

## Chapter 2

# Loss Distribution Approach

*Out of intense complexities intense simplicities emerge.*

Sir Winston Churchill

**Abstract** This chapter introduces a basic model for the Loss Distribution Approach. We discuss the main aspects of the model and basic probabilistic concepts of risk quantification. The essentials of the frequentist and Bayesian statistical approaches are introduced. Basic Markov chain Monte Carlo methods that allow sampling from the posterior distribution, when the sampling cannot be done directly, are also described.

### 2.1 Loss Distribution Model

A popular method under the AMA is the loss distribution approach (LDA). Under the LDA, banks quantify distributions for frequency and severity of operational risk losses for each risk cell (business line/event type) over a 1-year time horizon. The banks can use their own risk cell structure but must be able to map the losses to the Basel II risk cells. Various quantitative aspects of LDA modelling are discussed in King [134]; Cruz [65, 66]; McNeil, Frey and Embrechts [157]; Panjer [181]; Chernobai, Rachev and Fabozzi [55]; Shevchenko [216]. The commonly used LDA model for the total annual loss  $Z_t$  in a bank can be formulated as

$$Z_t = \sum_{j=1}^J Z_t^{(j)}; \quad Z_t^{(j)} = \sum_{i=1}^{N_t^{(j)}} X_i^{(j)}(t). \quad (2.1)$$

Here:

- $t = 1, 2, \dots$  is discrete time in annual units. If shorter time steps are used (e.g. quarterly steps to calibrate dependence structure between the risks), then extra summation over these steps can easily be added in (2.1).
- The annual loss  $Z_t^{(j)}$  in risk cell  $j$  is modelled as a compound (*aggregate*) loss over one year with the *frequency* (annual number of events)  $N_t^{(j)}$  implied by a counting process (e.g. Poisson process) and *severities*  $X_i^{(j)}(t)$ ,  $i = 1, \dots, N_t^{(j)}$ .
- Typically, the frequencies and severities are modelled by independent random variables.

Estimation of the annual loss distribution by modelling frequency and severity of losses is a well-known actuarial technique; see for example Klugman, Panjer and Willmot [136]. It is also used to model solvency requirements for the insurance industry; see Sandström [207] and Wüthrich and Merz [240]. Under model (2.1), the capital is defined as the 0.999 Value-at-Risk (VaR) which is the quantile of the distribution for the next year annual loss  $Z_{T+1}$ :

$$\text{VaR}_q[Z_{T+1}] = \inf\{z \in \mathbb{R} : \Pr[Z_{T+1} > z] \leq 1 - q\} \quad (2.2)$$

at the level  $q = 0.999$ . Here, index  $T + 1$  refers to the next year. The capital can be calculated as the difference between the 0.999 VaR and the expected loss if the bank can demonstrate that the expected loss is adequately captured through other provisions. If assumptions on correlations between some groups of risks (e.g. between business lines or between risk cells) cannot be validated then the capital should be calculated as the sum of the 0.999 VaRs over these groups. This is equivalent to the assumption of perfect positive dependence between annual losses of these groups.

Of course, instead of modelling frequency and severity to obtain the annual loss distribution, one can model aggregate loss per shorter time period (e.g. monthly total loss) and calculate the annual loss as a sum of these aggregate losses. However, the frequency/severity approach is more flexible and has good advantages, because some factors may affect frequency only while other factors may affect severity only. For example:

- As the business grows (e.g. volume of the transactions grows), the expected number of losses changes and this should be accounted for in forecasting the number of losses (frequency) over the next year.
- The general economic inflation affects the loss sizes (severity).
- The insurance for operational risk losses is more easily incorporated. This is because, typically, the insurance policies apply per event and affect the severity.

In this book, we focus on some statistical methods proposed in the literature for the LDA model (2.1). In particular we consider the problem of combining different data sources, modelling dependence and large losses, and accounting for parameter uncertainty.

## 2.2 Operational Risk Data

Basel II specifies the data that should be collected and used for AMA. In brief, a bank should have internal data, external data and expert opinion data. In addition, internal control indicators and factors affecting the businesses should be used. Development and maintenance of operational risk databases is a difficult and challenging task. Some of the main features of the required data are summarised as follows.

- *Internal data.* Internal data should be collected over a minimum five-year period to be used for capital charge calculations (when the bank starts the AMA, a three-year period is acceptable). Due to a short observation period, typically the internal data for many risk cells contain few low-frequency/high-severity losses or none. A bank must be able to map its historical internal loss data into the relevant Basel II risk cells; see Tables 1.1, 1.2 and 1.3. The data must capture all material activities and exposures from all appropriate sub-systems and geographic locations. A bank can have an appropriate low reporting threshold for internal loss data collection, typically of the order of EURO 10,000. Aside from information on gross loss amounts, a bank should collect information about the date of the event, any recoveries of gross loss amounts, as well as some descriptive information about the drivers or causes of the loss event.
- *External data.* A bank's operational risk measurement system must use relevant external data (either public data and/or pooled industry data). These external data should include data on actual loss amounts, information on the scale of business operations where the event occurred, and information on the causes and circumstances of the loss events. Industry data are available through external databases from vendors (e.g. Algo OpData provides publicly reported operational risk losses above USD 1million) and consortia of banks (e.g. ORX provides operational risk losses above EURO 20,000 reported by ORX members). External data are difficult to use directly due to different volumes and other factors. Moreover, the data have a survival bias as typically the data of all collapsed companies are not available. As discussed previously in Sect. 1.4, several Loss Data Collection Exercises (LDCE) for historical operational risk losses over many institutions were conducted and their analyses reported in the literature. In this respect, two papers are of high importance: Moscadelli [166] analysing 2002 LDCE and Dutta and Perry [77] analysing 2004 LDCE. In each case the data were mainly above EURO 10,000 and USD 10,000 respectively.
- *Scenario Analysis/expert opinion.* A bank must use scenario analysis in conjunction with external data to evaluate its exposure to high-severity events. Scenario analysis is a process undertaken by experienced business managers and risk management experts to identify risks, analyse past internal/external events, consider current and planned controls in the banks, etc. It may involve: workshops to identify weaknesses, strengths and other factors; opinions on the severity and frequency of losses; opinions on sample characteristics or distribution parameters of the potential losses. As a result some rough quantitative assessment of the risk frequency and severity distributions can be obtained. Scenario analysis is very subjective and should be combined with the actual loss data. In addition, it should be used for stress testing, for example to assess the impact of potential losses arising from multiple simultaneous loss events.
- *Business environment and internal control factors.* A bank's methodology must capture key business environment and internal control factors affecting operational risk. These factors should help to make forward-looking estimates, account for the quality of the controls and operating environments, and align capital assessments with risk management objectives.

Data important for modelling but often missing in external databases are risk exposure indicators and near-misses.

- *Exposure indicators.* The frequency and severity of operational risk events are influenced by indicators such as gross income, number of transactions, number of staff and asset values. For example, frequency of losses typically increases with increasing number of employees.
- *Near-miss losses.* These are losses that could occur but were prevented. Often these losses are included in internal datasets to estimate severity of losses but excluded in the estimation of frequency. For detailed discussion on management of near-misses, see Muermann and Oktem [167].

### 2.3 A Note on Data Sufficiency

Empirical estimation of the annual loss 0.999 quantile, using observed losses only, is impossible in practice. It is instructive to calculate the number of data points needed to estimate the 0.999 quantile empirically within the desired accuracy. Assume that independent data points  $X_1, \dots, X_n$  with common density  $f(x)$  have been observed. Then the quantile  $q_\alpha$  at confidence level  $\alpha$  is estimated empirically as  $\hat{Q}_\alpha = \tilde{X}_{\lfloor n\alpha \rfloor + 1}$ , where  $\tilde{\mathbf{X}}$  is the data sample  $\mathbf{X}$  sorted into the ascending order. The standard deviation of this empirical estimate is

$$\text{stdev}[\hat{Q}_\alpha] = \frac{\sqrt{\alpha(1-\alpha)}}{f(q_\alpha)\sqrt{n}}; \quad (2.3)$$

see Glasserman ([108], section 9.1.2, p. 490). Thus, to calculate the quantile within relative error  $\varepsilon = 2 \times \text{stdev}[\hat{Q}_\alpha]/q_\alpha$ , we need

$$n = \frac{4\alpha(1-\alpha)}{\varepsilon^2(f(q_\alpha)q_\alpha)^2} \quad (2.4)$$

observations. Suppose that the data are from the lognormal distribution  $\mathcal{LN}(\mu = 0, \sigma = 2)$ . Then using formula (2.4), we obtain that  $n = 140,986$  observations are required to achieve 10% accuracy ( $\varepsilon = 0.1$ ) in the 0.999 quantile estimate. In the case of  $n = 1,000$  data points, we get  $\varepsilon = 1.18$ , that is, the uncertainty is larger than the quantile we estimate.

Moreover, according to the regulatory requirements, the 0.999 quantile of the annual loss (rather than 0.999 quantile of the severity) should be estimated. As will be discussed many times in this book, operational risk losses are typically modelled by the so-called heavy-tailed distributions. In this case, the quantile at level  $q$  of the aggregate distributions can be approximated by the quantile of the severity distribution at level

$$p = 1 - \frac{1-q}{E[N]};$$

see Sect. 6.7. Here,  $E[N]$  is the expected annual number of events. For example, if  $E[N] = 10$ , then we obtain that the error of the annual loss 0.999 quantile is the same as the error of the severity quantile at the confidence level  $p = 0.9999$ . Again, using (2.4) we conclude that this would require  $n \approx 10^6$  observed losses to achieve 10% accuracy. If we collect annual losses then  $n/E[N] \approx 10^5$  annual losses should be collected to achieve the same accuracy of 10%. These amounts of data are not available even from the largest external databases and extrapolation well beyond the data is needed. Thus parametric models must be used.

For an excellent discussion on data sufficiency in operational risk, see Cope, Antonini, Mignola and Ugoccioni [62].

## 2.4 Insurance

Some operational risks can be insured. If a loss occurs and it is covered by an insurance policy, then part of the loss will be recovered. Under the AMA, banks are allowed to recognise the risk mitigating impact of insurance on the regulatory capital charge. The reduction in the capital due to insurance is limited to 20%; see BCBS ([17], p. 155).

A typical policy will provide a recovery  $R$  for a loss  $X$  subject to the excess amount (deductible)  $D$  and top cover limit amount  $U$  as follows:

$$R = \begin{cases} 0, & \text{if } 0 \leq X < D, \\ X - D, & \text{if } D \leq X < U + D, \\ U, & \text{if } D + U \leq X. \end{cases} \quad (2.5)$$

That is, the recovery will take place if the loss is larger than the excess and the maximum recovery that can be obtained from the policy is  $U$ . Note that in (2.5), the time of the event is not involved and the top cover limit applies for a recovery per risk event, that is, for each event the obtained recovery is subject of the top cover limit. Including insurance into the LDA is simple; the loss severity in (2.1) should be reduced by the amount of recovery (2.5) and can be viewed as a simple transformation of the severity. However, there are several difficulties in practice, namely that

- policies may cover several different risks;
- different policies may cover the same risk;
- the top cover limit may apply for the aggregated recovery over many events of one or several risks (e.g. the policy will pay the recovery for losses until the top cover limit is reached by accumulated recovery).

These aspects and special restrictions on insurance recoveries required by Basel II make recovery dependent on time. Thus accurate accounting for insurance requires modelling the loss event times. For example, one can use a Poisson process to model the event times.

*Remark 2.1* A convenient method to simulate event times from a Poisson process over a one-year time horizon is to simulate the annual number of events  $N$  from the Poisson distribution and then simulate the times of these  $N$  events as independent random variables from a uniform distribution  $\mathcal{U}(0, 1)$ .

It is not difficult to incorporate the insurance into an overall model if a Monte Carlo method<sup>1</sup> is used to quantify the annual loss distributions. The inclusion of the insurance will certainly reduce the capital charge, though the reduction is capped by 20% according to the Basel II requirement.

Finally, it is important to note that, incorporating insurance into the LDA is not only important for capital reduction but also beneficial for negotiating a fair premium with the insurer because the distribution of the recoveries and its characteristics can be estimated.

For implementation of insurance into the LDA, see Bazzarello, Crielaard, Piacenza and Soprano [22], Peters, Byrnes and Shevchenko [184]; also for guidelines on insurance within the AMA capital calculations, see Committee of European Banking Supervisors [59].

## 2.5 Basic Statistical Concepts

A concept of financial risk strongly relates to a notion of events that may occur and lead to financial consequences. Thus it is natural to model risks using probability theory. While a notion of randomness is very intuitive, it was only in 1933 that Kolmogorov [138] gave an axiomatic definition of randomness and probability. This theory gives a mathematical foundation to modern risk modelling. It is expected that the reader has a basic understanding of elementary statistics and probability. This section provides a description of essential concepts of probability theory used in the book and introduces relevant notation.

### 2.5.1 Random Variables and Distribution Functions

Hereafter, the following notation is used:

- Random variables are denoted by upper case symbols (capital letters) and their realisations are denoted by lower case symbols, e.g. random variable  $X$  and its realisation  $x$ .
- By convention, vectors are considered as column vectors and are written in bold, e.g.  $n$ -dimensional random vector  $\mathbf{X} = (X_1, X_2, \dots, X_n)'$ , where superscript ' $'$ ' denotes transposition.
- The realisations of random variables considered in this book are real numbers, so that  $\mathbf{x} = (x_1, x_2, \dots, x_n)'$  means a point in the  $n$ -dimensional Euclidean space of real numbers  $\mathbb{R}^n$ .

---

<sup>1</sup> Monte Carlo method is discussed in Sect. 3.2.

- To simplify notation, in general, the same symbol will be used to denote both a random variable and the space of its possible realisations. For example:  $\Theta$  is a random variable;  $\theta$  is realisation of  $\Theta$ ; and the space of all possible  $\theta$  values is also denoted as  $\Theta$ .
- Operators on random variables are written with square brackets, e.g. the variance of a random variable  $X$  is denoted as  $\text{Var}[X]$ .
- Notationally, an *estimator* is a function of the sample while an *estimate* is the realised value of an estimator for a given realisation of the sample. For example, given a sample of random variables  $X_1, X_2, \dots, X_n$  the estimator is a function of  $\mathbf{X}$  while the estimate is a function of the realisation  $\mathbf{x}$ .

A random variable has associated distribution function defined as follows.

**Definition 2.1 (Univariate distribution function)** The distribution function of a random variable  $X$ , denoted as  $F_X(x)$ , is defined as

$$F_X(x) = \Pr[X \leq x].$$

A corresponding *survival function (tail function)* is defined as

$$\bar{F}_X(x) = 1 - F_X(x) = \Pr[X > x].$$

**Definition 2.2 (Multivariate distribution function)** The multivariate distribution function of a random vector  $\mathbf{X} = (X_1, X_2, \dots, X_n)'$  is defined as

$$F_{\mathbf{X}}(x_1, x_2, \dots, x_n) = \Pr[X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n].$$

Often, for short notation we write  $F_{\mathbf{X}}(\mathbf{x})$ . A corresponding survival function is defined as

$$\bar{F}_{\mathbf{X}}(\mathbf{x}) = \Pr[\mathbf{X} > \mathbf{x}].$$

#### Remark 2.2

- Frequently used notation,  $X \sim F_X(x)$ , means a random variable  $X$  has a distribution function  $F_X(x)$ . Often, for simplicity of notation, we may drop the subscript and write  $X \sim F(\cdot)$ .
- All distributions used throughout the book are formally defined in Appendix A.

Random variables can be classified into different categories (*continuous, discrete or mixed*) according to their *support* (a set of all possible outcomes of a random variable). Precisely:

**Definition 2.3 (Support of a random variable)** The support of a random variable  $X$  with a distribution function  $F_X(\cdot)$  is defined as a set of all points, where  $F_X(\cdot)$  is strictly increasing.

**Definition 2.4 (Continuous random variable)** A continuous random variable  $X$  has its support on an interval, a union of intervals or real line (half-line). The distribution function of a continuous random variable can be written as

$$F_X(x) = \int_{-\infty}^x f_X(y)dy,$$

where  $f_X(x)$  is called the continuous *probability density function*.

**Definition 2.5 (Discrete random variable)** A discrete random variable  $X$  has a finite or countable number of values  $x_1, x_2, \dots$ . The distribution function of a discrete random variable has jump discontinuities at  $x_1, x_2, \dots$  and is constant between. The probability function (also called the *probability mass function*) of a discrete random variable is defined as

$$\begin{aligned} p_X(x_i) &= \Pr[X = x_i], \quad i = 1, 2, \dots \\ p_X(x) &= 0 \quad \text{for } x \neq x_1, x_2, \dots \end{aligned}$$

The corresponding probability density function can be written as

$$f_X(x) = \sum_{i \geq 1} p_X(x_i) \delta(x - x_i), \quad (2.6)$$

where  $\delta(x)$  is the *Dirac  $\delta$ -function* (also called the impulse  $\delta$ -function) defined next.

**Definition 2.6 (The Dirac  $\delta$ -function)** The Dirac  $\delta$ -function is a function which is zero everywhere except from the origin where it is infinite and its integral over any arbitrary interval containing the origin is equal to one:

$$\begin{aligned} \delta(x) &= 0 \text{ if } x \neq 0; \quad \delta(0) = \infty, \\ \int_{-\epsilon}^{\epsilon} \delta(x)dx &= 1 \text{ for any } \epsilon > 0. \end{aligned}$$

Note that, this implies that for any function  $g(x)$

$$\int_a^b g(x) \delta(x - x_0) dx = g(x_0) \quad \text{if } a < x_0 < b \quad (2.7)$$

and the integral is zero if  $(a, b)$  interval does not contain  $x_0$ . This definition of  $\delta$  function is merely a heuristic definition but it is enough for the purposes of this book. The use and theory of the Dirac  $\delta$ -function can be found in many books; see for example Pugachev ([196], section 9).

