) = \begin{cases} p, & \text{if } y = x + 1, \\ q, & \text{if } y = x - 1, \\ 1 - p - q, & \text{if } y = x, \\ 0, & \text{if } y \neq x - 1, x, x + 1, \end{cases}$$

which is illustrated in Fig. 3.

If the state space $S \subset \mathbb{R}^d$ is not only discrete but also finite, that is $S = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_r\}$, we can consider the *transition matrix* P defined by

$$P := \begin{pmatrix} P(\boldsymbol{x}_1, \boldsymbol{x}_1) & \ldots & P(\boldsymbol{x}_1, \boldsymbol{x}_r) \\ \vdots & & \vdots \\ P(\boldsymbol{x}_r, \boldsymbol{x}_1) & \ldots & P(\boldsymbol{x}_r, \boldsymbol{x}_r) \end{pmatrix}.$$

Higher order transition probabilities P^m for $m \geq 2$ can be obtained as follows

$$\begin{aligned} P^m(\boldsymbol{x}, \boldsymbol{y}) &:= P\big(\boldsymbol{\theta}^{(m)} = \boldsymbol{y}|\boldsymbol{\theta}^{(0)} = \boldsymbol{x}\big) \\ &= \sum_{\boldsymbol{x}_1 \in S} \cdots \sum_{\boldsymbol{x}_{m-1} \in S} P\big(\boldsymbol{\theta}^{(m)} = \boldsymbol{y}, \boldsymbol{\theta}^{(m-1)} = \boldsymbol{x}_{m-1}, \ldots, \boldsymbol{\theta}^{(1)} = \boldsymbol{x}_1|\boldsymbol{\theta}^{(0)} = \boldsymbol{x}\big) \\ &= \sum_{\boldsymbol{x}_1 \in S} \cdots \sum_{\boldsymbol{x}_{m-1} \in S} P\big(\boldsymbol{\theta}^{(m)} = \boldsymbol{y}|\boldsymbol{\theta}^{(m-1)} = \boldsymbol{x}_{m-1}\big) \cdots P\big(\boldsymbol{\theta}^{(1)} = \boldsymbol{x}_1|\boldsymbol{\theta}^{(0)} = \boldsymbol{x}\big) \\ &= \sum_{\boldsymbol{x}_1 \in S} \cdots \sum_{\boldsymbol{x}_{m-1} \in S} P(\boldsymbol{x}, \boldsymbol{x}_1)P(\boldsymbol{x}_1, \boldsymbol{x}_2) \cdots P(\boldsymbol{x}_{m-1}, \boldsymbol{y}), \end{aligned}$$

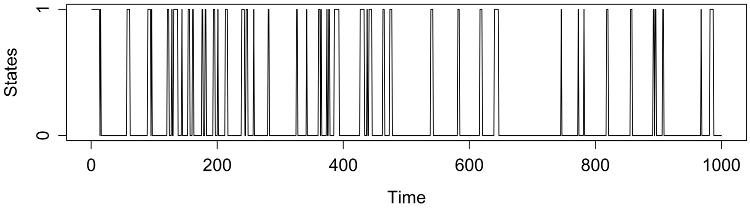


Fig. 4 Example of health states (0 = healthy, 1 = sick) of a policy holder over time ($p = 0.05$, $q = 0.3$, $\pi^{(0)}(0) = 0.8$, $\pi^{(0)}(1) = 0.2$)

where the second equality is due to the Markov property (2.3). In matrix notation we have $P^m = P \cdots P$ meaning matrix multiplication m times of the matrix P.

Further, let $\pi^{(0)}$ be the initial distribution of the chain, $\pi^{(0)}(\boldsymbol{x}) := P(\boldsymbol{\theta}^{(0)} = \boldsymbol{x})$. The marginal distribution after n time steps is given by

$$\begin{aligned}\pi^{(n)}(\boldsymbol{y}) := P\big(\boldsymbol{\theta}^{(n)} = \boldsymbol{y}\big) &= \sum_{\boldsymbol{x}\in S} P\big(\boldsymbol{\theta}^{(n)} = \boldsymbol{y}|\boldsymbol{\theta}^{(0)} = \boldsymbol{x}\big) P\big(\boldsymbol{\theta}^{(0)} = \boldsymbol{x}\big)\\ &= \sum_{\boldsymbol{x}\in S} P^n(\boldsymbol{x}, \boldsymbol{y})\pi^{(0)}(\boldsymbol{x}), \qquad (2.4)\end{aligned}$$

which can also be written as $\pi^{(n)} = \pi^{(0)}P^n = \pi^{(0)}P^{n-1}P = \pi^{(n-1)}P$.

Before we move on to discuss some major results which are the basis of MCMC methods, we consider an illustrative example.

Illustration 2.2 (Daily Allowance in Health Insurance) A health insurance company sells policies which pay a daily allowance to sick policy holders. In order to price the policies, the company sets up the following simplifying model. The health state of a person is modeled as a Markov chain $\{\theta^{(n)} : n \geq 0\}$ with states $S = \{\mathit{healthy}, \mathit{sick}\}$, denoted as $S = \{0, 1\}$, respectively. The initial distribution (the proportions of healthy and sick policy holders when the policy is sold) is denoted by $\pi^{(0)} = (\pi^{(0)}(0), \pi^{(0)}(1))'$ and the transition matrix P by

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} = \begin{pmatrix} P(0,0) & P(0,1) \\ P(1,0) & P(1,1) \end{pmatrix}.$$

That is, a healthy policy holder today is assumed to fall ill tomorrow with a probability of p versus staying healthy with a probability of $1 - p$. Similarly, a sick policy holder becomes healthy with a probability of q and stays sick with a probability of $1 - q$. An exemplary realization of this Markov chain is shown in Fig. 4

The probability that a person is healthy after n days (independent of whether or not he or she was sick in the meantime) is given by

$$\begin{aligned}P\big(\theta^{(n)} = 0\big) = P\big(\theta^{(n)} = 0|\theta^{(n-1)} = 0\big)P\big(\theta^{(n-1)} = 0\big)\\ + P\big(\theta^{(n)} = 0|\theta^{(n-1)} = 1\big)P\big(\theta^{(n-1)} = 1\big)\end{aligned}$$

$$
\begin{aligned}
&= (1-p)P\big(\theta^{(n-1)}=0\big) + qP\big(\theta^{(n-1)}=1\big)\\
&= (1-p-q)P\big(\theta^{(n-1)}=0\big) + q\\
&= (1-p-q)\big[(1-p-q)P\big(\theta^{(n-2)}=0\big)+q\big]+q\\
&\ \vdots\\
&= (1-p-q)^n\pi^{(0)}(0) + q\sum_{k=0}^{n-1}(1-p-q)^k.
\end{aligned}
$$

If $p = q = 0$, that is, healthy (sick) persons always stay healthy (sick), then $P(\theta^{(n)}=0)=\pi^{(0)}(0)$ and $P(\theta^{(n)}=1)=\pi^{(0)}(1)$. If $p+q>0$, using results for the finite geometric series gives

$$
\begin{aligned}
P\big(\theta^{(n)}=0\big) &= (1-p-q)^n\pi^{(0)}(0) + q\frac{1-[(1-p-q)^n]}{1-(1-p-q)}\\
&= (1-p-q)^n\left[\pi^{(0)}(0)-\frac{q}{p+q}\right]+\frac{q}{p+q}. \qquad (2.5)
\end{aligned}
$$

If the initial distribution is given by $\pi^{(0)} = (\frac{q}{p+q}, \frac{p}{p+q})'$, then the marginal probability $P(\theta^{(n)}=0)=\frac{q}{p+q}$ is the same for all time points n.

If $p+q<2$, then $(1-p-q)^n$ converges to zero as n goes to infinity and therefore

$$
\lim_{n\to\infty} P\big(\theta^{(n)}=0\big)=\frac{q}{p+q} \quad \text{and} \quad \lim_{n\to\infty} P\big(\theta^{(n)}=1\big)=\frac{p}{p+q},
$$

which shows that the initial distribution is obtained as the limiting distribution of the Markov chain. For the realizations of the Markov chain shown in Fig. 4 the convergence is illustrated in Table 1.

To obtain the probability that an initially healthy policy holder is also healthy after n days, denoted by $P^n(0,0)$, we assume that we always start in the healthy state, that is $\pi^{(0)}(0)=1$. Using (2.4) with $\pi^{(0)}(0)=1$ this gives

$$
\begin{aligned}
P^n(0,0) = P\big(\theta^{(n)}=0\big) &= (1-p-q)^n\left(1-\frac{q}{p+q}\right)+\frac{q}{p+q}\\
&= (1-p-q)^n\frac{p}{p+q}+\frac{q}{p+q}.
\end{aligned}
$$

Similarly, we compute $P^n(1,0)$, $P^n(0,1)$ and $P^n(1,1)$ to determine the nth order transition matrix P^n as

$$
P^n = \frac{(1-p-q)^n}{p+q}\begin{pmatrix} p & -q \\ -q & q \end{pmatrix} + \frac{1}{p+q}\begin{pmatrix} q & p \\ q & p \end{pmatrix}. \qquad (2.6)
$$

Table 1 Empirical marginal probabilities after different time points of the Markov chain shown in Fig. 4

Health states	Time										Limit prob.
	100	200	300	400	500	600	700	800	900	1000	
0	0.75	0.76	0.78	0.79	0.79	0.82	0.83	0.84	0.85	0.86	0.86
1	0.25	0.24	0.22	0.21	0.21	0.18	0.17	0.16	0.15	0.14	0.14

Finally, we denote by T_0 the first time that a person becomes healthy again. Given that he or she was healthy when taking out the policy, we have

$$\begin{aligned} P\big(T_0 = n|\theta^{(0)} = 0\big) &= P_0\big(\theta^{(n)} = 0, \theta^{(j)} \neq 0, 1 \leq j \leq n-1\big) \\ &= P(0,1)P(1,1)^{n-2}P(1,0) = p(1-q)^{n-2}q. \end{aligned}$$

Similarly let T_1 be the first time that a person falls ill. Then it holds that $P(T_1 = n|\theta^{(0)} = 0) = P(0,0)^{n-1}P(0,1) = p(1-p)^{n-1}$.