**Definition 2.7 (Mixed random variable)** Mixed random variable  $X$  is a continuous random variable with positive probability of occurrence on a countable set of exception points. Its distribution function  $F_X$  has jumps at these exception points and can be written as

$$F_X(x) = w F_X^{(d)}(x) + (1 - w) F_X^{(c)}(x)$$



where  $0 \leq w \leq 1$ ,  $F_X^{(c)}$  is a continuous distribution function and  $F_X^{(d)}(x)$  is a discrete distribution function. The corresponding density function can be written as

$$f_X(x) = w \sum_{i \geq 1} p_X(x_i) \delta(x - x_i) + (1 - w) f_X^{(c)}(x), \quad (2.8)$$

where  $f_X^{(c)}(x)$  is the continuous density function and  $p_X(x_i)$  is a probability mass function of a discrete distribution.

*Remark 2.3*

- A mixed random variable is common in modelling financial risk and in operational risk in particular, when there is a probability of non-occurrence loss during a period of time (giving finite probability mass at zero) while the loss amount is a continuous random variable.
- In general, every distribution function may be represented as a mixture of three different types: discrete distribution function, continuous distribution function and singular continuous distribution function. The last is a continuous distribution function with points of increase on a set of zero Lebesgue measure. This type of random variable will not be considered in the book. The case of mixed random variables with two components (discrete and continuous) covers all situations encountered in operational risk practice.

### 2.5.2 Quantiles and Moments

We use the following standard definition of a generalised inverse function (also called *quantile function*) for a distribution function.

**Definition 2.8 (Quantile function)** Given a distribution function  $F_X(x)$ , the inverse function  $F_X^{-1}$  of  $F_X$  is

$$F_X^{-1}(\alpha) = \inf\{x \in \mathbb{R} : F_X(x) \geq \alpha\} = \sup\{x \in \mathbb{R} : F_X(x) < \alpha\},$$

where  $0 < \alpha < 1$ .

Given a probability level  $\alpha$ ,  $F_X^{-1}(\alpha)$  is the  $\alpha$ -th quantile of  $X$  (often, it is denoted as  $q_\alpha$ ). This generalised definition is needed to define a quantile for cases such as discrete and mixed random variables. If  $F_X$  is continuous, then the quantile function is the ordinary inverse function.

The expected value (*mean*) of a random variable  $X$  is denoted as  $E[X]$ . A formal construction of the operator  $E[\cdot]$  is somewhat involved but for the purposes of this book we will use the following short definition.

**Definition 2.9 (Expected value)**

- If  $X$  is a continuous random variable with the density function  $f_X(x)$ , then

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx; \quad (2.9)$$

- If  $X$  is a discrete random variable with support  $x_1, x_2, \dots$  and probability mass function  $p_X(x)$ , then

$$E[X] = \sum_{j \geq 1} x_j p_X(x_j);$$

- In the case of a mixed random variable  $X$  (see Definition 2.7), the expected value is

$$E[X] = w \sum_{j \geq 1} x_j p_X(x_j) + (1 - w) \int_{-\infty}^{\infty} x f_X^{(c)}(x) dx.$$

*Remark 2.4*

- The expected value integral or sum may not converge to a finite value for some distributions. In this case it is said that the mean does not exist.
- The definition of the expected value (2.9) can also be used in the case of the discrete and mixed random variables if their density functions are defined as (2.6) and (2.8) respectively. This gives a unified notation for the expected value of the continuous, discrete and mixed random variables. Another way to introduce a unified notation is to use Riemann-Stieltjes integral

$$E[X] = \int_{-\infty}^{\infty} x dF_X(x). \quad (2.10)$$

See Carter and Van Brunt [48] for a good introduction on this topic.

The expected value is the first moment about the origin (also called the first raw moment). There are two standard types of moments: the raw moments and central moments, defined as follows.

**Definition 2.10 (Moments)**

- The  $k$ -th moment about the origin (raw moment) of a random variable  $X$  is the expected value of  $X^k$ , i.e.  $E[X^k]$ .
- The  $k$ -th central moment of a random variable  $X$  is the expected value of  $(X - E[X])^k$ , i.e.  $E[(X - E[X])^k]$ .

Typically,  $k$  is nonnegative integer  $k = 0, 1, 2, \dots$ . The expected value may not exist for some values of  $k$ ; then it is said that the  $k$ -th moment does not exist. The first four moments are most frequently used and the relevant characteristics are:

- *Variance* – The variance of a random variable  $X$  is the second central moment

$$\text{Var}[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2. \quad (2.11)$$

- *Standard deviation* – The standard deviation,

$$\text{stdev}[X] = \sqrt{\text{Var}[X]}, \quad (2.12)$$

is a measure of spread of the random variable around the mean. It is measured in the same units as the mean (i.e. the same units as the values of random variable).

- *Variational coefficient* – The *variational coefficient* (also called the *coefficient of variation*) is dimensionless quantity,

$$\text{Vco}[X] = \frac{\text{stdev}[X]}{E[X]}, \quad (2.13)$$

that measures the spread relative to the mean.

- *Skewness* – The skewness is a dimensionless quantity that measures an asymmetry of a random variable  $X$  and is defined as

$$\gamma_1 = \frac{E[(X - E[X])^3]}{(\text{stdev}[X])^3}. \quad (2.14)$$

For symmetric distributions, the skewness is zero.

- *Kurtosis* – The kurtosis is a dimensionless quantity that measures flatness of distribution relative to the normal distribution. It is defined as

$$\gamma_2 = \frac{E[(X - E[X])^4]}{(\text{stdev}[X])^4} - 3. \quad (2.15)$$

For the normal distribution, kurtosis is zero.

Again, for some distributions the above characteristics may not exist. Also, central moments can be expressed through the raw moments and vice-versa. Detailed discussion, definition and relationships for the above quantities can be found in virtually any statistical textbook. To conclude this section, we define the covariance and the linear correlation coefficient that measure the dependence between random variables.

**Definition 2.11 (Covariance and linear correlation)** The covariance of random variables  $X$  and  $Y$  is defined as

$$\text{Cov}[X, Y] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y].$$

The linear correlation between  $X$  and  $Y$  is

$$\rho[X, Y] = \text{Cov}[X, Y] / \sqrt{\text{Var}[X]\text{Var}[Y]}.$$

These quantities are popular measures of the dependence between  $X$  and  $Y$  but, as will be discussed in Chap. 7, the linear correlation can be a bad indicator of dependence. Also, for some distributions these measures may not exist.

## 2.6 Risk Measures

Using economic reasoning, a list of axiomatic properties for a good (*coherent*) risk measure was suggested in the seminal paper by Artzner, Delbaen, Eber and Heath [10].

**Definition 2.12 (A coherent risk measure)** A coherent risk measure,  $\varrho[X]$ , is defined to have the following properties for any two random variables  $X$  and  $Y$ :

- Subadditivity:  $\varrho[X + Y] \leq \varrho[X] + \varrho[Y]$ ;
- Monotonicity: if  $X \leq Y$  for all possible outcomes, then  $\varrho[X] \leq \varrho[Y]$ ;
- Positive homogeneity: for any positive constant  $c$ ,  $\varrho[cX] = c\varrho[X]$ ;
- Translation invariance: for any positive constant  $c$ ,  $\varrho[X + c] = \varrho[X] + c$ .

For detailed discussions of this topic, see McNeil, Frey and Embrechts [157]. Two popular risk measures are the so-called *Value-at-Risk* (VaR) and *expected shortfall* defined and discussed below.

**Definition 2.13 (Value-at-Risk)** The VaR of a random variable  $X \sim F_X(x)$  at the  $\alpha$ -th probability level,  $\text{VaR}_\alpha(X)$ , is defined as the  $\alpha$ -th quantile of the distribution of  $X$ , i.e.

$$\text{VaR}_\alpha[X] = F_X^{-1}(\alpha).$$

*Remark 2.5* VaR is not a coherent measure. In general, VaR possesses all the properties of a coherent risk measure in Definition 2.12 except subadditivity. For some cases, such as a multivariate normal distribution, VaR is subadditive. However, in general, the VaR of a sum may be larger than the sum of VaRs. For examples and discussions, see McNeil, Frey and Embrechts [157]. This has a direct implication for measuring operational risk and will be discussed in Chap. 7.

A VaR at a specified probability level  $\alpha$  does not provide any information about the fatness of the distribution upper tail. Often the management and regulators are concerned not only with probability of default but also with its severity. Therefore, other risk measures are often used. One of the most popular is *expected shortfall* (sometimes referred to as the tail Value-at-Risk), though, a formal Basel II regulatory requirement for operational risk capital charge refers to a VaR.

**Definition 2.14 (Expected shortfall)** The expected shortfall of a random variable  $X \sim F_X(x)$  at the  $\alpha$ -th probability level,  $\text{ES}_\alpha[X]$ , is

$$\text{ES}_\alpha[X] = \frac{1}{1 - \alpha} \int_\alpha^1 \text{VaR}_p[X] dp,$$

which is the “arithmetic average” of the VaRs of  $X$  from  $\alpha$  to 1.

*Remark 2.6* Expected shortfall is a coherent risk measure.

In the case of continuous distributions, it can be shown that  $ES_\alpha[X]$  is just expected loss given that the loss exceeds  $VaR_\alpha[X]$ .

**Proposition 2.1** *For a random variable  $X$  with a continuous distribution function  $F_X(x)$  we have*

$$ES_\alpha[X] = E[X|X \geq VaR_\alpha[X]],$$

*which is the conditional expected loss given that the loss exceeds  $VaR_\alpha[X]$ .*

*Proof* Using Definition 2.14, the proof is trivial: simply change the integration variable to  $x = F_X^{-1}(p)$ .  $\square$

*Remark 2.7* For a discontinuous distribution function  $F_X(x)$ , we have more general relation expression

$$ES_\alpha[X] = E[X|X \geq VaR_\alpha[X]] + \left( \frac{1}{1-\alpha} - \frac{1}{\overline{F}_X(VaR_\alpha[X])} \right) \times E[\max(X - VaR_\alpha[X], 0)]. \quad (2.16)$$

The quantity in brackets can be nonzero for some values of  $\alpha$ , where there are jumps in distribution function. For a proof, see Proposition 3.2 in Acerbi and Tasche [4].

## 2.7 Capital Allocation

After the total capital is measured by  $\varrho[\cdot]$ , it is important to answer the question on how much a risk cell  $j$  contributes to the total capital. Calculation of the bank overall capital  $\varrho[Z]$ , where

$$Z = \sum_{j=1}^J Z^{(j)}$$

is the annual loss in a bank over next year as defined by (2.1),<sup>2</sup> should be followed by an important procedure of allocation of the capital into risk cells in such a way that

$$\varrho[Z] = \sum_j^J AC_j. \quad (2.17)$$

---

<sup>2</sup> Here, for simplicity we drop the subscript indicating a year.

Here,  $AC_j$  denotes the capital allocated to the  $j$ -th risk cell. It can be used for performance measurement providing incentives for a business to improve its risk management practices. Naive choice  $AC_j = \varrho[Z^{(j)}]$  is certainly not appropriate because it disregards risk diversification. Also, the sum of  $\varrho[Z^{(j)}]$  adds up to  $\varrho[Z]$  only in the case of perfect positive dependence between risk cells.

Two popular methods, the *Euler principle* and *marginal contribution*, to allocate the capital are described below.

### 2.7.1 Euler Allocation

If risk measure  $\varrho$  is a positive homogeneous function (i.e.  $\varrho[hX] = h\varrho[X]$ ,  $h > 0$ ) and differentiable, then by the *Euler principle*

$$\varrho[Z] = \sum_{j=1}^J \varrho_j^{Euler}, \quad (2.18)$$

where

$$\varrho_j^{Euler} = \left. \frac{\partial \varrho[Z + hZ^{(j)}]}{\partial h} \right|_{h=0}. \quad (2.19)$$

For a proof, see Problem 2.4. The Euler principle is used by many practitioners to calculate the allocated capitals as

$$AC_j = \varrho_j^{Euler} = \left. \frac{\partial \varrho[Z + hZ^{(j)}]}{\partial h} \right|_{h=0}; \quad (2.20)$$

see Litterman [146], Tasche [232, 233] and McNeil, Frey and Embrechts ([157], section 6.3). These are called the Euler allocations and represent capital allocation per unit of exposure  $Z^{(j)}$ . Tasche [232] showed that it is the only allocation compatible with RORAC (return on risk adjusted capital, i.e. expected return divided by risk capital) measure of performance in portfolio management. Another justification of the Euler allocations was given in Denaut [75] using game-theoretic considerations.

*Standard deviation risk measure.* In the case of standard deviation as a risk measure,  $\varrho[Z] = \text{stdev}[Z]$ , it is easy to show that

$$\varrho_j^{Euler} = \frac{\text{Cov}[Z^{(j)}, Z]}{\sqrt{\text{Var}[Z]}}. \quad (2.21)$$

*VaR and expected shortfall risk measures.* For risk measures  $\text{VaR}_\alpha[\cdot]$  and  $\text{ES}_\alpha[\cdot]$ , the derivatives in (2.20) can be calculated as

$$\left. \frac{\partial \text{VaR}_\alpha[Z + hZ^{(j)}]}{\partial h} \right|_{h=0} = E[Z^{(j)} | Z = \text{VaR}_\alpha[Z]], \quad (2.22)$$

$$\left. \frac{\partial \text{ES}_\alpha[Z + hZ^{(j)}]}{\partial h} \right|_{h=0} = E[Z^{(j)} | Z \geq \text{VaR}_\alpha[Z]]. \quad (2.23)$$

It is easy to verify that

$$\begin{aligned} \sum_{j=1}^J E[Z^{(j)} | Z = \text{VaR}_\alpha[Z]] &= E[Z | Z = \text{VaR}_\alpha[Z]] = \text{VaR}_\alpha[Z], \\ \sum_{j=1}^J E[Z^{(j)} | Z \geq \text{VaR}_\alpha[Z]] &= E[Z | Z \geq \text{VaR}_\alpha[Z]] = \text{ES}_\alpha[Z]. \end{aligned}$$

In general, the Euler allocations should be calculated numerically. Assume that the total capital is quantified using Monte Carlo methods. That is, a sample of independent and identically distributed annual losses  $z_k^{(j)}$ ,  $k = 1, \dots, K$  is simulated for each risk cell  $j$  (here, the dependence between risk cells is allowed). Then, a sample  $z_1, \dots, z_K$ , where  $z_k = \sum_{j=1}^J z_k^{(j)}$ , can be calculated and  $\text{VaR}_\alpha[Z]$  is estimated using the sample in the usual way. Denote this estimate by  $\widehat{\text{VaR}}_\alpha[Z]$ . Then the Euler allocations in the case of expected shortfall (2.23) are

$$E[Z^{(j)} | Z \geq \text{VaR}_\alpha[Z]] \approx \frac{\sum_{k=1}^K z_k^{(j)} 1_{\{z_k \geq \widehat{\text{VaR}}_\alpha[Z]\}}}{\sum_{k=1}^K 1_{\{z_k \geq \widehat{\text{VaR}}_\alpha[Z]\}}}. \quad (2.24)$$

In the case of VaR, the Euler allocation can be difficult to estimate using the Monte Carlo sample, because  $\Pr[Z = \text{VaR}_\alpha[Z]] = 0$  in the case of continuous distributions. To handle this problem, the condition  $Z = \text{VaR}_\alpha[Z]$  can be replaced by  $|Z - \text{VaR}_\alpha[Z]| < \epsilon$  for some  $\epsilon > 0$  large enough to have  $\Pr[|Z - \text{VaR}_\alpha[Z]| < \epsilon] > 0$ . However, this condition will be satisfied by only a few Monte Carlo simulations and importance sampling techniques are needed to get an accurate estimation; see Glasserman [109]. For VaR, it can be somewhat easier to calculate the Euler allocations using the finite difference approximation

$$\left. \frac{\partial \varrho[Z + hZ^{(j)}]}{\partial h} \right|_{h=0} \approx \frac{\varrho[Z + \Delta Z^{(j)}] - \varrho[Z]}{\Delta} \quad (2.25)$$

with some small suitable  $\Delta \neq 0$ . Note that the choice of  $\Delta$  depends on the numerical accuracy of the estimator for  $\varrho[\cdot]$  and curvature of the  $\varrho[\cdot]$  with respect to  $h$ . So,  $\Delta$  should be neither very small nor too large. This is a typical problem with estimating derivatives via finite difference and details can be found in many books on numerical recipes; see for example Press, Teukolsky, Vetterling and Flannery ([195], section 5.7).

### 2.7.2 Allocation by Marginal Contributions

Another popular way to allocate capital is using marginal risk contribution

$$\varrho_j^{marg} = \varrho[Z] - \varrho[Z - Z^{(j)}], \quad (2.26)$$

which is the difference between total risk (across all risk cell) and total risk without risk cell  $j$ . This can be viewed as some crude approximation of Euler allocation derivatives (2.25) but of course differentiability is not required to calculate marginal contribution. The sum of marginal contributions may not add up to  $\varrho[Z]$ . In particular, in the case of subadditive risk measures, it can be shown that

$$\varrho_j^{marg} \leq \varrho_j^{Euler}, \quad \sum_{j=1}^J \varrho_j^{marg} \leq \varrho[Z]. \quad (2.27)$$

One can define

$$AC_j = \frac{\varrho_j^{marg}}{\sum_{i=1}^J \varrho_i^{marg}} \varrho[Z], \quad (2.28)$$

to ensure that allocated capitals add up to  $\varrho[Z]$ .

*Example 2.1* To illustrate, consider an example of three risk cells where the annual losses  $Z^{(j)}$  are independent random variables from the lognormal distribution  $\mathcal{LN}(0, \sigma_j)$  with  $\sigma_1 = 1.5$ ,  $\sigma_2 = 1.75$ , and  $\sigma_3 = 2$  respectively. Results based on  $4 \times 10^6$  Monte Carlo simulations are given in Table 2.1. Here, we estimate VaR of the total loss,  $\text{VaR}_{0.999}[\sum_j Z^{(j)}] \approx 556$ , and VaRs of individual risk cells  $\text{VaR}_{0.999}[Z^{(j)}]$ ,  $j = 1, 2, 3$ . The numerical error due to the finite number of simulations is of the order of 1%.  $\hat{\varrho}_j^{Euler}$  was estimated using finite difference approximation (2.25) with  $\Delta = 0.02$ . Due to this approximation,  $\sum_j \hat{\varrho}_j^{Euler} \approx 553$  is slightly different from  $\text{VaR}_{0.999}[\sum_j Z^{(j)}] \approx 556$ , so the final estimate for capital allocations using Euler principle is

$$AC_j^{Euler} = \frac{\hat{\varrho}_j^{Euler}}{\sum_i \hat{\varrho}_i^{Euler}} \text{VaR}_{0.999} \left[ \sum_i Z^{(i)} \right].$$

The total diversification

$$1 - \frac{\text{VaR}_{0.999}[\sum_j Z^{(j)}]}{\sum_i \text{VaR}_{0.999}[Z^{(i)}]} \quad (2.29)$$

is approximately 30%. It is easy to observe that, both marginal and Euler allocations  $AC_j$  are significantly less than corresponding  $\text{VaR}_{0.999}[Z^{(j)}]$ .



**Table 2.1** Allocation of capital  $C = \text{VaR}_{0.999}[\sum_j Z^{(j)}] \approx 556$  by marginal and Euler contributions. Here,  $Z^{(j)} \sim \mathcal{LN}(0, \sigma_j)$ . Estimated  $\text{AC}_j$  are given in absolute terms and as a percent of the total  $C$ . See Example 2.1 for details

$j$	$\sigma_j$	$\text{VaR}_{0.999}[Z^{(j)}]$	$\hat{\varrho}_j^{marg}$	$\text{AC}_j^{marg}$	$\hat{\varrho}_j^{Euler}$	$\text{AC}_j^{Euler}$
1	1.5	103	9	13\2 %	20	20\4 %
2	1.75	221	58	84\15 %	102	103\18 %
3	2.0	490	314	459\83 %	431	433\78 %
Total		814	381	556\100 %	553	556\100 %

Also,  $\hat{\varrho}_j^{marg} < \hat{\varrho}_j^{Euler}$ . Finally, it is important to note that the relative importance of risk cells cannot be measured by simple ratios

$$\frac{\text{VaR}_{0.999}[Z^{(j)}]}{\sum_i \text{VaR}_{0.999}[Z^{(i)}]}, \quad j = 1, 2, 3,$$

which are, in this example, 13%, 27% and 60% respectively and very different from  $\text{AC}_j / \sum_i \text{AC}_i$ .

## 2.8 Model Fitting: Frequentist Approach

Estimation of the frequency and severity distributions is a challenging task, especially for low-frequency/high-severity losses, due to very limited data for these risks. The main tasks involved in fitting the frequency and severity distributions using data are:

- finding the best point estimates for the distribution parameters;
- quantification of the parameter uncertainties; and
- assessing the model quality (model error).

In general, these tasks can be accomplished by undertaking either the so-called frequentist or Bayesian approaches briefly discussed in this and the next section.

Fitting distribution parameters using data via the frequentist approach is a classical problem described in many textbooks. For the purposes of this book it is worth to mention several aspects and methods. Firstly, under the frequentist approach one says that the model parameters are fixed while their estimators have associated uncertainties that typically converge to zero when a sample size increases. Several popular methods to fit parameters (finding point estimators for the parameters) of the assumed distribution are:

- method of moments – finding the parameter estimators to match the observed moments;
- matching certain quantiles of the empirical distribution;
- maximum likelihood method – finding parameter values that maximise the joint density of observed data; and

- estimating parameters by minimising a certain distance between empirical and theoretical distributions, e.g. Anderson-Darling or other statistics; see Ergashev [89].

A *point estimator* is a function of a sample. Notationally, an *estimator* is a function of the sample while an *estimate* is the realised value of an estimator for a realisation of the sample. For example, given a vector of random variables  $\mathbf{X} = (X_1, X_2, \dots, X_K)'$ , the estimator is a function of  $\mathbf{X}$  while the estimate is a function of the realisation  $\mathbf{x}$ .

Given a sample  $\mathbf{X} = (X_1, X_2, \dots, X_K)'$  from a density  $f(\mathbf{x}|\theta)$ , we try to find a point estimator  $\hat{\Theta}$  for a parameter  $\theta$ . In most cases different methods will lead to different point estimators. One of the standard ways to evaluate an estimator is to calculate its *mean squared error*.

**Definition 2.15 (Mean squared error)** The mean squared error (MSE) of an estimator  $\hat{\Theta}$  for a parameter  $\theta$  is defined as

$$\text{MSE}_{\hat{\Theta}}(\theta) = E[(\hat{\Theta} - \theta)^2].$$

Any increasing function of  $|\hat{\Theta} - \theta|$  can be used as a measure of the accuracy of the estimator but MSE is the most popular due to tractability and good interpretation. In particular, it can be written as

$$\text{MSE}_{\hat{\Theta}}(\theta) = \text{Var}[\hat{\Theta}] + (E[\hat{\Theta}] - \theta)^2, \quad (2.30)$$

where the first term is due to the uncertainty (variability) of the estimator and the second term is due to the bias. The latter is defined as follows

**Definition 2.16 (Bias of a point estimator)** The *bias* of a point estimator  $\hat{\Theta}$  for a parameter  $\theta$  is

$$\text{Bias}_{\hat{\Theta}}(\theta) = E[\hat{\Theta}] - \theta.$$

An estimator with zero bias, i.e.  $E[\hat{\Theta}] = \theta$  is called *unbiased*. The MSE of an unbiased estimator is reduced to  $\text{MSE}_{\hat{\Theta}}(\theta) = \text{Var}[\hat{\Theta}]$ .

*Example 2.2* Consider a sample of independent random variables  $N_1, N_2, \dots, N_M$  from  $\text{Poisson}(\lambda)$ , i.e.  $E[N_m] = \lambda$ , and an estimator  $\hat{\Lambda} = \frac{1}{M} \sum_{m=1}^M N_m$  (in this case it is a maximum likelihood estimator; see Sect. 2.8.1 below). Then

$$E[\hat{\Lambda}] = \frac{1}{M} E \left[ \sum_{m=1}^M N_m \right] = \lambda.$$

Thus the estimator  $\hat{\Lambda}$  is an unbiased estimator of  $\lambda$ .

It is important for the point estimator of a parameter to be a *consistent* estimator, i.e. converge to the “true” value of the parameter in probability as the sample size

increases. Formally, a property of consistency is defined for a sequence of estimators as follows.

**Definition 2.17 (Consistent estimator)** For a sample  $X_1, X_2, \dots$ , a sequence of estimators

$$\hat{\Theta}_n = \hat{\Theta}_n(X_1, \dots, X_n), \quad n = 1, 2, \dots$$

for the parameter  $\theta$  is a consistent sequence of estimators if for every  $\epsilon > 0$

$$\lim_{n \rightarrow \infty} \Pr[|\hat{\Theta}_n - \theta| < \epsilon] = 1.$$

A more informative estimation of the parameter (in comparison with the point estimator) is based on a confidence interval specifying the range of possible values.

**Definition 2.18 (Confidence interval)** Given a data realisation  $\mathbf{X} = \mathbf{x}$ , the  $1 - \alpha$  *confidence interval* for a parameter  $\theta$  is  $[L(\mathbf{x}), U(\mathbf{x})]$  such that

$$\Pr[L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})] \geq 1 - \alpha.$$

That is, the random interval  $[L, U]$ , where  $L = L(\mathbf{X})$  and  $U = U(\mathbf{X})$ , contains the true value of parameter  $\theta$  with at least probability  $1 - \alpha$ .

Typically, it is difficult to construct a confidence interval exactly. However, often it can be found approximately using Gaussian distribution approximation in the case of large data samples; see e.g. Sect. 2.8.1. Specifically, if a point estimator  $\hat{\Theta}$  is distributed from  $\mathcal{N}(\theta, \sigma(\theta))$ , then

$$\Pr \left[ -F_N^{-1}(1 - \alpha/2) \leq \frac{\hat{\Theta} - \theta}{\sigma(\theta)} \leq F_N^{-1}(1 - \alpha/2) \right] = 1 - \alpha,$$

where  $F_N^{-1}(\cdot)$  is the inverse of the standard normal distribution  $\mathcal{N}(0, 1)$ . Note that  $\sigma(\theta)$  depends on  $\theta$ . For a given data realisation, typically  $\sigma(\theta)$  is replaced by  $\sigma(\hat{\theta})$  to approximate a confidence interval by

$$\left[ \hat{\theta} - F_N^{-1}(1 - \alpha/2)\sigma(\hat{\theta}), \hat{\theta} + F_N^{-1}(1 - \alpha/2)\sigma(\hat{\theta}) \right]. \quad (2.31)$$

### 2.8.1 Maximum Likelihood Method

The most popular approach to fit the parameters of the assumed distribution is the maximum likelihood method. Given the model parameters  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_K)'$ , assume that the joint density of data  $\mathbf{X} = (X_1, X_2, \dots, X_n)'$  is  $f(\mathbf{x}|\boldsymbol{\theta})$ . Then the *likelihood function* is defined as the joint density  $f(\mathbf{x}|\boldsymbol{\theta})$  considered as a function of parameters  $\boldsymbol{\theta}$ .

**Definition 2.19 (Likelihood function)** For a sample  $\mathbf{X} = \mathbf{x}$  from the joint density  $f(\mathbf{x}|\boldsymbol{\theta})$  with the parameter vector  $\boldsymbol{\theta}$ , the *likelihood function* is a function of  $\boldsymbol{\theta}$ :

$$\ell_{\mathbf{x}}(\boldsymbol{\theta}) = f(\mathbf{x}|\boldsymbol{\theta}). \quad (2.32)$$

The *log-likelihood function* is  $\ln \ell_{\mathbf{x}}(\boldsymbol{\theta})$ .

Often it is assumed that  $X_1, X_2, \dots, X_n$  are independent with a common density  $f(x|\boldsymbol{\theta})$ ; then the likelihood function is  $\ell_{\mathbf{x}}(\boldsymbol{\theta}) = \prod_{i=1}^n f(x_i|\boldsymbol{\theta})$ .

The maximum likelihood estimators  $\hat{\boldsymbol{\theta}}^{\text{MLE}} = \hat{\boldsymbol{\theta}}(\mathbf{X})$  of the parameters  $\boldsymbol{\theta}$  are formally defined as follows.

**Definition 2.20 (Maximum likelihood estimator)** For a sample  $\mathbf{X}$ ,  $\hat{\boldsymbol{\theta}}(\mathbf{X})$  is the *maximum likelihood estimator* (MLE), if for each realisation  $\mathbf{x}$ ,  $\hat{\boldsymbol{\theta}}(\mathbf{x})$  is a value of parameter  $\boldsymbol{\theta}$  maximising the likelihood function  $\ell_{\mathbf{x}}(\boldsymbol{\theta})$  or equivalently maximising the log-likelihood function  $\ln \ell_{\mathbf{x}}(\boldsymbol{\theta})$ .

An important property of MLEs is their convergence to the true value in probability as the sample size increases, i.e. MLEs are *consistent* estimators.

**Theorem 2.1** For a sample  $X_1, X_2, \dots, X_n$  of independent and identically distributed random variables from  $f(x|\boldsymbol{\theta})$  and corresponding MLE  $\hat{\boldsymbol{\theta}}_n$ , under the suitable regularity conditions, as the sample size  $n$  increases,

$$\lim_{n \rightarrow \infty} \Pr[|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}| \geq \epsilon] = 0 \quad \text{for every } \epsilon > 0. \quad (2.33)$$

The required regularity conditions are:

- The parameter is identifiable:  $\boldsymbol{\theta} \neq \tilde{\boldsymbol{\theta}} \Rightarrow f(x|\boldsymbol{\theta}) \neq f(x|\tilde{\boldsymbol{\theta}})$ .
- The true parameter should be an interior point of the parameter space.
- The support of  $f(x|\boldsymbol{\theta})$  should not depend on  $\boldsymbol{\theta}$ .
- $f(x|\boldsymbol{\theta})$  should be differentiable in  $\boldsymbol{\theta}$ .

Asymptotically, for large sample size, under stronger conditions (that further require  $f(x|\boldsymbol{\theta})$  to be differentiable three times with respect to  $\boldsymbol{\theta}$  and to have continuous and bounded 3rd derivatives), the MLEs are distributed from the normal distribution:

**Theorem 2.2** Under the suitable regularity conditions, for a sample  $X_1, X_2, \dots, X_n$  of independent and identically distributed random variables from  $f(x|\boldsymbol{\theta})$ ,  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_K)'$ , and corresponding MLE  $\hat{\boldsymbol{\theta}}_n$ :

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \rightarrow \mathcal{N}_K \left( 0, [\mathbf{I}(\boldsymbol{\theta})]^{-1} \right), \quad (2.34)$$

as the sample size  $n$  increases. Here,  $[\mathbf{I}(\boldsymbol{\theta})]^{-1}$  is the inverse matrix of the expected Fisher information matrix for one observation  $\mathbf{I}(\boldsymbol{\theta})$ , whose matrix elements are

$$\begin{aligned}
\mathbf{I}(\boldsymbol{\theta})_{km} &= \mathbb{E} \left[ \frac{\partial}{\partial \theta_k} \ln f(X_1 | \boldsymbol{\theta}) \frac{\partial}{\partial \theta_m} \ln f(X_1 | \boldsymbol{\theta}) \right] \\
&= -\mathbb{E} \left[ \frac{\partial^2}{\partial \theta_k \partial \theta_m} \ln f(X_1 | \boldsymbol{\theta}) \right].
\end{aligned} \tag{2.35}$$

That is,  $\hat{\boldsymbol{\theta}}^{\text{MLE}}$  converges to  $\boldsymbol{\theta}$  as the sample size increases and asymptotically  $\hat{\boldsymbol{\theta}}^{\text{MLE}}$  is normally distributed with the mean  $\boldsymbol{\theta}$  and covariance matrix  $n^{-1} \mathbf{I}(\boldsymbol{\theta})^{-1}$ . For precise details on regularity conditions and proofs, see Lehmann ([143], Theorem 6.2.1 and 6.2.3); these can also be found in many other books such as Casella and Berger ([49], p. 516), Stuart, Ord and Arnold ([225], chapter 18), Ferguson ([93], part 4) or Lehmann and Casella ([144], section 6.3).

In practice, this asymptotic result is often used even for small samples and for the cases that do not formally satisfy the regularity conditions. Note that the mean and covariances depend on the unknown parameters  $\boldsymbol{\theta}$  and are usually estimated by replacing  $\boldsymbol{\theta}$  with  $\hat{\boldsymbol{\theta}}^{\text{MLE}}$  for a given realisation of data. Often in practice, the expected Fisher information matrix is approximated by the *observed information matrix*

$$\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}})_{km} = -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \ln f(x_i | \boldsymbol{\theta})}{\partial \theta_k \partial \theta_m} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -\frac{1}{n} \frac{\partial^2 \ln \ell_{\mathbf{x}}(\boldsymbol{\theta})}{\partial \theta_k \partial \theta_m} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{2.36}$$

for a given realisation of data. This should converge to the expected information matrix by the law of large numbers. It has been suggested in Efron and Hinkley [78], that the use of the observed information matrix leads to a better inference in comparison with the expected information matrix.

Though very useful and widely used, these asymptotic approximations are usually not accurate enough for small samples, that is the distribution of parameter errors can be materially different from normal and MLEs may have significant bias. Also, as for any asymptotic results, a priori, one cannot decide on a sample size that is large enough to use the asymptotic approximation.

To assess the quality of the fit, there are several popular goodness of fit tests including Kolmogorov-Smirnov, Anderson-Darling and Chi-square tests. Also, the likelihood ratio test and Akaike's information criterion are often used to compare models.

Usually maximisation of the likelihood (or minimisation of some distances in other methods) must be done numerically. Popular numerical optimisation algorithms include simplex method, Newton methods, expectation maximisation (EM) algorithm, and simulated annealing. It is worth mentioning that the last is attempting to find a global maximum while other methods find a local maximum. Also, EM is usually more stable and robust than the standard deterministic methods such as simplex or Newton methods.

Again, detailed descriptions of the above-mentioned methodologies can be found in many textbooks; for application in an operational risk context, see Panjer [181].