A fundamental problem for Markov chains in the context of simulation is the study of the asymptotic behavior of the chain as the number of steps or iterations n goes to infinity. A key concept for this is the *stationary distribution* π, which satisfies

$$\sum_{\boldsymbol{x} \in S} \pi(\boldsymbol{x})P(\boldsymbol{x}, \boldsymbol{y}) = \pi(\boldsymbol{y}) \quad \forall \boldsymbol{y} \in S, \tag{2.7}$$

and can be written in matrix notation as $\pi = \pi P$. The reason for the name is clear from the above equation. If the marginal distribution at any step n is π, then the distribution of the next step is πP. Once the chain reaches a stage where π is the distribution of the chain, the chain retains this distribution for all subsequent stages.

Illustration 2.3 (Illustration 2.2 Continued) Since the policies sold by the health insurance company are valid for the full lifetime of a policy holder, the company would like to investigate the long term expected proportions of healthy and sick persons. Since $S = \{0, 1\}$, in this case condition (2.7) is equivalent to

$$\pi(0)P(0,y) + \pi(1)P(1,y) = \pi(y), \quad y = 0, 1.$$

The solution is $\pi = (\frac{q}{p+q}, \frac{p}{p+q})$. Also for $p + q < 2$ it follows from (2.6) that

$$\lim_{n\to\infty} P^n = \frac{1}{p+q}\begin{pmatrix} q & p \\ q & p \end{pmatrix} = \begin{pmatrix} \pi(0) & \pi(1) \\ \pi(0) & \pi(1) \end{pmatrix}$$

and the distribution of $\theta^{(n)}$ converges to π at an exponential rate. This shows that for $p + q < 2$ the proportion of healthy and sick policy holders is asymptotically given by the stationary distribution π.

The case $p + q = 2$ still produces a stationary distribution π but this does not provide a unique limiting distribution since from (2.5) it follows that

$$P\left(\theta^{(n)} = 0\right) = (-1)^n \left(\pi^{(0)}(0) - \frac{q}{2}\right) + \frac{q}{2} \quad \forall n \geq 1.$$

This case is somewhat different, since the states are always alternating over time corresponding to the case that persons are healthy one day and always fall ill the next day which is evidently rather unrealistic. The chain has a periodic nature that will be addressed below.

Having established some basic properties of Markov chains, we are interested in characterizing the limiting behavior. For this a classification of the states of the Markov chain is necessary. For a more complete treatment see for example Chap. 2 of Resnick [26]. We define the first visit time to $\boldsymbol{y}$ as $T_{\boldsymbol{y}} = \inf\{n \geq 1 : \boldsymbol{\theta}^{(n)} = \boldsymbol{y}\}$ and the probability of visiting $\boldsymbol{y}$ after starting in $\boldsymbol{x}$ in finite time by $\rho_{\boldsymbol{x}\boldsymbol{y}} := P(T_{\boldsymbol{y}} < \infty | \boldsymbol{\theta}^{(0)} = \boldsymbol{x})$. Then a state $\boldsymbol{y} \in S$ is *recurrent* if and only if $\rho_{\boldsymbol{y}\boldsymbol{y}} = 1$, and—more strongly—*positive recurrent* if and only if $\boldsymbol{y}$ is recurrent and $E(T_{\boldsymbol{y}} | \boldsymbol{\theta}^{(0)} = \boldsymbol{y}) < \infty$. Further, the state $\boldsymbol{x}$ is said *to hit* $\boldsymbol{y}$ or $\boldsymbol{y}$ is *accessible* from $\boldsymbol{x}$, denoted by $\boldsymbol{x} \to \boldsymbol{y}$ if and only if $\rho_{\boldsymbol{x}\boldsymbol{y}} > 0$. One can show that $\boldsymbol{x} \to \boldsymbol{y}$ if and only if there exist an $n \geq 0$ such that $P^n(\boldsymbol{x}, \boldsymbol{y}) > 0$ (see Resnick [26, p. 78]). Let $\boldsymbol{x} \leftrightarrow \boldsymbol{y}$ if and only if $\boldsymbol{x} \to \boldsymbol{y}$ and $\boldsymbol{y} \to \boldsymbol{x}$. This is an equivalence relationship. The Markov chain is called *irreducible* if $\boldsymbol{x} \to \boldsymbol{y}$ for every pair $\boldsymbol{x}, \boldsymbol{y} \in S$.

Finally, to establish limit distributions one also needs to introduce the notation of periodicity. The *period* of state $\boldsymbol{x}$ is given by

$$d_{\boldsymbol{x}} = \text{largest common divisor of } \left\{n \geq 1 : P^n(\boldsymbol{x}, \boldsymbol{x}) > 0\right\}.$$

It follows that the condition $P(\boldsymbol{x}, \boldsymbol{x}) > 0$ implies $d_{\boldsymbol{x}} = 1$. Such a state is called *aperiodic*. Thus the states 0 and 1 in Illustration 2.3 are aperiodic if $p + q < 2$. On the other hand, if $p + q = 2$, it holds that $d_0 = d_1 = 2$, in other words, the states 0 and 1 are periodic with period 2.

A state $\boldsymbol{x}$ is called *ergodic* if it is aperiodic and positive recurrent. Similarly, a Markov chain is called *ergodic* if all states are aperiodic and positive recurrent. These concepts are sufficient to characterize the limiting distribution.

Theorem 2.4 (Limiting Distribution) *Let* $\{\boldsymbol{\theta}^{(n)}, n \geq 0\}$ *be an irreducible and ergodic Markov chain with stationary distribution* π, *then*

$$\lim_{n \to \infty} P^n(\boldsymbol{x}, \boldsymbol{y}) = \pi(\boldsymbol{y}) \quad \forall \boldsymbol{x}, \boldsymbol{y} \in S.$$

A proof can be found in Guttorp [18, Theorem 2.9]. This shows that the stationary distribution is also the limiting distribution under the assumptions of Theorem 2.4.

While the empirical mean converges to the population mean as the sample size increases for i.i.d. samples by the strong law of large numbers, a Markov chain equivalent will now be given.

Theorem 2.5 (Ergodic Theorem) *If the chain is ergodic and $E_\pi(t(\boldsymbol{\theta})) < \infty$ for the unique limiting distribution π then*

$$\bar{t}_n := \frac{1}{n}\sum_{i=1}^{n} t\big(\boldsymbol{\theta}^{(i)}\big) \xrightarrow{n\to\infty} E_\pi\big(t(\boldsymbol{\theta})\big) \quad \textit{with probability } 1.$$

A proof can be found on page 49 of Guttorp [18]. This theorem can be used as justification for using $\bar{t}_n$ as an estimate for $E_\pi(t(\boldsymbol{\theta}))$, see also the discussion in Sect. 2.1. A central limit theorem for Markov chains can also be formulated and is found for example in Gilks, Richardson, and Spiegelhalter [17]. It can be used for constructing asymptotic confidence intervals.

Having established the asymptotic theory of Markov chains, the final, and crucial, step is simulation. For this, consider an ergodic Markov chain $\{\boldsymbol{\theta}^{(n)}, n \geq 0\}$ with state space $S \subset \mathbb{R}^d$, transition probabilities $P(x, y)$ and initial distribution $\pi^{(0)}$. To generate values from this Markov chain the following algorithm can be used.

- Sample a starting value $\boldsymbol{\theta}^{(0)}$ from the initial distribution $\pi^{(0)}$.
- For $i = 1, \ldots, n$, sample value $\boldsymbol{\theta}^{(i)}$ from the probability mass function $f(\cdot) := P(\boldsymbol{\theta}^{(i-1)}, \cdot)$.

As n gets large the sampled values will have a distribution close to the limiting distribution π and can therefore be considered as an approximate sample from π. Note that all samples drawn after convergence are also samples from π since it is the stationary distribution. Here, convergence of a Markov chain means that the stationary distribution is approximated sufficiently accurately, which is difficult to assess. Relevant references will be given below. The values before convergence are called the *burn-in period* and will be deleted when considering the ergodic averages such as $\bar{t}_n$. The sampled values are dependent, since they arise from a Markov chain, however so-called thinning and batching methods can be applied to achieve an approximately i.i.d. sample. This general method of approximate sampling from the stationary distribution is called the *Markov Chain Monte Carlo* (MCMC) approach.

We can now use this approach to draw approximate samples from a complex posterior distribution $p(\cdot|\boldsymbol{x})$, which is analytically not tractable, by assuming that $p(\cdot|\boldsymbol{x})$ is the stationary distribution π of a Markov chain. The next two sections will study two famous MCMC algorithms in detail.

2.3 Gibbs Sampler

This chapter introduces and discusses the first widely used sampling scheme for constructing a Markov chain with prespecified limiting distribution π. It was first developed for approximately sampling from the *Gibbs distribution* used in image analysis. Geman and Geman [16] discussed this problem for several sampling schemes. Gelfand and Smith [14] were the first to point out to the statistical community at large that this sampling scheme could be used for other distributions than the Gibbs distribution. Before stating the sampling algorithm, we consider a small illustrative example.

Illustration 2.6 (Health States of a Couple) (Casella and George [2]) Let $S = \{(0,0)', (1,0)', (0,1)', (1,1)'\}$ be a two-dimensional state space with probability distribution π for the random vector $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ given by

$$\begin{aligned} &P(\theta_1 = 0, \theta_2 = 0) = \pi_{00}, \qquad &P(\theta_1 = 0, \theta_2 = 1) = \pi_{01}, \\ &P(\theta_1 = 1, \theta_2 = 0) = \pi_{10}, \qquad &P(\theta_1 = 1, \theta_2 = 1) = \pi_{11}. \end{aligned} \tag{2.8}$$

In view of Illustration 2.2 this can be interpreted as the healthy and sick states of a married couple. For example, if the first component corresponds to the health state of the husband and the second to that of his wife, then $\theta_1 = 1$ and $\theta_2 = 0$ indicates that the husband is sick, while his wife is healthy.