### 2.8.2 Bootstrap

Another popular method to estimate parameter uncertainties is the so-called *bootstrap*. This method is based on a simple idea: that we can learn about characteristics of a sample by taking resamples from the original sample and calculating the parameter estimates for each sample to assess the parameter variability. The bootstrap method was originally developed by Efron in the 1970s. For a good introduction to the method we refer the reader to Efron and Tibshirani [79]. Often the bootstrap estimators are reasonable and consistent. Two types of bootstrapping, *nonparametric bootstrap* and *parametric bootstrap*, are commonly used in practice.

*Nonparametric bootstrap.* Suppose we have a sample of independent and identically distributed random variables  $\mathbf{X} = (X_1, X_2, \dots, X_K)'$  and there is an estimator  $\hat{\Theta}(\mathbf{X})$ . Then:

- Draw  $M$  independent samples

$$\mathbf{X}^{(m)} = (X_1^{(m)}, X_2^{(m)}, \dots, X_K^{(m)})', \quad m = 1, \dots, M$$

with replacement from the original sample  $\mathbf{X}$ . That is  $X_k^{(m)}$ ,  $k = 1, \dots, K$ ,  $m = 1, \dots, M$  are independent and identically distributed, and drawn from the empirical distribution of the original sample  $\mathbf{X}$ .

- Calculate estimator  $\hat{\Theta}^{(m)} = \hat{\Theta}(\mathbf{X}^{(m)})$  for each resample  $m = 1, \dots, M$ .
- Calculate

$$\widehat{\text{Var}}[\hat{\Theta}] = \frac{1}{M-1} \sum_{m=1}^M \left( \hat{\Theta}^{(m)} - \mu \right)^2, \quad \text{where } \mu = \frac{1}{M} \sum_{m=1}^M \hat{\Theta}^{(m)}. \quad (2.37)$$

*Parametric bootstrap.* Suppose we have a sample of independent and identically distributed random variables  $\mathbf{X} = (X_1, X_2, \dots, X_K)'$  from  $f(x|\theta)$  and we can calculate some estimator  $\hat{\Theta}(\mathbf{X})$  (e.g. MLE) for  $\theta$ . Then:

- Draw  $M$  independent samples

$$\mathbf{X}^{(m)} = (X_1^{(m)}, X_2^{(m)}, \dots, X_K^{(m)})', \quad m = 1, \dots, M,$$

where  $X_k^{(m)}$ ,  $k = 1, \dots, K$ ,  $m = 1, \dots, M$  are independent and identically distributed from  $f(x|\hat{\theta})$ .

- Calculate estimator  $\hat{\Theta}^{(m)} = \hat{\Theta}(\mathbf{X}^{(m)})$  for each resample  $m = 1, \dots, M$ .
- Calculate  $\widehat{\text{Var}}[\hat{\Theta}] = \frac{1}{M-1} \sum_{m=1}^M \left( \hat{\Theta}^{(m)} - \mu \right)^2$ , where  $\mu = \frac{1}{M} \sum_{m=1}^M \hat{\Theta}^{(m)}$ .

The obtained  $\widehat{\text{Var}}[\hat{\Theta}]$  is used as an estimator for  $\text{Var}[\hat{\Theta}]$ . Typically, for independent and identically distributed samples, this estimator is consistent, i.e.

$$\widehat{\text{Var}}[\hat{\Theta}] \rightarrow \text{Var}[\hat{\Theta}], \quad \text{as } M \rightarrow \infty \text{ and } K \rightarrow \infty, \quad (2.38)$$

though in more general situations it may not occur.

*Remark 2.8* More accurate treatment of nonparametric bootstrap estimators involves an approximator

$$\widehat{\text{Var}}^*[\widehat{\Theta}] = \frac{1}{N-1} \sum_{m=1}^N \left( \widehat{\Theta}^{(m)} - \mu \right)^2, \quad \mu = \frac{1}{N} \sum_{m=1}^N \widehat{\Theta}^{(m)},$$

where  $N = K^K$  is the total number of nondistinct resamples.  $N$  is very large even for small  $K$ , e.g. for  $K = 10$ ,  $N = 10^{10}$ . Calculations of the variance estimators (2.37) with  $M \ll N$  is considered as approximation for  $\widehat{\text{Var}}^*$  variances. Then, convergence of bootstrap estimators is considered in two steps:  $\widehat{\text{Var}}[\widehat{\Theta}] \rightarrow \widehat{\text{Var}}^*[\widehat{\Theta}]$  as  $M \rightarrow \infty$ ; and  $\widehat{\text{Var}}^*[\widehat{\Theta}] \rightarrow \text{Var}[\widehat{\Theta}]$  as  $K \rightarrow \infty$ .

## 2.9 Bayesian Inference Approach

There is a broad literature covering Bayesian inference and its applications for the insurance industry as well as other areas. For a good introduction to the Bayesian inference method, see Berger [27] and Robert [200]. This approach is well suited for operational risk and will be a central topic in this book. It is sketched below to introduce some notation and concepts, and then it will be discussed in detail in Chap. 4.

Consider a random vector of data  $\mathbf{X} = (X_1, X_2, \dots, X_n)'$  whose density, for a given vector of parameters  $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_K)'$ , is  $f_{\mathbf{X}|\boldsymbol{\theta}}(\mathbf{x}|\boldsymbol{\theta})$ . In the Bayesian approach, both data and parameters are considered to be random. A convenient interpretation is to think that parameter is a random variable with some distribution and the true value (which is deterministic but unknown) of the parameter is a realisation of this random variable. Then the joint density of the data and parameters is

$$f_{\mathbf{X}, \boldsymbol{\Theta}}(\mathbf{x}, \boldsymbol{\theta}) = f_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x}|\boldsymbol{\theta})\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \pi_{\boldsymbol{\Theta}|\mathbf{X}}(\boldsymbol{\theta}|\mathbf{x})f_{\mathbf{X}}(\mathbf{x}), \quad (2.39)$$

where

- $\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$  is the density of parameters (a so-called *prior density*);
- $\pi_{\boldsymbol{\Theta}|\mathbf{X}}(\boldsymbol{\theta}|\mathbf{x})$  is the density of parameters given data  $\mathbf{X} = \mathbf{x}$  (a so-called *posterior density*);
- $f_{\mathbf{X}, \boldsymbol{\Theta}}(\mathbf{x}, \boldsymbol{\theta})$  is the joint density of the data and parameters;
- $f_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x}|\boldsymbol{\theta})$  is the density of the data given parameters  $\boldsymbol{\Theta} = \boldsymbol{\theta}$ . This is the same as a likelihood function, see (2.32), if considered as a function of  $\boldsymbol{\theta}$  for a given  $\mathbf{x}$ , i.e.  $\ell_{\mathbf{x}}(\boldsymbol{\theta}) = f_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x}|\boldsymbol{\theta})$ ;
- $f_{\mathbf{X}}(\mathbf{x})$  is the marginal density of  $\mathbf{X}$ . If  $\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$  is continuous, then

$$f_{\mathbf{X}}(\mathbf{x}) = \int f_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x}|\boldsymbol{\theta})\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta})d\boldsymbol{\theta}$$

and if  $\pi_{\Theta}(\theta)$  is a discrete, then the integration should be replaced by a corresponding summation.

*Remark 2.9* Typically,  $\pi_{\Theta}(\theta)$  depends on a set of further parameters, the so-called *hyper-parameters*, omitted here for simplicity of notation. The choice and estimation of the prior will be discussed later in Chap. 4.

Using (2.39), the well-known Bayes's theorem, Bayes [21], says that:

**Theorem 2.3 (Bayes's theorem)** *The posterior density can be calculated as*

$$\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) = f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\pi_{\Theta}(\theta)/f_{\mathbf{X}}(\mathbf{x}). \quad (2.40)$$

Here,  $f_{\mathbf{X}}(\mathbf{x})$  plays the role of a normalisation constant and the posterior can be viewed as a combination of prior knowledge (contained in  $\pi_{\Theta}(\theta)$ ) with information from the data (contained in  $f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)$ ).

Given that  $f_{\Theta}(\mathbf{x})$  is a normalisation constant, the posterior is often written as

$$\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x}) \propto f_{\mathbf{X}|\Theta}(\mathbf{x}|\theta)\pi_{\Theta}(\theta), \quad (2.41)$$

where “ $\propto$ ” means “is proportional to” with a constant of proportionality independent of the parameter  $\theta$ . Typically, in closed-form calculations, the right hand side of the equation is calculated as a function of  $\theta$  and then the normalisation constant is determined by integration over  $\theta$ .

Using the posterior  $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$ , one can easily construct a probability interval for  $\Theta$ , which is the analogue for confidence intervals (see Definition 2.18) under the frequentist approach.

**Definition 2.21 (Credibility interval)** Given a data realisation  $\mathbf{X} = \mathbf{x}$ , if  $\pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})$  is the posterior density of  $\Theta$  and

$$\Pr[a \leq \Theta \leq b|\mathbf{X} = \mathbf{x}] = \int_a^b \pi_{\Theta|\mathbf{X}}(\theta|\mathbf{x})d\theta \geq 1 - \alpha,$$

then the interval  $[a, b]$  contains the true value of parameter  $\theta$  with at least probability  $1 - \alpha$ . The interval  $[a, b]$  is called a *credibility interval* (sometimes referred to as *predictive interval* or *credible interval*) for parameter  $\theta$ .

*Remark 2.10*

- The inequality in the above definition is to cover the case of discrete posterior distributions.
- Typically, one chooses the smallest possible interval  $[a, b]$ . Also, one can consider one-sided intervals, e.g.  $\Pr[\Theta \leq b|\mathbf{X} = \mathbf{x}]$ .
- Extension to the multivariate case, i.e. parameter vector  $\theta$ , is trivial.
- Though the Bayesian credibility interval looks similar to the frequentist confidence interval (see Definition 2.18), these intervals are conceptually different. To determine a confidence (probability to contain the true value) the bounds of the frequentist confidence interval are considered to be random (functions of



random data) while bounds of the Bayesian credibility interval are functions of a data realisation. For some special cases the intervals are the same (for given data realisation) but in general they are different especially in the case of strong prior information.

If the data  $X_1, X_2, \dots$  are conditionally (given  $\Theta = \theta$ ) independent then the posterior can be calculated iteratively, i.e. the posterior distribution calculated after  $k-1$  observations can be treated as a prior distribution for the  $k$ -th observation. Thus the loss history over many years is not required, making the model easier to understand and manage, and allowing experts to adjust the priors at every step.

*For simplicity of notation, the density and distribution subscripts indicating random variables will often be omitted, e.g.  $\pi_{\Theta}(\theta)$  will be written as  $\pi(\theta)$ .*

### 2.9.1 Conjugate Prior Distributions

Sometimes the posterior density can be calculated in closed form, which is very useful in practice when Bayesian inference is applied. This is the case for the so-called conjugate prior distributions, where the prior and posterior distributions are of the same type.

**Definition 2.22 (Conjugate prior)** Let  $F$  denote a class of density functions  $f(\mathbf{x}|\theta)$ , indexed by  $\theta$ . A class  $U$  of prior densities  $\pi(\theta)$  is said to be a conjugate family for  $F$  and  $F - U$  is called a conjugate pair, if the posterior density  $\pi(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)\pi(\theta)/f(\mathbf{x})$ , where  $f(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta)d\theta$ , is in the class  $U$  for all  $f \in F$  and  $\pi \in U$ .

Formally, if the family  $U$  contains all distribution functions then it is conjugate to any family  $F$ . However, to make a model useful in practice it is important that  $U$  should be as small as possible while containing realistic distributions. In Chap. 4, we present  $F - U$  conjugate pairs (Poisson-gamma, lognormal-normal, Pareto-gamma) that are useful and illustrative examples of modelling frequencies and severities in operational risk. Several other pairs (binomial-beta, gamma-gamma, exponential-gamma) can be found for example in Bühlmann and Gisler [44]. In all these cases, the prior and posterior distributions have the same type and the posterior distribution parameters are easily calculated using the prior distribution parameters and observations (or recursively).

In general, if the posterior cannot be found in closed form or is difficult to evaluate, one can use Gaussian approximation or Markov chain Monte Carlo methods, discussed next.

### 2.9.2 Gaussian Approximation for Posterior

For a given data realisation  $\mathbf{X} = \mathbf{x}$ , denote the mode of the posterior  $\pi(\boldsymbol{\theta}|\mathbf{x})$  by  $\widehat{\boldsymbol{\theta}}$ . If the prior is continuous at  $\widehat{\boldsymbol{\theta}}$ , then a Gaussian approximation for the posterior is obtained by a second-order Taylor series expansion around  $\widehat{\boldsymbol{\theta}}$ :

$$\ln \pi(\boldsymbol{\theta}|\mathbf{x}) \approx \ln \pi(\widehat{\boldsymbol{\theta}}|\mathbf{x}) + \frac{1}{2} \sum_{i,j} \frac{\partial^2 \ln \pi(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_i \partial \theta_j} \bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} (\theta_i - \widehat{\theta}_i)(\theta_j - \widehat{\theta}_j). \quad (2.42)$$

Under this approximation,  $\pi(\boldsymbol{\theta}|\mathbf{x})$  is a multivariate normal distribution with the mean  $\widehat{\boldsymbol{\theta}}$  and covariance matrix

$$\boldsymbol{\Sigma} = \mathbf{I}^{-1}, \quad (\mathbf{I})_{ij} = -\frac{\partial^2 \ln \pi(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_i \partial \theta_j} \bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}. \quad (2.43)$$

*Remark 2.11* In the case of improper constant priors, this approximation is comparable to the Gaussian approximation for the MLEs (2.34). Also, note that in the case of constant priors, the mode of the posterior and the MLE are the same. This is also true if the prior is uniform within a bounded region, provided that the MLE is within this region.

### 2.9.3 Posterior Point Estimators

Once the posterior density  $\pi(\boldsymbol{\theta}|\mathbf{x})$  is found, for given data  $\mathbf{X}$ , one can define point estimators of  $\Theta$ . The mode and mean of the posterior are the most popular point estimators. These Bayesian estimators are typically referred to as the Maximum a Posteriori (MAP) estimator and the Minimum Mean Square Estimator (MMSE), formally defined as follows:

$$\text{MAP : } \widehat{\Theta}^{\text{MAP}} = \arg \max_{\boldsymbol{\theta}} [\pi(\boldsymbol{\theta} | \mathbf{X})], \quad (2.44)$$

$$\text{MMSE : } \widehat{\Theta}^{\text{MMSE}} = \mathbf{E}[\Theta|\mathbf{X}]. \quad (2.45)$$

The median of the posterior is also often used as a point estimator for  $\Theta$ . Also, note that if the prior  $\pi(\boldsymbol{\theta})$  is constant and the parameter range includes the MLE, then the MAP of the posterior is the same as the MLE; see Remark 2.11.

More formally, the choice of point estimators is considered using a *loss function*,  $l(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})$ , that measures the cost (loss) of a decision to use a particular point estimator  $\widehat{\Theta}$ . For example:

- quadratic loss:  $l(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^2$ ;
- absolute loss:  $l(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = |\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}|$ ;
- all or nothing loss:  $l(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = 0$  if  $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$  and  $l(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}) = 1$  otherwise;

- asymmetric loss function: e.g.  $l(\theta, \hat{\theta}) = \hat{\theta} - \theta$  if  $\hat{\theta} > \theta$  and  $l(\theta, \hat{\theta}) = -2(\hat{\theta} - \theta)$  otherwise.

Then the value of  $\hat{\Theta}$  that minimises  $E[l(\Theta, \hat{\Theta})|\mathbf{X}]$  is called a Bayesian point estimator of  $\Theta$ . Here, the expectation is calculated with respect to the posterior  $\pi(\theta|\mathbf{X})$ . In particular:

- The posterior mean is a Bayesian point estimator in the case of a quadratic loss function.
- In the case of an absolute loss function, the Bayesian point estimator is the median of the posterior.
- All or nothing loss function gives the mode of the posterior as the point estimator.

*Remark 2.12*  $\hat{\Theta} = \hat{\Theta}(\mathbf{X})$  is a function of data  $\mathbf{X}$  and thus it is referred to as estimator. For a given data realisation  $\mathbf{X} = \mathbf{x}$ , we get  $\hat{\Theta} = \hat{\theta}$  which is referred to as a point estimate.

Though the point estimators are useful, for quantification of operational risk annual loss distribution and capital we recommend the use of the whole posterior, as discussed in following chapters.

### 2.9.4 Restricted Parameters

In practice, it is not unusual to restrict parameters. In this case the posterior distribution will be a truncated version of the posterior distribution in the unrestricted case. That is, if  $\theta$  is restricted to some range  $[\theta_L, \theta_H]$  then the posterior distribution will have the same type as in the unrestricted case but truncated outside this range.

For example, we choose the lognormal distribution,  $\mathcal{LN}(\mu, \sigma)$  to model the data  $\mathbf{X} = (X_1, \dots, X_n)'$  and we choose a prior distribution for  $\mu$  to be the normal distribution  $\mathcal{N}(\mu_0, \sigma_0)$ . This case will be considered in Sect. 4.3.4. However, if we know that  $\mu$  cannot be negative, we restrict  $\mathcal{N}(\mu_0, \sigma_0)$  to nonnegative values only.

Another example is the Pareto-gamma case, where the losses are modelled by  $Pareto(\xi, L)$  and the prior distribution for the tail parameter  $\xi$  is  $Gamma(\alpha, \beta)$ ; see Sect. 4.3.6. The prior is formally defined for  $\xi > 0$ . However, if we do not want to allow infinite mean predicted loss, then the parameter should be restricted to  $\xi > 1$ .