The Markov chain now consists of a bivariate vector $\boldsymbol{\theta}^{(n)} = (\theta_1^{(n)}, \theta_2^{(n)})'$ and the following transition probabilities are assumed.

- For $\theta_1^{(n)}$ the probability of moving from $\theta_2^{(n-1)} = j$ to $\theta_1^{(n)} = 0$ and $\theta_1^{(n)} = 1$, respectively, is given by

$$\pi_1(0|j) = \frac{\pi_{0j}}{\pi_{0j} + \pi_{1j}} \quad \text{and} \quad \pi_1(1|j) = \frac{\pi_{1j}}{\pi_{0j} + \pi_{1j}}. \tag{2.9}$$

  Note that $\pi_1(\cdot|j)$ is the conditional probability function of θ_1 given $\theta_2 = j$, $j = 0, 1$.

- For $\theta_2^{(n)}$ the probability of moving from $\theta_1^{(n)} = i$ to $\theta_2^{(n)} = 0$ and $\theta_2^{(n)} = 1$, respectively, is given by

$$\pi_2(0|i) = \frac{\pi_{i0}}{\pi_{i0} + \pi_{i1}} \quad \text{and} \quad \pi_2(1|i) = \frac{\pi_{i1}}{\pi_{i0} + \pi_{i1}}. \tag{2.10}$$

  Note that $\pi_2(\cdot|i)$ is the conditional probability function of θ_2 given $\theta_1 = i$, $i = 0, 1$.

This means that the husband's health state depends on his wife's yesterday's state and today's health state of the wife depends on today's health state of the husband. For a transition from state (i, j) yesterday to state (k, l) today we have

$$\theta_2^{(n-1)} = j \quad \xrightarrow{\frac{\pi_{kj}}{\pi_{0j}+\pi_{1j}}} \quad \theta_1^{(n)} = k \quad \xrightarrow{\frac{\pi_{kl}}{\pi_{k0}+\pi_{k1}}} \quad \theta_2^{(n)} = l.$$

Therefore the overall transition probability is given by

$$\begin{aligned} P\big((i,j),(k,l)\big) &= P\big(\boldsymbol{\theta}^{(n)} = (k,l)|\boldsymbol{\theta}^{(n-1)} = (i,j)\big) \\ &= P\big(\theta_2^{(n)} = l|\theta_1^{(n)} = k\big) P\big(\theta_1^{(n)} = k|\theta_2^{(n-1)} = j\big) \\ &= \frac{\pi_{kl}}{\pi_{k0} + \pi_{k1}} \frac{\pi_{kj}}{\pi_{0j} + \pi_{1j}}, \end{aligned}$$

for $(i, j), (k, l) \in S$. Thus a 4×4 transition matrix P can be formed.

One can further show that $\{\boldsymbol{\theta}^{(n)} = (\theta_1^{(n)}, \theta_2^{(n)})', n \geq 0\}$ forms a Markov chain and that π defined in (2.8) is the stationary distribution of the chain. If all elements of π are positive, it is also a limiting distribution. In particular, chains formed by the superposition of the conditional distributions have a stationary distribution given by the joint distribution.

Illustration 2.6 can easily be extended to the case where $\boldsymbol{\theta}$ consists of d components with $m_1, \dots, m_d$ values.

In general, *Gibbs sampling* is an MCMC scheme where the transition probabilities are formed by the full conditional distributions. Assume as before that the distribution of interest is $\pi(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_d)'$. Each of the d components can be a scalar, vector or matrix. Further assume that for each $i \in \{1, \dots, n\}$ the *full conditional distribution* for $\boldsymbol{\theta}_i$

$$\pi_i^{FC}(\boldsymbol{\theta}_i) := \pi(\boldsymbol{\theta}_i | \boldsymbol{\theta}_{-i}) \quad \text{where } \boldsymbol{\theta}_{-i} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \dots, \boldsymbol{\theta}_d)'$$

is known and can be sampled, for example, using Eqs. (2.9) and (2.10) in the above example. The Gibbs sampling algorithm can now be described as follows.

1. Set the iteration counter to $j = 1$ and set initial values $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \dots, \boldsymbol{\theta}_d^{(0)})'$.
2. Obtain a new value $\boldsymbol{\theta}^{(j)} = (\boldsymbol{\theta}_1^{(j)}, \dots, \boldsymbol{\theta}_d^{(j)})'$ through successive generation of values

$$\begin{aligned} \boldsymbol{\theta}_1^{(j)} &\sim \pi\big(\boldsymbol{\theta}_1 \big| \boldsymbol{\theta}_2^{(j-1)}, \dots, \boldsymbol{\theta}_d^{(j-1)}\big), \\ \boldsymbol{\theta}_2^{(j)} &\sim \pi\big(\boldsymbol{\theta}_2 \big| \boldsymbol{\theta}_1^{(j)}, \boldsymbol{\theta}_3^{(j-1)}, \dots, \boldsymbol{\theta}_d^{(j-1)}\big), \\ &\vdots \\ \boldsymbol{\theta}_d^{(j)} &\sim \pi\big(\boldsymbol{\theta}_d \big| \boldsymbol{\theta}_1^{(j)}, \dots, \boldsymbol{\theta}_{d-1}^{(j)}\big). \end{aligned}$$

3. Change counter j to $j + 1$ and return to step 2 until convergence is reached.

When convergence is reached the resulting value $\boldsymbol{\theta}^{(j)}$ is a draw from π. Often convergence is assessed by choosing an error bound $\varepsilon > 0$ and assuming convergence when the distance between $\boldsymbol{\theta}^{(n+1)}$ and $\boldsymbol{\theta}^{(n)}$ is less than ε. A further example is given in the following.

Illustration 2.7 (Coal Mining Disasters) Carlin, Gelfand, and Smith [10] discuss the following problem: yearly numbers $Y_1, \dots, Y_M$ of British coal-mining disasters as measured over more than a century are unlikely to have stayed at a similar level due to better technology and increased safety requirements. It is therefore reasonable to assume the presence of a change point $m \in \{1, \dots, M\}$ at which the general level of disasters significantly changed. Therefore Carlin et al. [10] assume the number of coal-mining disasters before that (unknown) change point to be Poisson distributed with another intensity parameter than after. They consider the following hierarchical model:

$$\begin{aligned} Y_i|\lambda, m &\sim \text{Poisson}(\lambda) \quad \text{for } i = 1, \ldots, m \text{ (independent)},\\ Y_i|\phi, m &\sim \text{Poisson}(\phi) \quad \text{for } i = m+1, \ldots, M \text{ (independent)},\\ \lambda &\sim \text{Gamma}(\alpha, \beta), \qquad\qquad (2.11)\\ \phi &\sim \text{Gamma}(\gamma, \delta),\\ m &\sim \text{uniform over } \{1, \ldots, M\}, \end{aligned}$$

where α, β, γ and δ are known constants and the model is termed "hierarchical", since the parameters of the Poisson distributions are modeled as random themselves. That is, m is the year where there is a significant change in the number of disasters as modeled by $Y_1, \ldots, Y_M$ with different (random) intensities λ and ϕ depending on whether Y_i is measured before or after the change point m, respectively. Due to missing prior knowledge about the change point m its distribution is modeled as uniform.

The joint posterior density of λ, ϕ and m given data $\boldsymbol{y} = (y_1, \ldots, y_M)'$ satisfies

$$\begin{aligned} &\pi(\lambda, \phi, m|\boldsymbol{y})\\ &\quad\propto f(y_1, \ldots, y_M|\lambda, \phi, m)p(\lambda, \phi, m)\\ &\quad= \left[\prod_{i=1}^{m} f_P(y_i; \lambda)\right]\left[\prod_{i=m+1}^{M} f_P(y_i; \phi)\right] f_G(\lambda; \alpha, \beta) f_G(\phi; \gamma, \delta) 1_{\{1,\ldots,M\}}(m)\\ &\quad\propto \left[\prod_{i=1}^{m} e^{-\lambda}\lambda^{y_i}\right]\left[\prod_{i=m+1}^{M} e^{-\phi}\phi^{y_i}\right]\lambda^{\alpha-1}e^{-\beta\lambda}\phi^{\gamma-1}e^{-\delta\phi}1_{\{1,\ldots,M\}}(m)\\ &\quad\propto \lambda^{\alpha+(\sum_{i=1}^{m} y_i)-1}e^{-(\beta+m)\lambda}\phi^{\gamma+(\sum_{i=m+1}^{M} y_i)-1}e^{-(\delta+M-m)\phi}1_{\{1,\ldots,M\}}(m), \end{aligned}$$

where 1_A is the indicator function satisfying $1_A(m) = 1$ if $m \in A$ and $1_A(m) = 0$ otherwise. Further f_P and f_G denote the Poisson and Gamma density functions, respectively (see Glossary A.2).

Therefore the full conditionals can be calculated as

$$\begin{aligned} \pi_\lambda^{FC}(\lambda) := p(\lambda|\phi, m, \boldsymbol{y}) &= \frac{p(\lambda, \phi, m, \boldsymbol{y})}{p(\phi, m, \boldsymbol{y})} = \frac{f(\boldsymbol{y}|\lambda, \phi, m)p(\lambda, \phi, m)}{p(\phi, m, \boldsymbol{y})}\\ &\propto \pi(\lambda, \phi, m|\boldsymbol{y}) \quad \text{as function of } \lambda\\ &\propto \lambda^{\alpha+(\sum_{i=1}^{m} y_i)-1}e^{-(\beta+m)\lambda} \propto \text{Gamma}\left(\alpha + \sum_{i=1}^{m} y_i, \beta + m\right) \end{aligned}$$

and similarly, $\pi_\phi^{FC}(\phi) \propto \text{Gamma}(\gamma + \sum_{i=m+1}^M y_i, \delta + M - m)$, and for the discrete random parameter m for $m = 1, \ldots, M$ as

$$\pi_m^{FC}(m) = \frac{\lambda^{\alpha+\sum_{i=1}^m y_i - 1} e^{-(\beta+m)\lambda} \phi^{\gamma+\sum_{i=m+1}^M y_i - 1} e^{-(\delta+M-m)\phi}}{\sum_{l=1}^M \lambda^{\alpha+\sum_{i=1}^l y_i - 1} e^{-(\beta+l)\lambda} \phi^{\gamma+\sum_{i=l+1}^M y_i - 1} e^{-(\delta+M-l)\phi}}.$$

Therefore the Gibbs sampler for (λ, ϕ, m) draws $\lambda^{(n+1)}$ from Gamma$(\alpha + \sum_{i=1}^{m^{(n)}} y_i, \beta + m^{(n)})$, $\phi^{(n+1)}$ from Gamma$(\gamma + \sum_{i=m^{(n)}+1}^n y_i, \delta + M - m^{(n)})$ and chooses $m^{(n+1)} = m$ with probability $\pi_m^{FC}(m)$. Here $\pi_m^{FC}(m)$ depends on $\lambda^{(n+1)}$ and $\phi^{(n+1)}$.

To get a first impression on the behavior of this Gibbs sampler, we simulated data from the model (2.11) with $M = 50, \alpha = 5, \beta = 1, \gamma = 1$ and $\delta = 1$ (left panel of Fig. 5) and implemented the Gibbs sampler for 100 iterations. Note that the Gamma priors for λ and ϕ are quite informative, since the signal-to-noise ratio (mean divided by standard deviation) is 1. For illustration we used the true values as starting values. In the left panel of Fig. 5 the data is presented and the time plots of the MCMC iterations and posterior density estimates for each parameter are shown in the right panel of the same figure. The true values are indicated by a vertical dotted line.

The time plots (first column of right panel) indicate that the sampler is converged, which we expect since we used the true values as starting values. The true values of λ and ϕ are reasonably in the center of the sampled posterior distribution. The sampler has no difficulty finding the true break point. In general, the assessment of convergence is difficult especially for higher dimensions and convergence diagnostics have to be considered.

We now establish a few basic facts for the Gibbs sampler. First of all the Gibbs sampler defines a Markov chain, since the update step at iteration j involves only values of the chain at $j - 1$. Also the chain is homogeneous, since transitions are only affected by the iteration through the chain values. The transition kernel from $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_d)'$ to $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)'$ is given by

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{i=1}^d \pi(\boldsymbol{\phi}_i | \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_d). \tag{2.12}$$

The limiting distribution of a Markov chain with transition kernel (2.12) is π, which we established for $d = 2$ and the discrete case in Illustration 2.6. For the continuous case the exact conditions under which the Markov chain resulting from the Gibbs sampler has limiting distribution π are given in Roberts and Smith [27]. For the continuous case π-irreducibility and aperiodicity are sufficient conditions (see Nummelin [25]). However, there are Markov chains derived from the Gibbs sampler which are not irreducible, see, for example, Gilks et al. [17]. Finally it can also be shown that π is stationary.

Even though theoretical results assure the convergence of the Gibbs sampler, they are difficult to validate theoretically for many complex statistical problems. In these

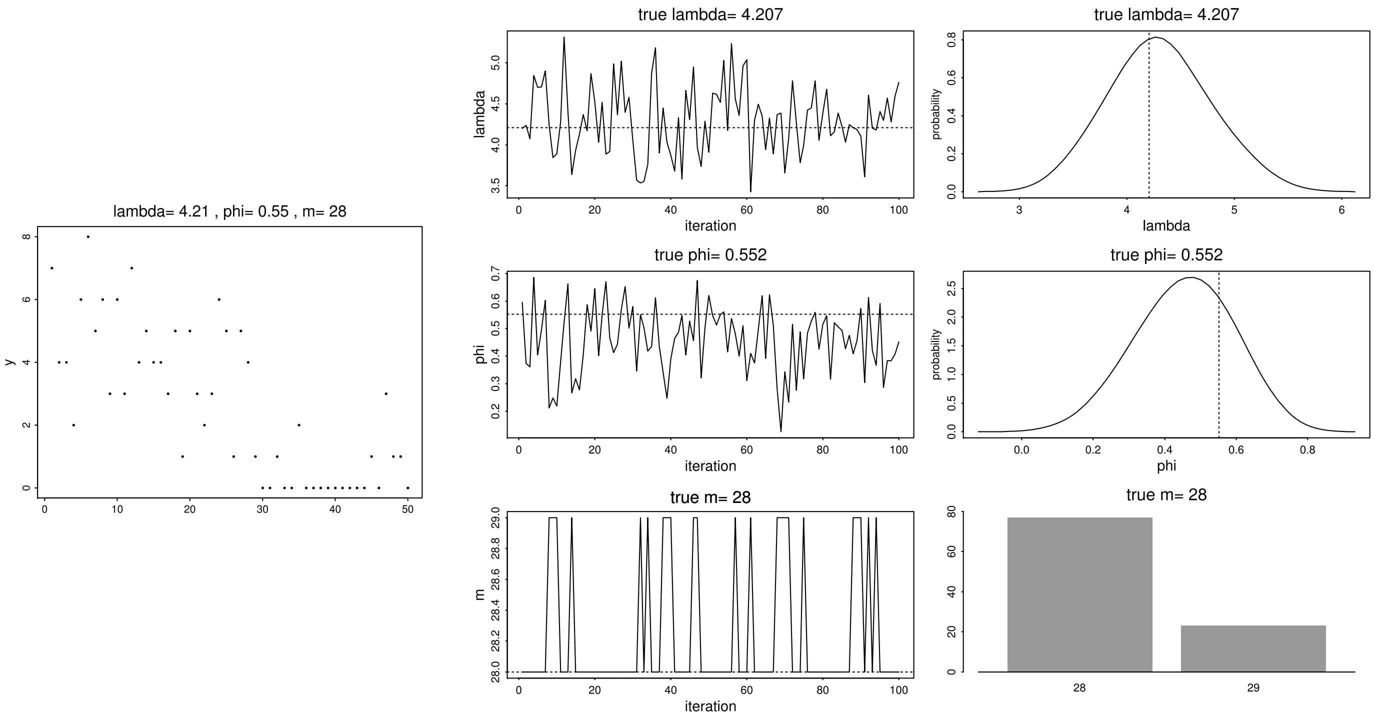


Fig. 5 *Left panel*: simulated data from model (2.11) with $M = 50$, $\alpha = 5$, $\beta = 1$, $\gamma = 1$ and $\delta = 1$. *Right panel*: time plot of MCMC iterations and posterior density estimates based on 100 iterations from the Gibbs sampler

cases a more practical approach is to assess the convergence by plotting n versus $\boldsymbol{\theta}^{(n)}$. If the variability of $\boldsymbol{\theta}^{(n)}$ for $n \geq n_0$ is approximately constant, then a burn-in of n_0 iterations is sufficient. Further MCMC sample based convergence assessments and comparison of several samplers with regard to burn-in iterations and required arithmetic operations are considered in Gilks et al. [17] and Marin and Robert [22] and the references therein.

Next, we draw attention to the use of the sample. For this, assume that we have a sample $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(n)}$ from the posterior distribution π now available as generated by the Gibbs sampler, after some burn-in period and possibly thinning or batching to reduce autocorrelation of the sampled MCMC iterates. Suppose we are interested in the posterior distribution of the statistics $\boldsymbol{\psi} = t(\boldsymbol{\theta})$. The standard estimator

$$\hat{\boldsymbol{\psi}} := \hat{E}_{\pi(\boldsymbol{\theta}|\boldsymbol{x})}(\boldsymbol{\psi}) := \frac{1}{n}\sum_{j=1}^{n} t\big(\boldsymbol{\theta}^{(j)}\big)$$

estimates the posterior mean $E_{\pi(\boldsymbol{\theta}|\boldsymbol{x})}(\boldsymbol{\psi})$ of $\boldsymbol{\psi}$, while the posterior variance $\sigma^2_{\boldsymbol{\psi}} := \mathrm{Var}_{\pi(\boldsymbol{\theta}|\boldsymbol{x})}(\boldsymbol{\psi}) = E_{\pi(\boldsymbol{\theta}|\boldsymbol{x})}(\boldsymbol{\psi}^2) - [E_{\pi(\boldsymbol{\theta}|\boldsymbol{x})}(\boldsymbol{\psi})^2]$ is estimated by

$$\hat{\sigma}^2_{\boldsymbol{\psi}} := \hat{E}_{\pi(\boldsymbol{\theta}|\boldsymbol{x})}\big(\boldsymbol{\psi}^2\big) - \big[\hat{E}_{\pi(\boldsymbol{\theta}|\boldsymbol{x})}(\boldsymbol{\psi})\big]^2 = \frac{1}{n}\sum_{j=1}^{n}\big[t\big(\boldsymbol{\theta}^{(j)}\big) - \hat{\boldsymbol{\psi}}\big]^2.$$

Moreover, *posterior credibility intervals* for $\boldsymbol{\psi}$ can be estimated by using sample quantiles as the estimates of the interval limits. For example if one is interested in estimating a 95 % credible interval for $\boldsymbol{\psi}$ and $n = 1000$, then the estimated credible interval is given as the interval between the 25th and 975th largest sampled value for $\boldsymbol{\psi}$. This section concludes with a continuation of the example on linear regression models.