These cases can be easily handled by using the truncated versions of the prior-posterior distributions. Assume that  $\pi(\theta)$  is the prior whose corresponding posterior density is  $\pi(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)\pi(\theta)/f(\mathbf{x})$ , where  $\theta$  is unrestricted. If the parameter is restricted to  $a \leq \theta \leq b$ , then we can consider the prior

$$\pi^{\text{tr}}(\theta) = \frac{\pi(\theta)}{\Pr[a \leq \theta \leq b]} 1_{\{a \leq \theta \leq b\}}, \quad \Pr[a \leq \theta \leq b] = \int_a^b \pi(\theta) d\theta, \quad (2.46)$$

for some  $a$  and  $b$  with  $\Pr[a \leq \theta \leq b] > 0$ .  $\Pr[a \leq \theta \leq b]$  plays the role of normalisation and thus the posterior density for this prior is simply

$$\pi^{\text{tr}}(\theta|\mathbf{x}) = \frac{\pi(\theta|\mathbf{x})}{\Pr[a \leq \theta \leq b|\mathbf{x}]} 1_{\{a \leq \theta \leq b\}}, \quad \Pr[a \leq \theta \leq b|\mathbf{x}] = \int_a^b \pi(\theta|\mathbf{x}) d\theta. \quad (2.47)$$

*Remark 2.13* It is obvious that if  $\pi(\theta)$  is a conjugate prior, then  $\pi^{\text{tr}}(\theta)$  is a conjugate prior too.

### 2.9.5 Noninformative Prior

Sometimes there is no prior knowledge about the model parameter  $\theta$ , or we would like to rely on data only and avoid an impact from any subjective information. In this case we need a *noninformative prior* (sometimes called *vague prior*) that attempts to represent a near-total absence of prior knowledge. A natural noninformative prior is the uniform density

$$\pi(\theta) \propto \text{const} \quad \text{for all } \theta. \quad (2.48)$$

If parameter  $\theta$  is restricted to a finite set, then this  $\pi(\theta)$  corresponds to a proper uniform distribution. For example, the parameter  $p$  in a binomial distribution  $\text{Bin}(n, p)$  is restricted to the interval  $[0, 1]$ . Then one can choose a noninformative constant prior which is the uniform distribution  $\mathcal{U}(0, 1)$ .

However, if the parameter  $\theta$  is not restricted, then a constant prior is not a proper density (since  $\int f(\theta)d\theta = \infty$ ). Such a prior is called an *improper prior*. For example, the parameter  $\mu$  (mean) of the normal distribution  $\mathcal{N}(\mu, \sigma)$  is defined on  $(-\infty, \infty)$ . Then, for any constant  $c > 0$ ,  $\pi(\mu) = c$  is not a proper density because  $\int \pi(\mu)d\mu = \infty$ . It is not a problem to use improper priors as long as the posterior is a proper distribution. Also, as noted in previous sections, if the prior  $\pi(\theta)$  is constant and the parameter range includes the MLE, then the mode of the posterior is the same as the MLE; see Remark 2.11.

A constant prior is often used as a noninformative prior, though it can be criticised for a lack of invariance under transformation. For example, if a constant prior is used for parameter  $\theta$  and model is reparameterised in terms of  $\tilde{\theta} = \exp(\theta)$ , then the prior density for  $\tilde{\theta}$  is proportional to  $1/\tilde{\theta}$ . Thus we cannot choose a constant prior for both  $\theta$  and  $\tilde{\theta}$ . In this case, one typically argues that some chosen parameterisation is the most intuitively reasonable and absence of prior information corresponds to a constant prior in this parameterisation. One can propose noninformative priors through consideration of problem transformations. This has been considered in many studies starting with Jeffreys [127]. For discussion on this topic, see Berger ([27], section 3.3). Here, we just mention that for a scale densities of the form  $\theta^{-1}f(x/\theta)$ , the recommended noninformative prior for a scale parameter  $\theta > 0$  is

$$\pi(\theta) \propto \frac{1}{\theta}, \quad (2.49)$$

which is an improper prior because  $\int_0^\infty \pi(\theta)d\theta = \infty$ .

## 2.10 Mean Square Error of Prediction

To illustrate the difference between the frequentist and Bayesian approaches, consider the so-called (conditional) mean square error of prediction (MSEP) which is often used for prediction of uncertainty.

Consider a sample  $X_1, X_2, \dots, X_n, \dots$  and assume that, given data

$$\mathbf{X} = (X_1, X_2, \dots, X_n)',$$

we are interested in prediction of a random variable  $R$  which is a some function of  $X_{n+1}, X_{n+2}, \dots$ . Assume that  $\hat{R}$  is a predictor for  $R$  and an estimator for  $E[R|\mathbf{X}]$ . Then, the conditional MSEP is defined by

$$\text{MSEP}_{R|\mathbf{X}}(\hat{R}) = E[(R - \hat{R})^2|\mathbf{X}]. \quad (2.50)$$

It allows for a good interpretation if decoupled into *process variance* and *estimation error* as

$$\begin{aligned} \text{MSEP}_{R|\mathbf{X}}(\hat{R}) &= \text{Var}[R|\mathbf{X}] + (E[R|\mathbf{X}] - \hat{R})^2 \\ &= \text{process variance} + \text{estimation error}. \end{aligned} \quad (2.51)$$

It is clear that the estimator  $\hat{R}$  that minimises conditional MSEP is  $\hat{R} = E[R|\mathbf{X}]$ . Assume that the model is parameterised by the parameter vector  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)'$ . Then under the frequentist and Bayesian approaches we get the following estimators of MSEP.

*Frequentist approach.* Unfortunately, in frequentist approach  $E[R|\mathbf{X}]$  is unknown and the second term in (2.51) is often estimated by  $\text{Var}[\hat{R}]$ ; see Wüthrich and Merz ([240], section 6.4.3). Under the frequentist approach,  $\text{Var}[R|\mathbf{X}]$  and  $E[R|\mathbf{X}]$  are functions of parameter  $\boldsymbol{\theta}$  and can be denoted as  $\text{Var}_{\boldsymbol{\theta}}[R|\mathbf{X}]$  and  $E_{\boldsymbol{\theta}}[R|\mathbf{X}]$  respectively. Typically these are estimated as  $\widehat{\text{Var}}_{\boldsymbol{\theta}}[R|\mathbf{X}] = \text{Var}_{\hat{\boldsymbol{\theta}}}[R|\mathbf{X}]$  and  $\widehat{E}_{\boldsymbol{\theta}}[R|\mathbf{X}] = E_{\hat{\boldsymbol{\theta}}}[R|\mathbf{X}]$ , where  $\hat{\boldsymbol{\theta}}$  is a point estimator of  $\boldsymbol{\theta}$  obtained by maximum likelihood or other methods. Also, typically one chooses  $\hat{R} = E_{\hat{\boldsymbol{\theta}}}[R|\mathbf{X}]$ , so that now  $\hat{R}$  is a function of  $\hat{\boldsymbol{\theta}}$ , that we denote as  $\hat{R}(\hat{\boldsymbol{\theta}})$ . The parameter uncertainty term  $\text{Var}_{\boldsymbol{\theta}}[\hat{R}]$  is usually estimated using the first-order Taylor expansion of  $\hat{R}(\hat{\boldsymbol{\theta}})$  around  $\boldsymbol{\theta}$

$$\hat{R}(\hat{\boldsymbol{\theta}}) \approx \hat{R}(\boldsymbol{\theta}) + \sum_i \left. \frac{\partial \hat{R}(\hat{\boldsymbol{\theta}})}{\partial \hat{\theta}_i} \right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} (\hat{\theta}_i - \theta_i)$$

leading to

$$\text{Var}_{\boldsymbol{\theta}}[\hat{R}(\hat{\boldsymbol{\theta}})] \approx \sum_{i,j} \left. \frac{\partial \hat{R}}{\partial \hat{\theta}_i} \right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} \left. \frac{\partial \hat{R}}{\partial \hat{\theta}_j} \right|_{\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}} \text{Cov}[\hat{\theta}_i, \hat{\theta}_j].$$

Estimating  $\theta$  by  $\hat{\Theta}$  gives the final estimator

$$\widehat{\text{Var}}_{\theta}[\hat{R}(\hat{\Theta})] = \text{Var}_{\hat{\Theta}}[\hat{R}(\hat{\Theta})].$$

Note that if the point estimators are unbiased, i.e.  $E[\hat{\Theta}_i - \theta_i] = 0$  then  $E[\hat{R}(\hat{\Theta})] \approx \hat{R}(\theta)$ . Finally, the estimator for conditional MSE is

$$\begin{aligned} \widehat{\text{MSEP}}_{R|X}[\hat{R}] &= \widehat{\text{Var}}[R|X] + \widehat{\text{Var}}[\hat{R}] \\ &= \text{process variance} + \text{estimation error}. \end{aligned} \quad (2.52)$$

The above estimators are typically consistent and unbiased in the limit of large sample size.

*Bayesian approach.* Under the Bayesian inference approach, where the unknown parameters  $\theta$  are modelled as random variables  $\Theta$ ,  $\text{Var}[R|X]$  can be decomposed as

$$\begin{aligned} \text{Var}[R|X] &= E[\text{Var}[R|\Theta, X]|X] + \text{Var}[E[R|\Theta, X]|X] \\ &= \text{average process variance} + \text{parameter estimation error} \end{aligned} \quad (2.53)$$

that equals  $\text{MSEP}_{R|X}[\hat{R}]$  if we choose  $\hat{R} = E[R|X]$ . Estimation of the terms involved requires knowledge of the posterior distribution for  $\Theta$  that can be obtained either analytically or approximated accurately using Markov chain Monte Carlo methods discussed in the next section.

## 2.11 Markov Chain Monte Carlo Methods

As has already been mentioned, the posterior distribution is often not known in closed form. Thus, typically, estimation of the posterior empirically by direct simulation is also problematic. Then, in general, Markov chain Monte Carlo methods (hereafter referred to as MCMC methods) can be used. These are described below.

Simulation from the known density function can be accomplished using well-known generic methods such as the inverse transform, or accept-reject methods; see Glasserman ([108], section 2.2).

**Corollary 2.1 (The inverse transform)** *If  $U \sim \mathcal{U}(0, 1)$ , then the distribution of the random variable  $X = F^{-1}(U)$  is  $F(x)$ .*

*Remark 2.14* That is, to simulate  $X$  from the distribution  $F(x)$  using the inverse transform, generate  $U \sim \mathcal{U}(0, 1)$  and calculate  $X = F^{-1}(U)$ .

**Corollary 2.2** *Simulating  $X$  from the density  $f(x)$  is equivalent to simulating  $(X, U)$  from the uniform distribution on  $(x, u)$ , where  $0 \leq u \leq f(x)$ .*

*Remark 2.15* This means that to simulate  $X$  from the density  $f(x)$ , generate  $(X, U)$  from the uniform distribution under the curve of  $f(x)$ . The latter is typically done through accept-reject algorithm (or sometimes called as rejection sampling).

**Corollary 2.3 (Accept-reject method)** Assume that the density  $f(x)$  is bounded by  $M$  (i.e.  $f(x) \leq M$ ) and defined on the support  $a \leq x \leq b$ . Then, to simulate  $X$  with the density  $f(x)$ :

- draw  $X \sim \mathcal{U}(a, b)$  and  $U \sim \mathcal{U}(0, M)$ ;
- accept the sample of  $X$  if  $U \leq f(X)$ , otherwise repeat the above steps.

If another density  $g(x)$  such that  $Mg(x) \geq f(x)$  can be found for constant  $M$ , then to simulate  $X$  with the density  $f(x)$ :

- draw  $X$  from  $g(x)$  and  $U \sim \mathcal{U}(0, Mg(X))$ ;
- accept the sample of  $X$  if  $U \leq f(X)$ , otherwise repeat the above steps.

The inverse method cannot be used if the normalisation constant is unknown, and the above accept-reject method cannot be used if you cannot easily find the bounds for the density. These difficulties are typical for the posterior densities. In general, estimation (sampling) of the posterior  $\pi(\theta|\mathbf{x})$  numerically can be accomplished using MCMC methods; for a good introduction see Robert and Casella [201]. MCMC has almost unlimited applicability though its performance depends on the problem particulars. The idea of MCMC methods is based on a simple observation that to obtain an acceptable approximation to some integrals depending on a distribution of interest  $\pi(\theta|\mathbf{x})$ , it is enough to sample a sequence (Markov chain)  $\{\theta^{(1)}, \theta^{(2)}, \dots\}$ , whose limiting density is the density of interest  $\pi(\theta|\mathbf{x})$ . This idea appeared as early as the original Monte Carlo method but became very popular and practical in the last few decades only when fast computing platforms became available.

A Markov chain is a sequence of random variables defined as follows:

**Definition 2.23 (Markov chain)** A sequence of random variables

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

is a *Markov chain* if, for any  $l$ , the conditional distribution of  $\Theta^{(l+1)}$  given  $\Theta^{(i)}$ ,  $i = 0, 1, \dots, l$  is the same as the conditional distribution of  $\Theta^{(l+1)}$  given  $\Theta^{(l)}$ . A conditional probability density of  $\Theta^{(l+1)}$  given  $\Theta^{(l)}$  is called *transition kernel* of the chain and is usually denoted as  $K(\Theta^{(l)}, \Theta^{(l+1)})$ .

The MCMC approach produces an *ergodic Markov chain* with a *stationary distribution* (which is also a *limiting distribution*). These chains are also *recurrent* and *irreducible*. The precise definitions of these properties are somewhat involved and can be found for example in Robert and Casella [201]. For the purposes of this book we remark as follows:

*Remark 2.16*

- We are interested in the case when the chain stationary distribution corresponds to the posterior density  $\pi(\theta|\mathbf{x})$ .

- The *ergodic* property means that the distribution of  $\Theta^{(l)}$  converges to a *limiting distribution*  $\pi(\theta|\mathbf{x})$  for almost any starting value of  $\Theta^{(0)}$ . Therefore for large  $l$ ,  $\Theta^{(l)}$  is approximately distributed from  $\pi(\theta|\mathbf{x})$  regardless of the starting point. Of course the problem is to decide what is large  $l$ . This can formally be accomplished by running diagnostic tests on the stationarity of the chain.
- A Markov chain is said to have a *stationary distribution* if there is a distribution  $\pi(\theta|\mathbf{x})$  such that if  $\Theta^{(l)}$  is distributed from  $\pi(\theta|\mathbf{x})$  then  $\Theta^{(l+1)}$  is distributed from  $\pi(\theta|\mathbf{x})$  too.
- A Markov chain is *irreducible* if it is guaranteed to visit any set  $\mathcal{A}$  of the support of  $\pi(\theta|\mathbf{x})$ . This property implies that the chain is *recurrent*, i.e. that the average number of visits to an arbitrary set  $\mathcal{A}$  is infinite and even *Harris recurrent*. The latter means that the chain has the same limiting behaviour for *every* starting value rather than *almost every* starting value.
- Markov chains considered in MCMC algorithms are almost always *homogeneous*, i.e. the distribution of  $\Theta^{(l_0+1)}, \Theta^{(l_0+2)}, \dots, \Theta^{(l_0+k)}$  given  $\Theta^{(l_0)}$  is the same as the distribution of  $\Theta^{(1)}, \Theta^{(2)}, \dots, \Theta^{(k)}$  given  $\Theta^{(0)}$  for any  $l_0 \geq 0$  and  $k > 0$ .
- Another important stability property is called *reversibility* that means that the direction of the chain does not matter. That is, the distribution of  $\Theta^{(l+1)}$  conditional on  $\Theta^{(l+2)} = \theta$  is the same as the distribution of  $\Theta^{(l+1)}$  conditional on  $\Theta^{(l)} = \theta$ . The chain is *reversible* if the transition kernel satisfies the *detailed balance condition*:

$$K(\theta, \theta')\pi(\theta|\mathbf{x}) = K(\theta', \theta)\pi(\theta'|\mathbf{x}). \quad (2.54)$$

The detailed balance condition is not necessary but sufficient condition for  $\pi(\theta|\mathbf{x})$  to be stationary density associated with the transitional kernel  $K(\cdot, \cdot)$  that usually can easily be checked for MCMC algorithms.

Of course, the samples  $\Theta^{(1)}, \Theta^{(2)}, \dots$  are not independent. However, the independence is not required if we have to calculate some functionals of  $\pi(\theta|\mathbf{x})$ , because the Ergodic Theorem implies that for large  $L$ , the average

$$\frac{1}{L} \sum_{l=1}^L g(\Theta^{(l)}) \quad (2.55)$$

converges to  $E[g(\Theta)|X = \mathbf{x}]$  (if this expectation is finite), where expectation is calculated with respect to  $\pi(\theta|\mathbf{x})$ .