Illustration 2.8 (Linear Regression with Ar(1) Disturbances) Sometimes the observed risk quantities are not independent, but might depend on previous observations. For example if we consider monthly plant growth rates, then the growth rate might depend on the variety but also on the previous month growth rate. Therefore we extend the linear regression model of Sect. 1.3 to include autoregressive lag 1 (AR(1)) disturbances, that is, the response variables are no longer assumed independent but dependent upon the previous response. We change indices from i to t to acknowledge the time dependencies. Similar to (1.13) the model is then given by

$$Y_t = x_{t1}\beta_1 + \cdots + x_{td}\beta_d + u_t \quad \text{where } u_t = \rho u_{t-1} + \varepsilon_t$$

for a time series of responses Y_t with possibly time dependent covariates $\boldsymbol{x}_t = (x_{t1}, \dots, x_{td})' \in \mathbb{R}^d$ for $t = 1, \dots, T$. Further we assume $|\rho| < 1$ and $\epsilon_t \sim N(0, \sigma^2)$ are i.i.d. As an initial condition we use $u_0 \sim N(0, \frac{\sigma^2}{1-\rho^2})$. The following informative priors can be used:

- $\boldsymbol{\beta}|\sigma^2 \sim N_d(\boldsymbol{\beta}_0, \sigma^2 A_0^{-1})$

- $\sigma^2 \sim$ Inverse Gamma$(\frac{n_0}{2}, \frac{\delta_0}{2})$
- $\rho \sim N(\rho_0, R_0^{-1})$ truncated to $(-1, 1)$, where a truncated normal distribution is a normal distribution whose values are bounded below, above or both. Thus the usual normal density is multiplied with an indicator function $1_{(a,b)}$ for an interval with endpoints $a < b$ and rescaled appropriately to ensure that it integrates to 1.

In the following we determine the full conditional distributions of the parameters, which can be used in a corresponding Gibbs sampling scheme.

1. *Regression parameter:* To update the vector of regression parameters $\boldsymbol{\beta}$ consider the following transformations

$$\boldsymbol{Y}^* := \begin{pmatrix} \sqrt{1-\rho^2}Y_1 \\ Y_2 - \rho Y_1 \\ Y_3 - \rho Y_2 \\ \vdots \\ Y_T - \rho Y_{T-1} \end{pmatrix} \quad \text{and} \quad X^* := \begin{pmatrix} \sqrt{1-\rho^2}\boldsymbol{x}_1' \\ \boldsymbol{x}_2' - \rho \boldsymbol{x}_1' \\ \boldsymbol{x}_3' - \rho \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_T' - \rho \boldsymbol{x}_{T-1}' \end{pmatrix}.$$

   Therefore $\boldsymbol{Y}^*$ follows a standard linear model with

$$\boldsymbol{Y}^* = X^*\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{where } \boldsymbol{\varepsilon} \sim N_T\big(0, \sigma^2 I_T\big).$$

   Since the full conditional for $\boldsymbol{\beta}$ given $\boldsymbol{Y}, X, \rho$ and σ^2 is the same as the full conditional for $\boldsymbol{\beta}$ given $\boldsymbol{Y}^*, X^*, \rho$ and σ^2, we can use Theorem 1.6 to show that

$$\boldsymbol{\beta}|\boldsymbol{Y}, X, \rho, \sigma^2 \sim N_p\big(\boldsymbol{\beta}_1, \sigma^2 B_1^{-1}\big),$$

   with $B_1 = (A_0 + X^{*\prime}X^*)^{-1}$ and $\boldsymbol{\beta}_1 = B_1(A_0\boldsymbol{\beta}_0 + X^{*\prime}\boldsymbol{Y}^*)$.
2. AR(1) *error variance:* By again considering the precision $\phi := \frac{1}{\sigma^2}$ and using the equality of the following conditional distributions $\phi|\boldsymbol{Y}, X, \boldsymbol{\beta}, \rho = \phi|\boldsymbol{Y}^*, X^*, \boldsymbol{\beta}, \rho$, it can be shown that

$$\sigma^2|\boldsymbol{Y}, X, \boldsymbol{\beta}, \rho \sim \text{Inverse Gamma}\left(\frac{n_1}{2}, \frac{\delta_1}{2}\right),$$

   with $n_1 = T + n_0 + d$ and $\delta_1 = \delta_0 + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'X^{*\prime}X^*(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + (\boldsymbol{Y}^* - X^*\hat{\boldsymbol{\beta}})'(\boldsymbol{Y}^* - X^*\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)'A_0(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$, where $\hat{\boldsymbol{\beta}} = (X^{*\prime}X^*)^{-1}X^{*\prime}\boldsymbol{Y}^*$.
3. *Correlation parameter:* Finally for updating the parameter ρ we can use Bayes' theorem to show that

$$\rho|\boldsymbol{Y}, X, \boldsymbol{\beta}, \sigma^2 \sim N(\tilde{\rho}, \tilde{R}) \text{ truncated to } (-1, 1),$$

   where $\tilde{R} := \sigma^{-2}(\sum_{t=1}^T u_{t-1}^2 + R_0)$ and $\tilde{\rho} := \tilde{R}^{-1}(\sigma^{-2}\sum_{t=1}^T u_t u_{t-1} + R_0\rho_0)$.

2.4 Metropolis Hastings Algorithms

The final MCMC algorithms presented here are the *Metropolis Hastings algorithms* (Metropolis et al. [23]; Hastings [19]). A nice introduction to the Metropolis Hastings algorithms is given in Chib and Greenberg [3]. As before, we are interested in constructing a Markov Chain with given stationary distribution π. First we consider a small example to motivate the discussion below.

Illustration 2.9 (Metropolis Hastings Algorithms) Consider a distribution π for $\boldsymbol{x} \in S$, where $S \subset \mathbb{R}^d$, $d \geq 1$. For a possible application recall Illustration 2.6, where we investigated the health states of a couple as modeled by the two-dimensional state space $S = \{0,1\}^2$ and the probability distribution π.

Our aim is to construct a Markov chain with stationary and limiting distribution π. For this, let Q be any four-dimensional irreducible transition matrix on S satisfying the symmetry condition $Q(\boldsymbol{x},\boldsymbol{y}) = Q(\boldsymbol{y},\boldsymbol{x})$ $\forall \boldsymbol{x},\boldsymbol{y} \in S$ and define a Markov chain $\{\boldsymbol{\theta}^{(n)}, n \geq 0\}$ as having transitions from $\boldsymbol{x}$ to $\boldsymbol{y}$ proposed according to the probabilities $Q(\boldsymbol{x},\boldsymbol{y})$. This proposed value for $\boldsymbol{\theta}^{(n+1)}$ is accepted with probability $\min\{1, \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}\}$ and rejected otherwise, leaving the chain in $\boldsymbol{x}$. This implies that for $\boldsymbol{x} \neq \boldsymbol{y}$

$$\begin{aligned} P(\boldsymbol{x},\boldsymbol{y}) &= P\big(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{y},\ \text{transition accepted}|\boldsymbol{\theta}^{(n)} = \boldsymbol{x}\big) \\ &= P\big(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{y}|\boldsymbol{\theta}^{(n)} = \boldsymbol{x}\big) P(\text{transition accepted}) \\ &= Q(\boldsymbol{x},\boldsymbol{y}) \min\left\{1, \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}\right\} \end{aligned}$$

and for $\boldsymbol{x} = \boldsymbol{y}$

$$\begin{aligned} P(\boldsymbol{x},\boldsymbol{x}) = P\big(&\boldsymbol{\theta}^{(n+1)} = \boldsymbol{x},\ \text{accepted}|\boldsymbol{\theta}^{(n)} = \boldsymbol{x}\big) \\ &+ P\big(\boldsymbol{\theta}^{(n+1)} \neq \boldsymbol{x},\ \text{not accepted}|\boldsymbol{\theta}^{(n)} = \boldsymbol{x}\big) \\ = P\big(&\boldsymbol{\theta}^{(n+1)} = \boldsymbol{x}|\boldsymbol{\theta}^{(n)} = \boldsymbol{x}\big) P(\text{accep.}) \\ &+ \sum_{\boldsymbol{y} \neq \boldsymbol{x}} P\big(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{y}|\boldsymbol{\theta}^{(n)} = \boldsymbol{x}\big) P(\text{not accep.}) \\ = Q(&\boldsymbol{x},\boldsymbol{x}) \min\left\{1, \frac{\pi(\boldsymbol{x})}{\pi(\boldsymbol{x})}\right\} + \sum_{\boldsymbol{y} \neq \boldsymbol{x}} Q(\boldsymbol{x},\boldsymbol{y}) \left[1 - \min\left\{1, \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}\right\}\right]. \end{aligned}$$

Further observe that if we assume that $\pi(\boldsymbol{y}) > \pi(\boldsymbol{x})$ for $\boldsymbol{x} \neq \boldsymbol{y}$, then

$$\begin{aligned} \pi(\boldsymbol{x}) P(\boldsymbol{x},\boldsymbol{y}) = \pi(\boldsymbol{x}) Q(\boldsymbol{x},\boldsymbol{y}) \min\left\{1, \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}\right\} &= \pi(\boldsymbol{x}) Q(\boldsymbol{x},\boldsymbol{y}) \\ = \pi(\boldsymbol{y}) \min\left\{1, \frac{\pi(\boldsymbol{x})}{\pi(\boldsymbol{y})}\right\} Q(\boldsymbol{y},\boldsymbol{x}) &= \pi(\boldsymbol{y}) P(\boldsymbol{y},\boldsymbol{x}), \end{aligned}$$

and similarly if $\pi(\boldsymbol{y}) < \pi(\boldsymbol{x})$. This result is referred to as *reversibility* of a Markov chain and ensures that π constitutes the stationary distribution of the chain. If Q is aperiodic, so will be P and the stationary distribution is also the limiting distribution.

In general, Metropolis Hastings algorithms also exploit the concept of reversibility as in Illustration 2.9. That is, in order to construct a Markov chain with stationary distribution π we require the following *reversibility condition* for the transition kernel $P(\boldsymbol{\theta}, \boldsymbol{\phi})$:

$$\pi(\boldsymbol{\theta})P(\boldsymbol{\theta}, \boldsymbol{\phi}) = \pi(\boldsymbol{\phi})P(\boldsymbol{\phi}, \boldsymbol{\theta}) \quad \forall \boldsymbol{\theta}, \boldsymbol{\phi}.$$

Hastings [19] proposes to define the acceptance probability in such a way that when combined with an arbitrary transition probability, it defines a reversible chain. Such an acceptance probability is given by

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) = \begin{cases} \min\{1, \frac{\pi(\boldsymbol{\phi})Q(\boldsymbol{\phi},\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})Q(\boldsymbol{\theta},\boldsymbol{\phi})}\}, & \text{if } \pi(\boldsymbol{\theta})Q(\boldsymbol{\theta}, \boldsymbol{\phi}) > 0, \\ 1, & \text{otherwise.} \end{cases} \tag{2.13}$$

Algorithms based on (2.13) are called *Metropolis Hastings* (*MH*) *algorithms*. MH algorithms define reversible chains with stationary distribution π if $P(\boldsymbol{\theta}, \boldsymbol{\phi}) > 0$. Roberts and Smith [27] show that if Q is irreducible and aperiodic and $\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) > 0$ for all $(\boldsymbol{\theta}, \boldsymbol{\phi})$, then the algorithm defines an irreducible and aperiodic Markov chain with limiting distribution π. The MH algorithm can now be described as follows:

1. Set iteration counter $j = 1$ and arbitrary initial value $\boldsymbol{\theta}^{(0)}$.
2. Move the chain to a new value $\boldsymbol{\phi}$ generated from the density $Q(\boldsymbol{\theta}^{(j-1)}, \cdot)$.
3. Evaluate the acceptance probability of the move given by $\alpha(\boldsymbol{\theta}^{(j-1)}, \boldsymbol{\phi})$ in (2.13). If the move is accepted, then $\boldsymbol{\theta}^{(j)} = \boldsymbol{\phi}$. If the move is not accepted, then $\boldsymbol{\theta}^{(j)} = \boldsymbol{\theta}^{(j-1)}$ and the chain does not move.
4. Change the counter from j to $j + 1$ and return to Step 2 until convergence is reached.

Step 3 can easily be performed by generating an independent uniform quantity u. If $u \le \alpha$, then the move is accepted and else it is not.

Note that you do not need to know the often complicated normalizing constant of the stationary distribution π to perform the MH algorithm. Further, when using a symmetric proposal probability as in Illustration 2.9, (2.13) simplifies to $\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) = \min\{1, \frac{\pi(\boldsymbol{\phi})}{\pi(\boldsymbol{\theta})}\}$ if $\pi(\boldsymbol{\theta}) > 0$ and $\alpha(\boldsymbol{\theta}, \boldsymbol{\phi}) = 1$ otherwise. Other common choices for Q lead to a *random walk* (new value = old value + disturbance; Illustration 2.1), *independence* (new value chosen independently of old value) or *hybrid chains* (Metropolis within Gibbs algorithm).

We close our discussion of MCMC methods with an example resuming and extending the Poisson model for claim frequencies of Illustration 1.3.

Illustration 2.10 (Claim Frequencies) Scollnik [29] considered the following model for modeling claim frequency data for group insurance policies: let X_{ij} be the number of claims for the ith group of policy holders in the jth policy year and P_{ij} the payroll count for the ith group of company employees in the jth policy year for $i=1,\dots,I$, $j=1,\dots,J$. The payroll counts give the number of employees which are at risk to incur a claim. The dependency among the claim counts over different years for the same policy i is modeled by introducing an unobserved random unit rate θ_i which has a common distribution for all policies. In particular Scollnik [29] assumed that X_{ij} given θ_i are independent with

$$X_{ij}|P_{ij},\theta_i \sim \text{Poisson}(P_{ij}\theta_i), \qquad \alpha \sim \text{Gamma}(5,5),$$
$$\theta_i|\alpha,\beta \sim \text{Gamma}(\alpha,\beta), \qquad \beta \sim \text{Gamma}(25,1).$$

The prior specification for α and β are rather arbitrary, but they imply that each θ_i has a prior mean and standard deviation approximately equal to 0.041 and 0.048, which might not be unreasonable in this context according to Scollnik [29]. Denote by $\boldsymbol{X}_i=(X_{i1},\dots,X_{iJ})'$ the number of claims vector of policy group i over all years and $\boldsymbol{X}=(\boldsymbol{X}_1',\dots,\boldsymbol{X}_I')'$ the total number of claims vector. Further, let $\boldsymbol{\theta}=(\theta_1,\dots,\theta_I)'$. Then the joint distribution of $(\boldsymbol{X},\boldsymbol{\theta},\alpha,\beta)$ can be written as follows:

$$p(\boldsymbol{X},\boldsymbol{\theta},\alpha,\beta)=\left[\prod_{j=1}^{J}\prod_{i=1}^{I} f_P(X_{ij}|P_{ij},\theta_i)\right]\left[\prod_{i=1}^{I} f_G(\theta_i|\alpha,\beta)\right]p(\alpha)p(\beta).$$

To update the unobserved latent rates θ_i we have as full conditional

$$\begin{aligned} p(\theta_i|\boldsymbol{X},\boldsymbol{\theta}_{-i},\alpha,\beta) &\propto \left[\prod_{j=1}^{J} f_P(X_{ij}|P_{ij},\theta_i)\right] f_G(\theta_i|\alpha,\beta) \\ &\propto \theta_i^{\alpha+\sum_{j=1}^{J}X_{ij}-1}\exp\left[-\left[\beta+\sum_{j=1}^{J}P_{ij}\right]\theta_i\right], \end{aligned}$$

which is a Gamma distribution with parameters $\alpha+\sum_{j=1}^{J}X_{ij}$ and $\beta+\sum_{j=1}^{J}P_{ij}$ and where $\boldsymbol{\theta}_{-i}=(\theta_1,\dots,\theta_{i-1},\theta_{i+1},\dots,\theta_I)'$. We see that these conditionals are actually independent of $\boldsymbol{\theta}_{-i}$. For updating α note that

$$p(\alpha|\boldsymbol{X},\boldsymbol{\theta},\beta)\propto\prod_{i=1}^{I} f_G(\theta_i|\alpha,\beta)p(\alpha)\propto\left[\frac{\beta^\alpha}{\Gamma(\alpha)}\right]^I\left[\prod_{i=1}^{I}\theta_i\right]^\alpha\alpha^4\exp(-5\alpha).$$

This is not a standard distribution and an MH step is needed.

Finally, to update β, we obtain for $\beta|\boldsymbol{X},\boldsymbol{\theta},\alpha$ again a Gamma distribution with parameters $I\alpha+25$ and $\sum_{i=1}^{I}\theta_i+1$.

According to Scollnik [29] we implemented a hybrid chain for the small data set with $I=3$ and $J=5$ shown in Table 2 using `WinBUGS` (**B**ayesian inference **U**sing **G**ibbs **S**ampling; http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml), which can be called directly from the statistical computing environment `R` (see

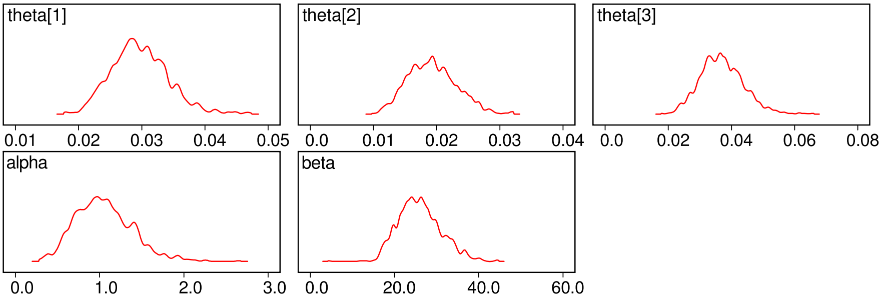


Fig. 6 Estimated posterior densities of 1000 iterations for $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)'$, α and β

Table 2 Data set of claim numbers and payroll counts for groups of policy holders and policy years

Year	Group 1		Group 2		Group 3	
	Payroll	Claims	Payroll	Claims	Payroll	Claims
1	280	9	260	6	267	6
2	320	7	275	4	145	8
3	265	6	240	2	120	3
4	340	13	265	8	105	4
5	325	10	285	5	115	7

Ntzoufras [7] for more information). The estimated posterior densities of 1000 iterations are shown in Fig. 6.

3 Food for Thought

There is software for Bayesian inference based on MCMC methods available in specialized problems. To the interested reader we particularly recommend to have a look at the above mentioned software `WinBUGS` and the illustrative book by Ntzoufras [7]. The recent book by Lunn et al. [6] also covers software for Bayesian statistical methods.

Another important issue of MCMC methods which could not be treated here appropriately are burn-in diagnostics which were briefly mentioned in Sect. 2.2 and provide tools for determining when we consider the values of the sampler as realizations from the posterior distribution. Further information can be found for example in Cowles and Carlin [11] and in Brooks and Roberts [9]. Related to this is the theoretical study of convergence questions.

Other areas of interest are, on the one hand, so-called ABC (Approximate Bayesian computation) methods which were developed for computationally very complex problems such as large-scale applications. Roberts et al. [28] and Frühwirth-Schnatter and Sögner [13], on the other hand, use MCMC methods for estimating stochastic volatility models commonly used in financial applications.

4 Summary

In this chapter, we gave a brief introduction to the main concepts of Bayesian statistics. After discussing the fundamental Bayes' theorem and three illustrating examples, we examined the problem of an appropriate prior choice in more detail and introduced Bayesian inference techniques. The first section closed with the commonly used linear regression model.