### 2.11.1 Metropolis-Hastings Algorithm

Metropolis-Hastings algorithm is almost a universal algorithm used to generate a Markov chain with a stationary distribution  $\pi(\theta|\mathbf{x})$ . It has been developed by Metropolis et al. [161] in mechanical physics and generalised by Hastings [116]



in a statistical setting. It can be applied to a variety of problems since it requires the knowledge of the distribution of interest up to a constant only. Given a density  $\pi(\theta|\mathbf{x})$ , known up to a normalisation constant, and a conditional density  $q(\theta^*|\theta)$ , the method generates the chain  $\{\theta^{(1)}, \theta^{(2)}, \dots\}$  using the following algorithm:

**Algorithm 2.1 (Metropolis-Hastings algorithm)**

1. Initialise  $\theta^{(l=0)}$  with any value within a support of  $\pi(\theta|\mathbf{x})$ ;
2. For  $l = 1, \dots, L$ 
  - a. Set  $\theta^{(l)} = \theta^{(l-1)}$ ;
  - b. Generate a proposal  $\theta^*$  from  $q(\theta^*|\theta^{(l)})$ ;
  - c. Accept proposal with the acceptance probability

$$p(\theta^{(l)}, \theta^*) = \min \left\{ 1, \frac{\pi(\theta^*|\mathbf{x})q(\theta^{(l)}|\theta^*)}{\pi(\theta^{(l)}|\mathbf{x})q(\theta^*|\theta^{(l)})} \right\}, \quad (2.56)$$

i.e. simulate  $U$  from the uniform distribution function  $\mathcal{U}(0, 1)$  and set  $\theta^{(l)} = \theta^*$  if  $U < p(\theta^{(l)}, \theta^*)$ . Note that the normalisation constant of the posterior does not contribute here;

3. Next  $l$  (i.e. do an increment,  $l = l + 1$ , and return to step 2).

*Remark 2.17*

- The density  $\pi(\theta|\mathbf{x})$  is called the *target* or *objective density*.
- $q(\theta^*|\theta)$  is called the *proposal density* and will be discussed shortly.

### 2.11.2 Gibbs Sampler

The Gibbs sampler is a technique for generating random variables from a distribution indirectly, without having to calculate the density. The method takes its name from the Gibbs random fields in image-processing models starting with the paper of Geman and Geman [101]. Its roots can be traced back to the 1950s; see Robert and Casella [201] for a brief summary of the early history.

To illustrate the idea of the Gibbs sampler, consider the case of two random variables  $X$  and  $Y$  that have a joint bivariate density  $f(x, y)$ . Assume that simulation of  $X$  from  $f(x)$  cannot be done directly but we can easily sample  $X$  from  $f(x|y)$  and  $Y$  from  $f(y|x)$ . Then, the Gibbs sampler generates samples as follows:

**Algorithm 2.2 (Gibbs sampler, bivariate case)**

1. Initialise  $y^{(l=0)}$  with an arbitrary value within a support of  $Y$ .
2. For  $l = 1, \dots, L$ 
  - a. simulate  $x^{(l)}$  from  $f(x|y^{(l-1)})$ ;

- b. simulate  $y^{(l)}$  from  $f(y|x^{(l)})$ ;
- 3. Next  $l$  (i.e. do an increment,  $l = l + 1$ , and return to step 2).

Under quite general conditions  $f(x, y)$  is a stationary distribution of the chain  $\{(x^{(l)}, y^{(l)}), l = 1, 2, \dots\}$ ; and the chain is ergodic with a limiting distribution  $f(x, y)$ , that is the distribution of  $x^{(l)}$  converges to  $f(x)$  for large  $l$ .

Gibbs sampling can be thought of as a practical implementation of the fact that knowledge of the conditional distributions is sufficient to determine a joint distribution (if it exists!).

The generalisation of the Gibbs sampling to a multidimensional case is as follows. Consider a random vector  $\mathbf{X}$  with a joint density  $f(\mathbf{x})$ . Denote full conditionals  $f_i(x_i|\mathbf{x}_{-i}) = f(x_i|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N)$ . Then, do the following steps:

**Algorithm 2.3 (Gibbs sampler, multivariate case)**

- Initialise  $x_2^{(l=0)}, \dots, x_N^{(l=0)}$  with an arbitrary value.
- For  $l = 1, \dots, L$ 
  - 1) simulate  $x_1^{(l)}$  from  $f_1(x_1|x_2^{(l-1)}, \dots, x_N^{(l-1)})$ ;
  - 2) simulate  $x_2^{(l)}$  from  $f_2(x_2|x_1^{(l)}, x_3^{(l-1)}, \dots, x_N^{(l-1)})$ ;
  - $\vdots$
  - $N$ ) simulate  $x_N^{(l)}$  from  $f_N(x_N|x_1^{(l)}, \dots, x_{N-1}^{(l-1)})$ ;
- Next  $l$ .

Again, under general conditions the joint density  $f(\mathbf{x})$  is a stationary distribution of the generated chain  $\{\mathbf{x}^{(l)}, l = 1, 2, \dots\}$ ; and the chain is ergodic, that is  $f(\mathbf{x})$  is a limiting distribution of the chain.

### 2.11.3 Random Walk Metropolis-Hastings Within Gibbs

The *Random Walk Metropolis-Hastings (RW-MH) within Gibbs* algorithm is easy to implement and often efficient if the likelihood function can be easily evaluated. It is referred to as *single-component Metropolis-Hastings* in Gilks, Richardson and Spiegelhalter ([106], section 1.4). The algorithm is not well known among operational risk practitioners and we would like to mention its main features; see Shevchenko and Temnov [217] for application in the context of operational risk and Peters, Shevchenko and Wüthrich [186] for application in the context of a similar problem in the insurance.

The RW-MH within Gibbs algorithm creates a reversible Markov chain with a stationary distribution corresponding to our target posterior distribution. Denote by  $\boldsymbol{\theta}^{(l)}$  the state of the chain at iteration  $l$ . The algorithm proceeds by proposing to move the  $i$ -th parameter from the current state  $\theta_i^{(l-1)}$  to a new proposed state  $\theta_i^*$  sampled from the MCMC proposal transition kernel. Typically the parameters are restricted by simple ranges,  $\theta_i \in [a_i, b_i]$ , and proposals are sampled from the normal distribution. Then, the logical steps of the algorithm are as follows.

**Algorithm 2.4 (RW-MH within Gibbs)**

1. Initialise  $\theta_i^{(l=0)}$ ,  $i = 1, \dots, I$  by e.g. using MLEs.
2. For  $l = 1, \dots, L$ 
  - a. Set  $\boldsymbol{\theta}^{(l)} = \boldsymbol{\theta}^{(l-1)}$ .
  - b. For  $i = 1, \dots, I$ 
    - i. Sample proposal  $\theta_i^*$  from the transition kernel, e.g. from the truncated normal density

$$f_N^{\text{tr}}(\theta_i^* | \theta_i^{(l)}, \sigma_i) = \frac{f_N(\theta_i^* | \theta_i^{(l)}, \sigma_i)}{F_N(b_i | \theta_i^{(l)}, \sigma_i) - F_N(a_i | \theta_i^{(l)}, \sigma_i)}, \quad (2.57)$$

where  $f_N(x | \mu, \sigma)$  and  $F_N(x | \mu, \sigma)$  are the normal density and its distribution with mean  $\mu$  and standard deviation  $\sigma$ .

- ii. Accept proposal with the acceptance probability

$$p(\boldsymbol{\theta}^{(l)}, \boldsymbol{\theta}^*) = \min \left\{ 1, \frac{\pi(\boldsymbol{\theta}^* | \mathbf{x}) f_N^{\text{tr}}(\theta_i^{(l)} | \theta_i^*, \sigma_i)}{\pi(\boldsymbol{\theta}^{(l)} | \mathbf{x}) f_N^{\text{tr}}(\theta_i^* | \theta_i^{(l)}, \sigma_i)} \right\}, \quad (2.58)$$

where  $\boldsymbol{\theta}^* = (\theta_1^{(l)}, \dots, \theta_{i-1}^{(l)}, \theta_i^*, \theta_{i+1}^{(l-1)}, \dots)$ , i.e. simulate  $U$  from the uniform  $\mathcal{U}(0, 1)$  and set  $\theta_i^{(l)} = \theta_i^*$  if  $U < p(\boldsymbol{\theta}^{(l)}, \boldsymbol{\theta}^*)$ . Note that the normalisation constant of the posterior does not contribute here.

c. Next  $i$

3. Next  $l$ .

This procedure builds a set of correlated samples from the target posterior distribution. One of the most useful asymptotic properties is the convergence of ergodic averages constructed using the Markov chain samples to the averages obtained under the posterior distribution. The chain has to be run until it has sufficiently converged to the stationary distribution (the posterior distribution) and then one obtains samples from the posterior distribution. General properties of this algorithm, including convergence results, can be found in Robert and Casella ([201], sections 6–10).

The RW-MH algorithm is simple in nature and easy to implement. However, for a bad choice of the proposal distribution, the algorithm gives a very slow convergence to the stationary distribution. There have been several recent studies regarding the optimal scaling of the proposal distributions to ensure optimal convergence rates; see Bedard and Rosenthal [24]. The suggested asymptotic acceptance rate optimising the efficiency of the process is 0.234. Usually, it is recommended that the  $\sigma_i$  in (2.57) are chosen to ensure that the acceptance probability is roughly close to 0.234. This requires some tuning of the  $\sigma_i$  prior to the final simulations.

### 2.11.4 ABC Methods

The standard MCMC described above assumes that the likelihood of the data for given model parameters can easily be evaluated. If this is not the case, but synthetic data are easily simulated from the model for given parameters, then the so-called *approximate Bayesian computation* (ABC) methods can be utilised to estimate the model. For example, this is the case when the severity is modelled by the  $\alpha$ -stable or g-and-h distributions that can easily be simulated but the density is not available in closed form. ABC methods are relatively recent developments in computational statistics; see Beaumont, Zhang and Balding [23] and Tavaré, Marjoram, Molitor and Plagnol [234]. For applications in the context of operational risk and insurance; see Peters and Sisson [188], and Peters, Wüthrich and Shevchenko [190].

Consider the data  $\mathbf{X}$  and denote the model parameters by  $\theta$ . Then the posterior from which we wish to draw samples is  $\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$ . The purpose of ABC is to sample from the posterior  $\pi(\theta|\mathbf{x})$  without evaluating computationally intractable  $f(\mathbf{x}|\theta)$ . The logical steps of the simplest ABC algorithm are as follows.

#### Algorithm 2.5 (Rejection Sampling ABC)

1. Choose a small tolerance level  $\epsilon$ .
2. For  $l = 1, 2, \dots$ 
  - a. Draw  $\theta^*$  from the prior  $\pi(\cdot)$ .
  - b. Simulate a synthetic dataset  $\mathbf{x}^*$  from the model given parameters  $\theta^*$ , i.e. simulate from  $f(\cdot|\theta^*)$ .
  - c. Rejection condition: calculate a distance metric  $\rho(\mathbf{x}, \mathbf{x}^*)$  that measures a difference between  $\mathbf{x}$  and  $\mathbf{x}^*$ . Accept the sample, i.e. set  $\theta^{(l)} = \theta^*$  if  $\rho(\mathbf{x}, \mathbf{x}^*) \leq \epsilon$ , otherwise return to step a).
3. Next  $l$ .

It is easy to show that, if the support of the distributions on  $\mathbf{x}$  is discrete and the rejection condition  $\rho(\mathbf{x}, \mathbf{x}^*) \leq \epsilon$  is a simplest condition accepting the proposal only if  $\mathbf{x}^* = \mathbf{x}$ , then the obtained  $\theta^{(1)}, \theta^{(2)}, \dots$  are samples from  $\pi(\theta|\mathbf{x})$ . For general case, the obtained samples  $\theta^{(l)}$ , are from

$$\pi_{ABC}(\boldsymbol{\theta}|\mathbf{x}, \epsilon) \propto \int \pi(\boldsymbol{\theta})\pi(\mathbf{x}^*|\boldsymbol{\theta})g_{\epsilon}(\mathbf{x}|\mathbf{x}^*)d\mathbf{x}^*, \quad (2.59)$$

where the weighting function

$$g_{\epsilon}(\mathbf{x}|\mathbf{x}^*) \propto \begin{cases} 1, & \text{if } \rho(\mathbf{x}, \mathbf{x}^*) \leq \epsilon, \\ 0, & \text{otherwise.} \end{cases} \quad (2.60)$$

As  $\epsilon \rightarrow 0$ , for appropriate choices of distance  $\rho(\cdot, \cdot)$ ,

$$\pi_{ABC}(\boldsymbol{\theta}|\mathbf{x}, \epsilon) \rightarrow \pi(\boldsymbol{\theta}|\mathbf{x}).$$

Of course, for a finite  $\epsilon$  we obtain an approximation for  $\pi(\boldsymbol{\theta}|\mathbf{x})$ .

To improve the efficiency,  $\rho(\mathbf{x}, \mathbf{x}^*)$  is often replaced by  $\rho(S(\mathbf{x}), S(\mathbf{x}^*))$ , where  $S(\mathbf{x})$  is a summary statistic of the data sample. Other weighting functions can be used. In general, the procedure is simple: given a realisation of the model parameters, a synthetic dataset  $\mathbf{x}^*$  is simulated and compared to the original dataset  $\mathbf{x}$ . Then the summary statistic  $S(\mathbf{x}^*)$  is calculated for simulated dataset  $\mathbf{x}^*$  and compared to the summary statistic of the observed data  $S(\mathbf{x})$ ; and a distance  $\rho(S(\mathbf{x}), S(\mathbf{x}^*))$  is calculated. Finally, a greater weight is given to the parameter values producing  $S(\mathbf{x}^*)$  close to  $S(\mathbf{x})$  according to the weighting function  $g_{\epsilon}(\mathbf{x}|\mathbf{x}^*)$ . The obtained sample is from  $\pi_{ABC}(\boldsymbol{\theta}|\mathbf{x}, \epsilon)$  that converges to the target posterior  $\pi(\boldsymbol{\theta}|\mathbf{x})$  as  $\epsilon \rightarrow 0$ , assuming that  $S(\mathbf{x})$  is a *sufficient statistic*<sup>3</sup> and the weighting function converges to a point mass on  $S(\mathbf{x})$ . The tolerance,  $\epsilon$  is typically set as small as possible for a given computational budget. One can calculate the results for subsequently reduced values of  $\epsilon$  until the further reduction does not make material difference for the model outputs. The described ABC can be viewed as a general augmented model

$$\pi(\boldsymbol{\theta}, \mathbf{x}, \mathbf{x}^*) = \pi(\mathbf{x}|\mathbf{x}^*, \boldsymbol{\theta})\pi(\mathbf{x}^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$

where  $\pi(\mathbf{x}|\mathbf{x}^*, \boldsymbol{\theta})$  is replaced by  $g(\mathbf{x}|\mathbf{x}^*)$ .

To improve the performance of ABC algorithm, it can be combined with MCMC producing the stationary distribution  $\pi_{ABC}(\boldsymbol{\theta}|\mathbf{x}, \epsilon)$ . For example, the MCMC-ABC can be implemented as follows.

#### Algorithm 2.6 (MCMC-ABC)

1. Initialise  $\boldsymbol{\theta}^{(l=0)}$ .
2. For  $l = 1, \dots, L$ 
  - a. Draw proposal  $\boldsymbol{\theta}^*$  from the proposal density  $q(\cdot|\boldsymbol{\theta}^{(l-1)})$ .

<sup>3</sup> A sufficient statistic is a function of the dataset  $\mathbf{x}$  which summarises all the available sample information about  $\boldsymbol{\theta}$ ; for a formal definition, see Berger ([27], section 1.7).

- b. Simulate a synthetic dataset  $\mathbf{x}^*$  from the model given parameters  $\boldsymbol{\theta}^*$ .
- c. Accept the proposal with the acceptance probability

$$p(\boldsymbol{\theta}^{(l-1)}, \boldsymbol{\theta}^*) = \min \left\{ 1, \frac{\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(l-1)}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^{(l-1)})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(l-1)})} 1_{\{\rho(S(\mathbf{x}), S(\mathbf{x}^*)) \leq \epsilon\}} \right\},$$

i.e. simulate  $U$  from the uniform  $(0,1)$  and set  $\boldsymbol{\theta}^{(l)} = \boldsymbol{\theta}^*$  if  $U \leq p(\boldsymbol{\theta}^{(l-1)}, \boldsymbol{\theta}^*)$ , otherwise set  $\boldsymbol{\theta}^{(l)} = \boldsymbol{\theta}^{(l-1)}$ . Here,  $1_{\{\cdot\}}$  is a standard indicator function.

3. Next  $l$ .

Various summary statistics of the dataset  $x_1, \dots, x_N$  are used in practice. For example, the statistic  $S(\mathbf{x})$  can be defined as the following vectors:

- $\mathbf{S} = (\tilde{\mu}, \tilde{\sigma})$ , where  $\tilde{\mu}$  and  $\tilde{\sigma}$  are empirical mean and standard deviation of the dataset  $\mathbf{x}$  respectively;
- $\mathbf{S} = (x_1, \dots, x_N)$ , i.e. all data points in the dataset.

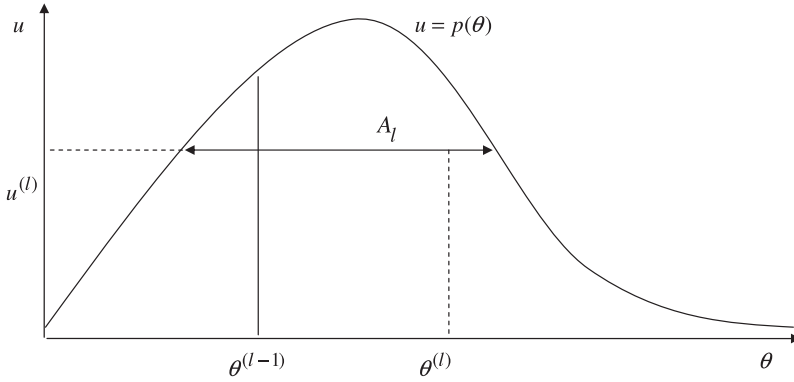
Popular choices for the distance metrics,  $\rho(\mathbf{S}, \mathbf{S}^*)$ , include:

- Euclidean distance:  $\rho(\mathbf{S}, \mathbf{S}^*) = \sum_{l=1}^L (S_l - S_l^*)^2$ ;
- $\mathcal{L}^1$ -distance  $\rho(\mathbf{S}, \mathbf{S}^*) = \sum_{l=1}^L |S_l - S_l^*|$ .