In the second section, we introduced the important class of MCMC methods, which are increasingly becoming popular for estimating parameters in complex statistical models. They are based on Monte Carlo techniques and properties of Markov chains, which were discussed before turning to the two most common MCMC algorithms, namely the Gibbs sampler and the Metropolis Hastings algorithms. There were discussed and illustrated using relevant examples involving risk quantities on different scales and with different contexts.

Appendix: Glossary

A.1 Foundations

Symbol	Explanation
X	random variable (r.v.)
$X = x$	realization or observed value of r.v. X
X continuous	r.v. X takes on any value in an interval (e.g., $X =$ annual crop yield $\in [0, \infty)$)
X discrete	r.v. X takes on only finite or countable many values (e.g., $X =$ number of mining disasters $\in \{0, 1, 2, \ldots\}$)
i.i.d.	independent and identically distributed
θ	unknown parameter of a distribution (e.g., $\theta =$ probability of occurrence of a complication after a medical treatment)
$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$	unknown parameters of a distribution (e.g., $\boldsymbol{\theta} = (\mu, \sigma^2)$, μ mean, σ^2 variance of a normal distribution)
$P_{\boldsymbol{\theta}}(A)$	probability that event A occurs when parameters $\boldsymbol{\theta}$ are true
$F(x\|\boldsymbol{\theta})$	cumulative distribution function (cdf) of r.v. X, i.e., $F(x\|\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(X \leq x)$
$f(x\|\boldsymbol{\theta})$	probability density function (pdf), when X continuous, i.e., $f(x\|\boldsymbol{\theta}) \geq 0$, $\int_{-\infty}^{\infty} f(x\|\boldsymbol{\theta})dx = 1$, $P_{\boldsymbol{\theta}}(X \leq x) = \int_{-\infty}^{x} f(x\|\boldsymbol{\theta})dx$
$f(x\|\boldsymbol{\theta})$	probability mass function (pmf), when X discrete, i.e., $f(x\|\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(X = x)$
$\mu = E(X)$	mean or expectation of r.v. X ($E(X) = \int_{-\infty}^{\infty} xf(x\|\boldsymbol{\theta})dx$ for X continuous)
$\sigma^2 = \text{Var}(X)$	variance of r.v. X ($\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x\|\boldsymbol{\theta})dx$ for X continuous)
$\phi = \frac{1}{\sigma^2}$	precision of r.v. X
$X \sim F(\cdot\|\boldsymbol{\theta})$	X has cdf $F(\cdot\|\boldsymbol{\theta})$

Symbol	Explanation
$X \sim f(\cdot\mid\boldsymbol{\theta})$	X has pdf/pmf $f(\cdot\mid\boldsymbol{\theta})$
$(X, Y) \sim f(\cdot,\cdot\mid\boldsymbol{\theta})$	r.v.s X and Y have joint pdf/pmf $f(\cdot,\cdot\mid\boldsymbol{\theta})$
$f_X(x\mid\boldsymbol{\theta})$ $(f_Y(y\mid\boldsymbol{\theta}))$	marginal pdf for X (Y): $f_X(x\mid\boldsymbol{\theta}) = \int_{-\infty}^{\infty} f(x, y\mid\boldsymbol{\theta})dy$ $(f_Y(y\mid\boldsymbol{\theta}) = \int_{-\infty}^{\infty} f(x, y\mid\boldsymbol{\theta})dx)$
$f_X(x\mid\boldsymbol{\theta})$ $(f_Y(y\mid\boldsymbol{\theta}))$	marginal pmf for X (Y): $f_X(x\mid\boldsymbol{\theta}) = \sum_{i=1}^{\infty} f(x, y_i\mid\boldsymbol{\theta})$ $(f_Y(y\mid\boldsymbol{\theta}) = \sum_{i=1}^{\infty} f(x_i, y\mid\boldsymbol{\theta}))$
$P_{\boldsymbol{\theta}}(A\mid B)$	conditional probability of A given B: $P_{\boldsymbol{\theta}}(A\mid B) = \frac{P_{\boldsymbol{\theta}}(A\cap B)}{P_{\boldsymbol{\theta}}(B)}$ if $P_{\boldsymbol{\theta}}(B) > 0$
x_α	α-quantile of continuous r.v. X: $P_{\boldsymbol{\theta}}(X \le x_\alpha) = \alpha$
$x_{0.5}$	median of continuous r.v. X
x_{mode}	mode of continuous r.v. X, that is the value which maximizes $f(x\mid\boldsymbol{\theta})$ over x
$\boldsymbol{X} = (X_1, \ldots, X_n)'$	$\boldsymbol{X}$ random vector, where $X_1, \ldots, X_n$ r.v.s
$F(\boldsymbol{x}\mid\boldsymbol{\theta})$ $(f(\boldsymbol{x}\mid\boldsymbol{\theta}))$	cdf (pdf/pmf) of $\boldsymbol{X}$
$E(\boldsymbol{X}) = (E(X_1), \ldots, E(X_n))$	mean vector of random vector $\boldsymbol{X}$
$\Sigma = (\Sigma_{ij})_{i,j=1,\ldots,n}$	covariance matrix of random vector $\boldsymbol{X}$ with $\Sigma_{ij} = \text{Cov}(X_i, X_j) = E((X_i - \mu_i)(X_j - \mu_j))$
Σ^{-1}	precision matrix of random vector $\boldsymbol{X}$
$I(\boldsymbol{\theta}) = (I(\boldsymbol{\theta})_{ij})_{i,j=1,\ldots,n}$	Fisher information matrix with $I(\boldsymbol{\theta})_{ij} = E(\frac{\partial^2 \ln f(\boldsymbol{X}\mid\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j})$

A.2 Distributions

Symbol	Explanation
$X \sim N(\mu, \sigma^2)$	X is normally distributed with mean μ, variance σ^2 and pdf $f(x\mid\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$, $x \in \mathbb{R}$
$X \sim \text{Bernoulli}(\theta)$	X is Bernoulli distributed with success probability $\theta \in (0, 1)$ and pmf $f(x\mid\theta) = \theta^x(1-\theta)^{1-x}$, $x = 0, 1$, $E(X) = \theta$, $\text{Var}(X) = \theta(1-\theta)$
$X \sim \text{Beta}(\alpha, \beta)$	X is Beta distributed with parameters $\alpha > 0$, $\beta > 0$ and pdf $f(x\mid\alpha, \beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$, $x \in (0, 1)$, $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$, $E(X) = \frac{\alpha}{\alpha+\beta}$, $\text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
$X \sim \text{Poisson}(\theta)$	X is Poisson distributed with parameter $\theta > 0$ and pmf $f(x\mid\theta) = \frac{\theta^x}{x!}e^{-x}$, $x \in \{0, 1, 2, \ldots\}$, $E(X) = \text{Var}(X) = \theta$
$X \sim \text{Gamma}(\alpha, \beta)$	X is Gamma distributed with parameters $\alpha > 0$, $\beta > 0$ and pdf $f(x\mid\alpha, \beta) = \frac{1}{\Gamma(\alpha)}\beta^\alpha x^{\alpha-1}e^{-\beta x}$, $x > 0$, $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$, $E(X) = \frac{\alpha}{\beta}$, $\text{Var}(X) = \frac{\alpha}{\beta^2}$
$X \sim N(0, 1)$	X is standard normal with pdf $\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}x^2\}$, and cdf $\Phi(x) = \int_{-\infty}^{x} \varphi(u)du$, $E(X) = 0$, $\text{Var}(X) = 1$
$\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \Sigma)$	$\boldsymbol{X}$ is multivariate normally distributed with mean vector $\boldsymbol{\mu}$, covariance matrix Σ and pdf $f(\boldsymbol{x}\mid\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{n/2}}\lvert\Sigma\rvert^{-1/2} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})' \times \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$, $\boldsymbol{x} \in \mathbb{R}^n$, $E(\boldsymbol{X}) = \boldsymbol{\mu}$, $\text{Var}(\boldsymbol{X}) = \Sigma$

A.3 Classical Statistics

Symbol	Explanation
θ ($\boldsymbol{\theta}$)	unknown fixed parameter to be estimated
$(x_1, \dots, x_n)'$	i.i.d. sample (realizations) from r.v. X
$\hat{\theta}$ ($\hat{\boldsymbol{\theta}}$)	estimate of θ ($\boldsymbol{\theta}$) based on data $\boldsymbol{x} = (x_1, \dots, x_n)$
$\ell(\boldsymbol{\theta} \mid \boldsymbol{x})$	likelihood for $\boldsymbol{\theta}$ based on data $\boldsymbol{x}$ from $X \sim f(\cdot \mid \boldsymbol{\theta})$ given as $\ell(\boldsymbol{\theta} \mid \boldsymbol{x}) = f(\boldsymbol{x} \mid \boldsymbol{\theta})$
$\hat{\boldsymbol{\theta}}_{ML}$	maximum likelihood estimator of $\boldsymbol{\theta}$: maximizes the likelihood $\ell(\boldsymbol{x} \mid \boldsymbol{\theta})$ over $\boldsymbol{\theta}$
$I^{-1}(\boldsymbol{\theta})$	inverse Fisher information matrix, corresponds to asymptotic covariance matrix of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{ML}$
$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$	sample or empirical mean for the i.i.d. sample $(x_1, \dots, x_n)$
$s^2 := \frac{1}{n-1} \times \sum_{i=1}^{n}(x_i - \bar{x})^2$	sample variance for the i.i.d. sample $(x_1, \dots, x_n)$
$Y_i \sim N(x_{i1}\beta_1 + \cdots + x_{id}\beta_d, \sigma^2)$ independent for $i = 1, \dots, d$	linear regression model for response Y_i, covariates $x_{i1}, \dots, x_{id}$ and unknown regression coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)$
$\hat{\boldsymbol{\beta}}_{LS}$	least square estimator of $\boldsymbol{\beta}$, given by minimizing $Q(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - \cdots - x_{id}\beta_d)^2$ for observed responses $y_1, \dots, y_n$
$[l(\boldsymbol{x}), u(\boldsymbol{x})]$	$100(1-\alpha)$ % confidence interval for θ if $P_\theta(l(\boldsymbol{x}) \leq \theta \leq u(\boldsymbol{x})) \geq 1 - \alpha$, that is, the random interval $[l(\boldsymbol{x}), u(\boldsymbol{x})]$ covers the true parameter θ in $100(1-\alpha)$ % of times

A.4 Bayesian Statistics

Symbol	Explanation
θ ($\boldsymbol{\theta}$)	unknown random parameter
$p(\boldsymbol{\theta})$	prior pdf/pmf for $\boldsymbol{\theta}$
$p(\boldsymbol{\theta} \mid \boldsymbol{x})$	posterior pdf/pmf of $\boldsymbol{\theta}$ given the observed sample $\boldsymbol{x}$ from $\boldsymbol{X} \sim f(\cdot \mid \boldsymbol{\theta})$ Bayes' theorem: $p(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{\ell(\boldsymbol{\theta} \mid \boldsymbol{x})p(\boldsymbol{\theta})}{\int_{-\infty}^{\infty}\ell(\boldsymbol{\theta} \mid \boldsymbol{x})p(\boldsymbol{\theta})d\boldsymbol{\theta}}$
$\boldsymbol{\theta}_{\text{mode}}(\boldsymbol{x})$	posterior mode = mode of posterior distribution
$\boldsymbol{\theta}_{\text{mean}}(\boldsymbol{x})$	posterior mean = mean of posterior distribution
$I(\boldsymbol{x})$	$100(1-\alpha)$ % credible interval for θ if $\int_{I(\boldsymbol{x})} p(\theta \mid \boldsymbol{x})d\theta = 1 - \alpha$
$\theta_\alpha(\boldsymbol{x})$ ($\hat{\theta}_\alpha(\boldsymbol{x})$)	(empirical) α-quantile of posterior distribution
$[\hat{\theta}_{\alpha/2}(\boldsymbol{x}), \hat{\theta}_{1-\alpha/2}(\boldsymbol{x})]$	$100(1-\alpha)$ % credible interval based on empirical quantiles
$f(y \mid \boldsymbol{x})$	predictive density of future observation y given the observations $\boldsymbol{x}$: $f(y \mid \boldsymbol{x}) = \int f(y \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \boldsymbol{x})d\boldsymbol{\theta}$ if Y is independent of $\boldsymbol{X}$ given $\boldsymbol{\theta}$

A.5 MCMC Methods

Symbol	Explanation
$\{\boldsymbol{\theta}^{(t)} : t \in T\}$	stochastic process with random vectors $\boldsymbol{\theta}^{(t)}$ taking values in the state space S for each t out of the index set T
$\{\boldsymbol{\theta}^{(n)} : n = 1, 2, \ldots\}$	Markov chain (MC) if (2.3) holds
$\boldsymbol{\theta}^{(n)}$ homogeneous	if (2.3) does not depend on n
$P(\boldsymbol{x}, \boldsymbol{y}) := P(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{y} \mid \boldsymbol{\theta}^{(n)} = \boldsymbol{x})$	transition probability of homogeneous MC $\boldsymbol{\theta}^{(n)}$ with discrete state space S
$P = (P(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j=1,\ldots,r}$	transition matrix for a homogeneous MC $\boldsymbol{\theta}^{(n)}$ with finite state space $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r\}$: $P(\boldsymbol{x}_i, \boldsymbol{x}_j) = P(\boldsymbol{\theta}^{(n+1)} = \boldsymbol{x}_j \mid \boldsymbol{\theta}^{(n)} = \boldsymbol{x}_i)$
$P^m(\boldsymbol{x}, \boldsymbol{y}) := P(\boldsymbol{\theta}^{(n+m)} = \boldsymbol{y} \mid \boldsymbol{\theta}^{(n)} = \boldsymbol{x})$	mth order transition probability for $m > n$
$\pi^{(0)}(\boldsymbol{x}) = P(\boldsymbol{\theta}^{(0)} = \boldsymbol{x})$	initial distribution of MC $\boldsymbol{\theta}^{(n)}$
$\pi^{(n)}(\boldsymbol{x}) = P(\boldsymbol{\theta}^{(n)} = \boldsymbol{x})$	nth step marginal distribution of MC $\boldsymbol{\theta}^{(n)}$
π stationary	if (2.7) holds
$T_{\boldsymbol{y}}$	first visit of MC $\boldsymbol{\theta}^{(n)}$ to $\boldsymbol{y}$
$\rho_{\boldsymbol{x}\boldsymbol{y}}$	probability of visiting $\boldsymbol{y}$ after starting in $\boldsymbol{x}$
$\boldsymbol{y} \in S$ (positive) recurrent	$\rho_{\boldsymbol{y}\boldsymbol{y}} = 1$ ($\rho_{\boldsymbol{y}\boldsymbol{y}} = 1$ and $E(T_{\boldsymbol{y}} \mid \boldsymbol{\theta}^{(0)} = \boldsymbol{y}) < \infty$)
$\boldsymbol{\theta}^{(n)}$ irreducible	$\rho_{\boldsymbol{x}\boldsymbol{y}} > 0$, $\rho_{\boldsymbol{y}\boldsymbol{x}} > 0$ $\forall \boldsymbol{x}, \boldsymbol{y} \in S$
$\boldsymbol{\theta}^{(n)}$ aperiodic	if largest common divisor of $\{n \geq 1 : P^n(\boldsymbol{x}, \boldsymbol{x}) > 0\} = 1$ $\forall \boldsymbol{x} \in S$
$\boldsymbol{\theta}^{(n)}$ ergodic	if $\boldsymbol{\theta}^{(n)}$ aperiodic and irreducible
Full conditionals of random parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)'$	conditional distributions of θ_i, $i = 1, \ldots, d$, given all other components different from i
Autocorrelation of lag k	correlation $Cor(\boldsymbol{\theta}^{(n)}, \boldsymbol{\theta}^{(n+k)})$ in homogeneous MC $\boldsymbol{\theta}^{(n)}$

References

Selected Bibliography

1. W.M. Bolstad, *Introduction to Bayesian Statistics* (Wiley, Hoboken, 2004)
2. G. Casella, E.I. George, Explaining the Gibbs sampler. Am. Stat. **46**, 167–174 (1992)
3. S. Chib, E. Greenberg, Understanding the Metropolis-Hastings algorithm. Am. Stat. **49**, 327–335 (1995)
4. D. Gamerman, H.F. Lopes, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference* (Taylor & Francis, Boca Raton, 2006)
5. P.M. Lee, *Bayesian Statistics: An Introduction*, 4th edn. (Wiley, Hoboken, 2012)
6. D. Lunn, C. Jackson, N. Best, A. Thomas, D. Spiegelhalter, *The BUGS Book—A Practical Introduction to Bayesian Analysis* (Chapman & Hall/CRC, London, 2012)
7. I. Ntzoufras, *Bayesian Modeling Using WinBUGS* (Wiley, Hoboken, 2009)

Additional Literature

8. J.O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd edn. (Springer, Berlin, 1985)
9. S.P. Brooks, G.O. Roberts, Assessing convergence of Markov chain Monte Carlo algorithms. Stat. Comput. **8**, 319–335 (1998)
10. B.P. Carlin, A.E. Gelfand, A.F.M. Smith, Hierarchical Bayesian analysis of changepoint problems. Appl. Stat. **41**, 389–405 (1992)
11. M.K. Cowles, B.P. Carlin, Markov chain Monte Carlo convergence diagnostics: a comparative review. J. Am. Stat. Assoc. **91**, 883–904 (1996)
12. R. Durrett, *Probability: Theory and Examples*, 4th edn. (Cambridge University Press, Cambridge, 2010)
13. S. Frühwirth-Schnatter, L. Sögner, Bayesian estimation of stochastic volatility models based on OU processes with marginal Gamma law. Ann. Inst. Stat. Math. **61**(1), 159–179 (2009)
14. A.E. Gelfand, A.F.M. Smith, Sampling-based approaches to calculating marginal densities. J. Am. Stat. Assoc. **85**, 398–409 (1990)
15. A. Gelman, J.B. Carlin, H.S. Stern, D.B.R. Rubin, *Bayesian Data Analysis*, 2nd edn. (Chapman & Hall, London, 2003)
16. S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721–741 (1984)
17. W.R. Gilks, S. Richardson, D.J. Spiegelhalter, *Markov Chain Monte Carlo in Practice* (Chapman & Hall, London, 1996)
18. P. Guttorp, *Stochastic Modeling of Scientific Data* (Chapman & Hall/CRC, London, 1995)
19. W.K. Hastings, Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57**, 97–109 (1970)
20. P.D. Hoff, *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics (Springer, New York, 2009)
21. A. Jeffreys, *The Theory of Probability* (Cambridge University Press, Cambridge, 1961)
22. J.-M. Marin, C.P. Robert, *Bayesian Core: A Practical Approach to Computational Bayesian Statistics* (Springer, New York, 2007)
23. N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1092 (1953)
24. S.P. Meyn, R.L. Tweedie, *Markov Chains and Stochastic Stability*, 2nd edn. (Cambridge University Press, Cambridge, 2009)
25. E. Nummelin, *General Irreducible Markov Chains and Non-negative Operators* (Cambridge University Press, Cambridge, 1984)
26. S.I. Resnick, *Adventures in Stochastic Processes* (Birkhäuser, Boston, 1992)
27. G.O. Roberts, A.F.M. Smith, Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. Stoch. Process. Appl. **49**, 207–216 (1994)
28. G.O. Roberts, O. Papaspiliopoulos, P. Dellaportas, Bayesian inference for non-Gaussian Ornstein-Uhlenbeck stochastic volatility processes. J. R. Stat. Soc., Ser. B **66**, 369–393 (2004)
29. D.P.M. Scollnik, Actuarial modeling with MCMC and BUGS. N. Am. Actuar. J. **5**, 96–124 (2001)