### 2.11.5 Slice Sampling

Often, the full conditional distributions in Gibbs sampler do not take standard explicit closed forms and typically the normalising constants are not known in closed form. Therefore this will exclude straightforward simulation using the inversion method (see Corollary 2.1) or basic rejection sampling (see Corollaries 2.2 and 2.3). In this case, for sampling, one may adopt a Metropolis-Hastings within Gibbs algorithm (described in Sect. 2.11.3). This typically requires tuning of the proposal for a given target distribution that becomes computationally expensive, especially for high dimensional problems. To overcome this problem one may use an adaptive Metropolis-Hastings within Gibbs sampling algorithm; see Atchade and Rosenthal [11] and Rosenthal [205]. An alternative approach, which is more efficient in some cases, is known as a univariate *slice sampler*; see Neal [170]. The latter was developed with the intention of providing a “black box” approach for sampling from a target distribution which may not have a simple form.

A single iteration of the slice sampler algorithm for a toy example is presented in Fig. 2.1. The intuition behind the slice sampling arises from the fact that sampling



**Fig. 2.1** Markov chain created for  $\Theta$  and auxiliary random variable  $U$ ,  $(u^{(1)}, \theta^{(1)}), \dots, (u^{(l-1)}, \theta^{(l-1)}), (u^{(l)}, \theta^{(l)}), \dots$  has a stationary distribution with the desired marginal density  $p(\theta)$

from a univariate density  $p(\theta)$  can always be achieved by sampling uniformly from the region under the density  $p(\theta)$ .

#### Algorithm 2.7 (Univariate slice sampler)

1. Initialise  $\theta^{(0)}$  by any value within the support of  $p(\theta)$ .
2. For  $l = 1, 2, \dots$ 
  - a. Sample a value  $u^{(l)} \sim \mathcal{U}(0, p(\theta^{(l-1)}))$ .
  - b. Sample a value  $\theta^{(l)}$  uniformly from the level set  $A_l = \{\theta : p(\theta) > u^{(l)}\}$ , i.e.  $\theta^{(l)} \sim \mathcal{U}(A_l)$ .
3. Next  $l$ .

By discarding the auxiliary variable sample  $u^{(l)}$ , one obtains correlated samples  $\theta^{(l)}$  from  $p(\cdot)$ . Neal [170], demonstrates that a Markov chain  $(U, \Theta)$  constructed in this way will have a stationary distribution defined by a uniform distribution under  $p(\theta)$  and the marginal of  $\Theta$  has desired stationary density  $p(\theta)$ . Additionally, Mira and Tierney [165] proved that the slice sampler algorithm, assuming a bounded target density  $p(\theta)$  with bounded support, is uniformly ergodic.

There are many approaches that could be used in the determination of the level sets  $A_l$  for the density  $p(\cdot)$ ; see Neal ([170], section 4). For example, one can use a stepping out and a shrinkage procedure; see Neal ([170], p. 713, Figure 1). The basic idea is that given a sampled vertical level  $u^{(l)}$ , the level sets  $A_l$  can be found by positioning an interval of width  $w$  randomly around  $\theta^{(l-1)}$ . This interval is expanded in step sizes of width  $w$  until both ends are outside the slice. Then a new state is obtained by sampling uniformly from the interval until a point in the slice  $A_l$  is obtained. Points that fail can be used to shrink the interval.

Additionally, it is important to note that we only need to know the target full conditional posterior up to normalisation; see Neal ([170], p. 710). To make more precise the intuitive description of the slice sampler presented above, we briefly detail the argument made by Neal on this point. Suppose we wish to sample a random vector  $\Theta$  whose density  $p(\theta)$  is proportional to some function  $f(\theta)$ . This can be achieved by sampling uniformly from the  $(n + 1)$ -dimensional region that lies under the plot of  $f(\theta)$ . This is formalised by introducing the auxiliary random variable  $U$  and defining a joint distribution over  $\Theta$  and  $U$  (which is uniform over the region  $\{(\Theta, U) : 0 < u < f(\theta)\}$  below the surface defined by  $f(\theta)$ ) given by

$$p(\theta, u) = \begin{cases} 1/Z, & \text{if } 0 < u < f(\theta), \\ 0, & \text{otherwise,} \end{cases} \quad (2.61)$$

where  $Z = \int f(\theta) d\theta$ . Then the target marginal density for  $\Theta$  is given by

$$p(\theta) = \int_0^{f(\theta)} \frac{1}{Z} du = \frac{f(\theta)}{Z}, \quad (2.62)$$

as required.

The simplest way to apply the slice sampler in a multivariate case is by applying the univariate slice sampler for each full conditional distribution within the Gibbs sampler, as in the example in Sect. 7.13.1.

## 2.12 MCMC Implementation Issues

There are several numerical issues when implementing MCMC. In practice, a MCMC run consists of three stages: *tuning*, *burn-in* and *sampling* stages. Also, it is important to assess the numerical errors of the obtained estimators due to finite number of MCMC iterations.

### 2.12.1 Tuning, Burn-in and Sampling Stages

*Tuning.* The use of MCMC samples can be very inefficient for an arbitrary chosen proposal distribution. Typically, parameters of a chosen proposal distribution are adjusted to achieve a reasonable acceptance rate for each component. There have been several studies regarding the optimal scaling of proposal distributions to ensure optimal convergence rates. Gelman, Gilks and Roberts [100], Bedard and Rosenthal [24] and Roberts and Rosenthal [202] were the first authors to publish theoretical results for the optimal scaling problem in RW-MH algorithms with Gaussian proposals. For the  $d$ -dimensional target distributions with independent and identically distributed components, the asymptotic acceptance rate optimising the efficiency of the process is 0.234 independent of the target density. Though for most problems the posterior parameters are not independent Gaussian, it provides a practical guide.



There is no need to be very precise in this stage. In practice, the chains with acceptance rate between 0.2 and 0.8 work well. Typically, tuning is easy. In an ad-hoc procedure, one can initialise the proposal distribution parameters with the values corresponding to the proposal with a very small variability; and start the chain. This will lead to a very high acceptance rate. Then run the chain and gradually change the parameters towards the values that correspond to the proposal with a large uncertainty. This will gradually decrease the acceptance rate. Continue this procedure until the acceptance rate is within 0.2–0.8 range. For example, for Gaussian proposal choose a very small standard deviation parameter. Then increase the standard deviation in small steps and measure the average acceptance rate over the completed iterations until the rate is within 0.2–0.8 range. One can apply a reverse procedure, that is start with parameter values corresponding to a very uncertain proposal resulting in a very low acceptance rate. Then gradually change the parameters towards the values corresponding to the proposal with small variability. Many other alternative ways can be used in this spirit.

Gaussian proposals are often useful with the covariance matrix given by (2.43), that is using Gaussian approximation for the posterior, or just MLE observed information matrix (2.36) in the case of constant prior. An alternative approach is to utilise a new class of Adaptive MCMC algorithms recently proposed in the literature; see Atchade and Rosenthal [11], and Rosenthal [204].

*Burn-in stage.* Subject to regularity conditions, the chain converges to the stationary target distribution. The number of iterations required for the chain to converge should be discarded and called *burn-in* iterations. Again, we do not need to identify this quantity precisely. Rough approximations of the order of magnitude work well. Visual inspections of the chain plot is the most commonly used method. If the chain is run for long enough then the impact of these *burn-in* iterations on the final estimates is not material. There are many formal *convergence diagnostics* that can be used to determine the length of *burn-in*; for a review, see Cowles and Carlin [63].

*Sampling stage.* Consider the chain  $\{\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(L)}\}$  and the number of *burn-in* iterations is  $L_b$ . Then,  $\theta^{(L_b+1)}, \theta^{(L_b+2)}, \dots, \theta^{(L)}$  are considered as dependent samples from the target distribution  $\pi(\theta|\mathbf{x})$  and used for estimation purposes. For example,  $E[g(\Theta)|\mathbf{X} = \mathbf{x}]$  is estimated as

$$E[g(\Theta)|\mathbf{X} = \mathbf{x}] = \int g(\theta)\pi(\theta|\mathbf{x})d\theta \approx \frac{1}{L - L_b} \sum_{l=L_b+1}^L g(\theta^{(l)}). \quad (2.63)$$

Typically, when we calculate the posterior characteristics using MCMC samples, we assume that the samples are taken after burn-in and  $L_b$  is dropped in corresponding formulas to simplify notation.

In addition to visual inspection of MCMC, checking that after the burn-in period the samples are mixing well over the support of the posterior distribution, it is useful to monitor the serial correlation of the MCMC samples. For a given chain sample  $\theta_i^{(1)}, \dots, \theta_i^{(L)}$ , the autocorrelation at lag  $k$  is estimated as

$$\widehat{\text{ACF}}[\theta_i, k] = \frac{1}{(L-k)\widehat{s}^2} \sum_{l=1}^{L-k} (\theta_i^{(l)} - \widehat{\mu})(\theta_i^{(l+k)} - \widehat{\mu}), \quad (2.64)$$

where  $\widehat{\mu}$  and  $\widehat{s}^2$  are the mean and variance of a sample  $\theta_i^{(1)}, \dots, \theta_i^{(L)}$ . In well mixed MCMC samples, the autocorrelation falls to near zero quickly and stays near zero at larger lags. It is useful to find a lag  $k^{\max}$  where the autocorrelations seem to have “died out”, that is fallen to near zero (for some interesting discussion on this issue, see for example Kass, Carlin, Gelman and Neal [133]). It is not unusual to choose a  $k_i^{\max}$  for each component such that the autocorrelation at lag  $k_i^{\max}$  has reduced to less than 0.01.

*Example 2.3* To illustrate the above described stages, consider a dataset of the annual counts  $\mathbf{n} = (9, 12, 7, 9)$  simulated from  $Poisson(10)$ . Then, we obtain the chain  $\lambda^{(0)}, \lambda^{(1)}, \dots$  using RW-MH algorithm with the Gaussian proposal distribution for the  $Poisson(\lambda)$  model and constant prior on a very wide range  $[0.1, 100]$ . Figure 2.2 shows the chains in the case of different starting values  $\lambda^{(0)}$  and different standard deviations  $\sigma_{RW}$  of the Gaussian proposal. One can see that after the burn-in stage indicated by the vertical broken line, the chain looks like stationary. Figure 2.2a, b were obtained when  $\sigma_{RW} = \widehat{\text{stdev}}[\widehat{\lambda}^{\text{MLE}}] \approx 1.521$  leading to the acceptance probability approximately 0.7, while Fig. 2.2c, d were obtained when  $\sigma_{RW} = 0.4$  and  $\sigma_{RW} = 30$  leading to the acceptance probability about 0.91 and 0.10 respectively. The MLE was calculated in the usual way as  $\widehat{\text{stdev}}[\widehat{\lambda}^{\text{MLE}}] = (\sum_{i=1}^m n_i / m)^{1/2} / \sqrt{m}$ , where  $m = 4$ . The impact of the value of  $\sigma_{RW}$  is easy to see: the chains on Fig. 2.2c, d are *mixing* slowly (moves slowly around the support of the posterior) while the chains on Fig. 2.2a, b are mixing rapidly. Slow mixing means that much longer chain should be run to get good estimates.

### 2.12.2 Numerical Error

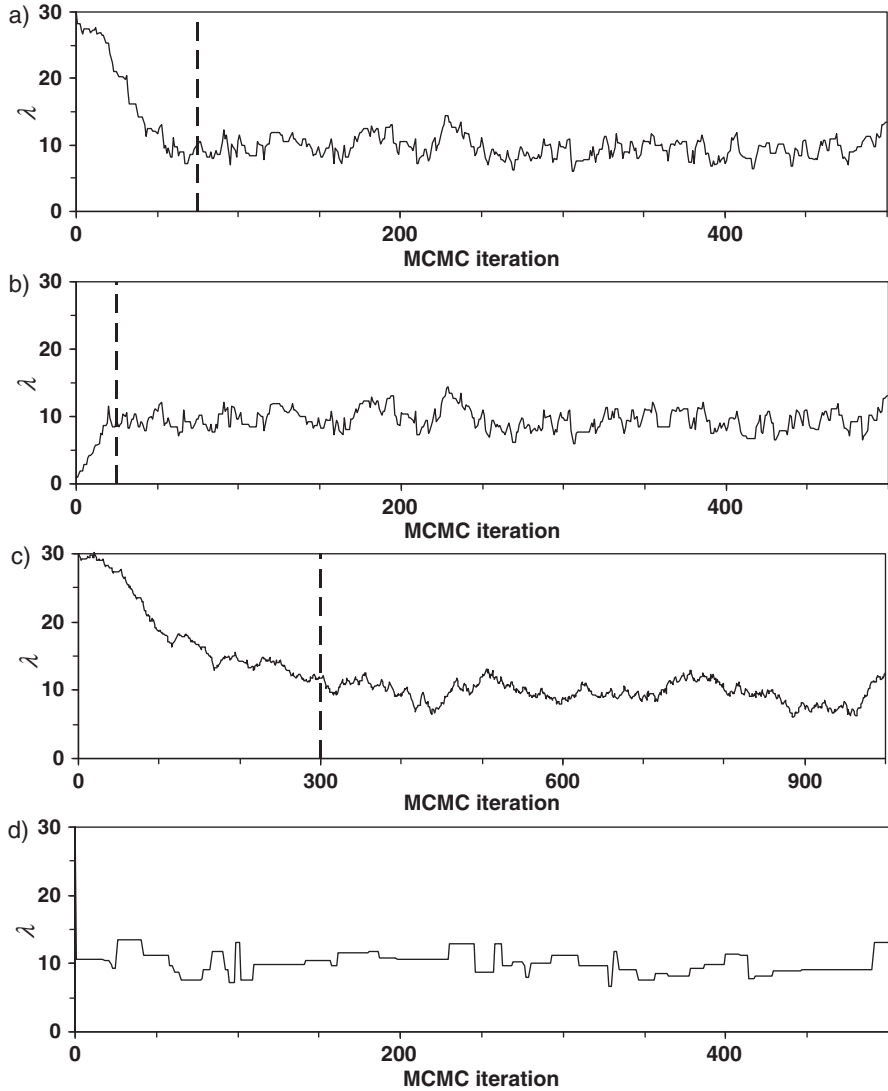
Due to the finite number of iterations, MCMC estimates have numerical error that reduces as the chain length increases. Consider the estimator

$$\widehat{\Omega} = \widehat{\mathbb{E}}[g(\Theta)|\mathbf{X} = \mathbf{x}] = \frac{1}{L} \sum_{l=1}^L g(\Theta^{(l)}). \quad (2.65)$$

If the samples  $\Theta^{(1)}, \dots, \Theta^{(L)}$  are independent and identically distributed then the standard error of  $\widehat{\Omega}$  (due to the finite  $L$ ) is estimated using

$$\text{stdev}[\widehat{\Omega}] = \text{stdev}[g(\Theta)|\mathbf{X} = \mathbf{x}] / \sqrt{L},$$

where  $\text{stdev}[g(\Theta)|\mathbf{X}]$  is estimated by the standard deviation of the sample  $g(\Theta^{(l)})$ ,  $l = 1, \dots, L$ . This formula does not work for MCMC samples due to serial



**Fig. 2.2** MCMC chains of  $\lambda$  parameter of  $Poisson(\lambda)$  model in the case of different starting points  $\lambda^{(0)}$  and different standard deviations of the Gaussian proposal distribution: **(a)** starting point  $\lambda^{(0)} = 30$  and  $\sigma_{RW} = 1.521$ ; **(b)**  $\lambda^{(0)} = 1$  and  $\sigma_{RW} = 1.521$ ; **(c)**  $\lambda^{(0)} = 30$  and  $\sigma_{RW} = 0.4$ ; **(d)**  $\lambda^{(0)} = 30$  and  $\sigma_{RW} = 30$ . The *burn-in* stage is to the left of the vertical broken line. The dataset consisting of the annual number of events (9, 12, 7, 9) over 4 years was simulated from  $Poisson(10)$

correlations between the samples. Of course one can keep every  $k_{\max}$ -th sample from the chain to get approximately independent samples, but it is always a suboptimal approach; see MacEachern and Berliner [152].

*Effective sample size.* If there is only one parameter  $\theta$ , then one of the popular approaches is to calculate *effective sample size*,  $T_{\text{eff}} = T/\tau$ , where  $\tau$  is autocorrelation time

$$\tau = 1 + 2 \sum_{k=1}^{\infty} \text{ACF}[\theta, k]. \quad (2.66)$$

To estimate  $\tau$ , it is necessary to cut off the sum in (2.66) at a value of  $k = k^{\max}$ , where the autocorrelations seem to have fallen to near zero. Then the standard error of the  $\hat{\Omega}$  (2.65) is estimated using

$$\text{stdev}[\hat{\Omega}] = \frac{\text{stdev}[g(\Theta)]}{\sqrt{L/\tau}};$$

see Ripley [199], Neal [168].

*Batch sampling.* Probably the most popular approach to estimate the numerical error of the MCMC posterior averages is a so-called *batch sampling*; see section 3.4.1 in Gilks, Richardson and Spiegelhalter [106]. Consider MCMC posterior samples  $\Theta^{(1)}, \dots, \Theta^{(L)}$  of  $\Theta$  with the length  $L = K \times N$ , and an estimator  $\hat{\Omega} = \sum_{l=1}^L g(\Theta^{(l)})$  of  $E[g(\Theta)]$ . If  $N$  is sufficiently large, the means

$$\hat{\Omega}_j = \frac{1}{N} \sum_{i=(j-1)N+1}^{j \times N} g(\Theta^{(i)}), \quad j = 1, \dots, K \quad (2.67)$$

are approximately independent and identically distributed. Then the overall estimator and its variance are

$$\begin{aligned} \hat{\Omega} &= \frac{1}{K} (\hat{\Omega}_1 + \dots + \hat{\Omega}_K), \\ \text{Var}[\hat{\Omega}] &= \frac{1}{K^2} (\text{Var}[\hat{\Omega}_1] + \dots + \text{Var}[\hat{\Omega}_K]) = \frac{\sigma^2}{K}, \end{aligned}$$

where  $\sigma^2 = \text{Var}[\hat{\Omega}_1] = \dots = \text{Var}[\hat{\Omega}_K]$ . In the limit of large  $K$ , by the central limit theorem (i.e. we also assume that  $\sigma^2$  is finite), the distribution of  $\hat{\Omega}$  is normal with the standard deviation  $\sigma/\sqrt{K}$ . The latter is referred to as the standard error of  $\hat{\Omega}$ . Finally,  $\sigma^2$  can be estimated using sample variance

$$\hat{\sigma}^2 = \frac{1}{K-1} \sum_{j=1}^K (\hat{\Omega}_j - \hat{\Omega})^2. \quad (2.68)$$

Note that  $K$  is the number of quasi-independent bins, and  $N = L/K$  is the size of each bin or batch. Typically, in practice  $K \geq 20$  and  $N \geq 100k^{\max}$ , where  $k^{\max} = \max(k_1^{\max}, k_2^{\max}, \dots)$  is the maximum of the cut-off lags over components. In general, we would like to run the chain until the numerical error is not material. So, one can set  $N$  using  $k^{\max}$  identified during tuning and burning stages, e.g. set  $N = 100k^{\max}$ , then run the chain in batches until the numerical error of the estimates is less than the desired accuracy.

### 2.12.3 MCMC Extensions

Sometimes, in the developed Bayesian models, there is a strong correlation between the model parameters in the posterior. In extreme cases, this can cause slow rates of convergence in the Markov chain to reach the ergodic regime, translating into longer Markov chain simulations. In such a situation several approaches can be tried to overcome this problem.

The first involves the use of a mixture transition kernel combining local and global moves. For example, one can perform local moves via a univariate slice sampler and global moves via an independent Metropolis-Hastings sampler with adaptive learning of its covariance structure. Such an approach is known as a hybrid sampler; see comparisons in Brewer, Aitken and Talbot [36]. Alternatively, for the global move, if determination of level sets in multiple dimensions is not problematic (for the model under consideration), then some of the multivariate slice sampler approaches designed to account for correlation between parameters can be incorporated; see Neal [170] for details.

Another approach to break correlation between parameters in the posterior is via the transformation of the parameter space. If the transformation is effective this will reduce correlation between parameters of the transformed target posterior. Sampling can then proceed in the transformed space, and then samples can be transformed back to the original space. It is not always straightforward to find such transformations.

A third alternative is based on *simulated tempering*, introduced by Marinari and Parisi [153] and discussed extensively in Geyer and Thompson [103]. In particular a special version of simulated tempering, first introduced by Neal [169], can be utilised in which one considers a sequence of target distributions  $\{\pi_l\}$  constructed such that they correspond to the objective posterior in the following way,

$$\pi_l = (\pi(\boldsymbol{\theta}|\mathbf{x}))^{\gamma_l} \quad (2.69)$$

with sequence  $\{\gamma_l\}$ . Then one can use the standard MCMC algorithms (e.g. slice sampler), where  $\pi$  is replaced with  $\pi_l$ .

Running a Markov chain such that at each iteration  $l$  we target the posterior  $\pi_l$  and then only keeping samples from the Markov chain corresponding to situations in which  $\gamma_l = 1$  can result in a significant improvement in exploration around the posterior support. This can overcome slow mixing arising from a univariate

sampling regime. The intuition for this is that for values of  $\gamma_l \ll 1$  the target posterior is almost uniform over the space, resulting in large moves being possible around the support of the posterior. Then as  $\gamma_l$  returns to a value of 1, several iterations later, it will be in potentially new unexplored regions of the posterior support.

For example, one can utilise a sine function,

$$\gamma_l = \min \left( \sin \left( \frac{2\pi}{K} l \right) + 1, 1 \right)$$

with large  $K$  (e.g.  $K = 1,000$ ), which has its amplitude truncated to ensure it ranges between 0 and 1. That is the function is truncated at  $\gamma_l = 1$  for extended iteration periods for our simulation index  $l$  to ensure the sampler spends significant time sampling from the actual posterior distribution.

In the application of tempering one must discard many simulated states of the Markov chain, whenever  $\gamma_l \neq 1$ . There is, however, a computational way to avoid discarding these samples; see Gramacy, Samworth and King [111].

Finally, we note that there are several alternatives to a Metropolis-Hastings within Gibbs sampler such as a basic Gibbs sampler combined with *adaptive rejection sampling* (ARS), Gilks and Wild [107]. Note that ARS requires distributions to be log-concave. Alternatively an adaptive version of this known as the adaptive Metropolis rejection sampler could be used; see Gilks, Best and Tan [105].

## 2.13 Bayesian Model Selection

Consider a model  $M$  with parameter vector  $\theta$ . The model likelihood with data  $\mathbf{x}$  can be found by integrating out the parameter  $\theta$

$$\pi(\mathbf{x}|M) = \int \pi(\mathbf{x}|\theta, M) \pi(\theta|M) d\theta, \quad (2.70)$$

where  $\pi(\theta|M)$  is the prior density of  $\theta$  in  $M$ . Given a set of  $K$  competing models  $(M_1, \dots, M_K)$  with parameters  $\theta_{[1]}, \dots, \theta_{[K]}$  respectively, the Bayesian alternative to traditional hypothesis testing is to evaluate and compare the posterior probability ratio between the models. Assuming we have some prior knowledge about the model probability  $\pi(M_i)$ , we can compute the posterior probabilities for all models using the model likelihoods

$$\pi(M_i|\mathbf{x}) = \frac{\pi(\mathbf{x}|M_i) \pi(M_i)}{\sum_{k=1}^K \pi(\mathbf{x}|M_k) \pi(M_k)}. \quad (2.71)$$

Consider two competing models  $M_1$  and  $M_2$ , parameterised by  $\theta_{[1]}$  and  $\theta_{[2]}$  respectively. The choice between the two models can be based on the posterior model probability ratio, given by

$$\frac{\pi(M_1|\mathbf{x})}{\pi(M_2|\mathbf{x})} = \frac{\pi(\mathbf{x}|M_1) \pi(M_1)}{\pi(\mathbf{y}|M_2) \pi(M_2)} = \frac{\pi(M_1)}{\pi(M_2)} B_{12}, \quad (2.72)$$

where  $B_{12} = \pi(\mathbf{x}|M_1)/\pi(\mathbf{x}|M_2)$  is the Bayes factor, the ratio of the posterior odds of model  $M_1$  to that of model  $M_2$ . As shown by Lavin and Scherrish [142], an accurate interpretation of the Bayes factor is that the ratio  $B_{12}$  captures the change of the odds in favour of model  $M_1$  as we move from the prior to the posterior. Jeffreys [127] recommended a scale of evidence for interpreting the Bayes factors, which was later modified by Wasserman [238]. A Bayes factor  $B_{12} > 10$  is considered strong evidence in favour of  $M_1$ . Kass and Raftery [131] give a detailed review of the Bayes factors.

Typically, the integral (2.70) required by the Bayes factor is not analytically tractable, and sampling based methods must be used to obtain estimates of the model likelihoods. There are quite a few methods in the literature for direct computation of the Bayes factor or indirect construction of the Bayesian model selection criterion, both based on MCMC outputs. The popular methods are direct estimation of the model likelihood thus the Bayes factor; indirect calculation of an asymptotic approximation as the model selection criterion; and direct computation of the posterior model probabilities, as discussed below.

Popular model selection criteria, based on simplifying approximations, include the Deviance information criterion (DIC) and Bayesian information criterion (BIC); see e.g. Robert ([200], chapter 7).

In general, given a set of possible models  $(M_1, \dots, M_K)$ , the model uncertainty can be incorporated in Bayesian framework via considering the joint posterior for the model and the model parameters  $\pi(M_k, \theta_{[k]}|\mathbf{x})$ , where  $\theta_{[k]}$  is a vector of parameters for model  $k$ . Subsequently calculated posterior model probabilities  $\pi(M_k|\mathbf{x})$  can be used to select an optimal model as the model with the largest probability or average over possible models according to the full joint posterior.

Accurate estimation of the required posterior distributions usually involves development of a Reversible Jump MCMC framework. This type of Markov chain sampler is complicated to develop and analyse. It goes beyond the scope of this book but interested reader can find details in Green [112]. In the case of small number of models, Congdon [60] suggests to run a standard MCMC (e.g. RW-MH) for each model separately and use the obtained MCMC samples to estimate  $\pi(M_k|\mathbf{x})$ . It was adopted in Peters, Shevchenko and Wüthrich [186] for modelling claims reserving problem in the insurance. Using the Markov chain results for each model, in the case of equiprobable nested models, this procedure calculates the posterior model probabilities  $\pi(M_i|\mathbf{x})$  as

$$\pi(M_i|\mathbf{x}) = \frac{1}{L} \sum_{l=1}^L \frac{f(\mathbf{x}|M_i, \theta_{[i]}^{(l)})}{\sum_{j=1}^K f(\mathbf{x}|M_j, \theta_{[j]}^{(l)})}, \quad (2.73)$$

where  $\theta_{[i]}^{(l)}$  is the MCMC posterior sample at Markov chain step  $l$  for model  $M_i$ ,  $f(\mathbf{x}|M_i, \theta_{[i]}^{(l)})$  is the joint density of the data  $\mathbf{x}$  given the parameter vector  $\theta_{[i]}^{(l)}$  for model  $M_i$ , and  $L$  is the total number of MCMC steps after burn-in period.

### 2.13.1 Reciprocal Importance Sampling Estimator

Given MCMC samples  $\boldsymbol{\theta}^{(l)}$ ,  $l = 1, \dots, L$  from the posterior distribution obtained through MCMC, Gelfand and Dey [99] proposed the *reciprocal importance sampling estimator* (RISE) to approximate the model likelihood

$$\hat{p}_{RI}(\mathbf{x}) = \left[ \frac{1}{L} \sum_{l=1}^L \frac{h(\boldsymbol{\theta}^{(l)})}{\pi(\mathbf{x}|\boldsymbol{\theta}^{(l)}) \pi(\boldsymbol{\theta}^{(l)})} \right]^{-1}, \quad (2.74)$$

where  $h$  plays the role of an importance sampling density roughly matching the posterior. Gelfand and Dey [99] suggested the multivariate normal or  $t$  distribution density with mean and covariance fitted to the posterior sample.

The RISE estimator can be regarded as a generalisation of the *harmonic mean estimator* suggested by Newton and Raftery [175]. The latter is obtained from the RISE estimator by setting  $h = 1$ . Other estimators include the *bridge sampling* proposed by Meng and Wong [159], and the *Chib's candidate's estimator* in Chib [56]. In a recent comparison study by Miazhyńska and Dorffner [162], these estimators were employed as competing methods for Bayesian model selection on GARCH-type models, along with the reversible jump MCMC. It was demonstrated that the RISE estimator (either with normal or  $t$  importance sampling density), the bridge sampling method, and the Chib's algorithm gave statistically equal performance in model selection. Also, the performance more or less matched the much more involved reversible jump MCMC.

### 2.13.2 Deviance Information Criterion

For a dataset  $\mathbf{X} = \mathbf{x}$  generated by the model with the posterior density  $\pi(\boldsymbol{\theta}|\mathbf{x})$ , define the deviance

$$D(\boldsymbol{\theta}) = -2 \ln \pi(\mathbf{x}|\boldsymbol{\theta}) + C, \quad (2.75)$$

where the constant  $C$  is common to all candidate models. Then the *deviance information criterion* (DIC) is calculated as

$$\begin{aligned} DIC &= 2E[D(\boldsymbol{\Theta})|\mathbf{X} = \mathbf{x}] - D(E[\boldsymbol{\Theta}|\mathbf{X} = \mathbf{x}]) \\ &= E[D(\boldsymbol{\Theta})|\mathbf{X} = \mathbf{x}] + (E[D(\boldsymbol{\Theta})|\mathbf{X} = \mathbf{x}] - D(E[\boldsymbol{\Theta}|\mathbf{X} = \mathbf{x}])), \end{aligned} \quad (2.76)$$

where

- $E[\cdot|\mathbf{X} = \mathbf{x}]$  is the expectation with respect to the posterior density of  $\boldsymbol{\Theta}$ .
- The expectation  $E[D(\boldsymbol{\Theta})|\mathbf{X} = \mathbf{x}]$  is a measure of how well the model fits the data; the smaller this is, the better the fit.
- The difference  $E[D(\boldsymbol{\Theta})|\mathbf{X} = \mathbf{x}] - D(E[\boldsymbol{\Theta}|\mathbf{X} = \mathbf{x}])$  can be regarded as the effective number of parameters. The larger this difference, the easier it is for the model to fit the data.



The DIC criterion favours the model with a better fit but at the same time penalises the model with more parameters. Under this setting the model with the smallest DIC value is the preferred model.

DIC is a Bayesian alternative to BIC (*Schwarz's criterion* and also called the *Bayesian information criterion*, Schwarz [209]) and AIC (*Akaike's information criterion*, Akaike [5]). For more details on the above-mentioned criteria, see e.g. Robert ([200], chapter 7).

## Problems<sup>4</sup>

**2.1 (★)** Given independent and identically distributed data  $N_1, N_2, \dots, N_m$  from  $Poisson(\lambda)$ , find the maximum likelihood estimator  $\hat{\lambda}^{MLE}$  (for parameter  $\lambda$ ) and its variance. Show that this variance is the same as the one obtained from a large sample size normal approximation for MLE.

**2.2 (★ ★ ★)** Suppose there are independent and identically distributed data  $\mathbf{N} = (N_1, \dots, N_m)'$  from  $Poisson(\lambda)$ .

- Find in closed form the mean and variance of the posterior  $\pi(\lambda|\mathbf{N})$ . Compare these with the MLE and its variance calculated in Problem 2.1.
- Simulate Markov chain  $\{\lambda^{(1)}, \lambda^{(2)}, \dots, \lambda^{(L)}\}$  for parameter  $\lambda$  using RW-MH MCMC and dataset  $\mathbf{N}$  as in Example 2.3. Estimate the mean and variance of the chain samples and compare with the above calculated closed form posterior mean and variance. Assume that  $L = 1000$ .

**2.3 (★ ★ ★)** For a Markov chain  $\{\lambda^{(1)}, \lambda^{(2)}, \dots, \lambda^{(L)}\}$ ,  $L = 1000$ , simulated in Problem 2.2, estimate the numerical error of the posterior mean that was estimated using the chain samples. Repeat calculations for  $L = 4 \times 10^3$ ,  $L = 16 \times 10^3$  and compare results.

**2.4 (★★)** Consider random variables  $L_1, \dots, L_J$  and  $L = L_1 + \dots + L_J$ . If risk measure  $\varrho[L]$  is positively homogeneous, i.e.  $\varrho[hZ] = h\varrho[Z]$  for  $h > 0$  and differentiable, show that

$$\varrho[L] = \sum_{j=1}^J \frac{\partial \varrho[L + hL_j]}{\partial h} \Big|_{h=0}. \quad (2.77)$$

**2.5 (★★)** Given three independent risks,  $Z_i \sim \text{Gamma}(\alpha_i, \beta)$ , with  $\alpha_1 = 0.5$ ,  $\alpha_2 = 1$ ,  $\alpha_3 = 1.5$  respectively and the scale parameter  $\beta = 1$ , find:

- the 0.999 VaR for each risk,  $\text{VaR}_{0.999}[Z_i]$ ,  $i = 1, 2, 3$ ;
- the 0.999 VaR of the total risk,  $\text{VaR}_{0.999}[Z_1 + Z_2 + Z_3]$ ; and
- diversification

---

<sup>4</sup> Problem difficulty is indicated by asterisks: (★) – low; (★★) – medium, (★★★) – high.

$$1 - \text{VaR}_{0.999} \left[ \sum_j Z_j \right] / \sum_j \text{VaR}_{0.999}[Z_j].$$

Hint: use the fact that the sum of two independent random variables,  $X_1 \sim \text{Gamma}(\alpha_1, \beta)$  and  $X_2 \sim \text{Gamma}(\alpha_2, \beta)$ , is distributed from  $\text{Gamma}(\alpha_1 + \alpha_2, \beta)$ .

**2.6 (★)** Show that expected shortfall of a continuous random variable  $X$  (see Definition 2.14) can be calculated as

$$\text{ES}_\alpha[X] = \text{E}[X | X \geq \text{VaR}_\alpha[X]].$$

That is, prove Proposition 2.1.

**2.7 (★)** Calculate mean, variance and 0.9 quantile of a random variable  $X$  that has:

- a finite mass at zero,  $\text{Pr}[X = 0] = 0.5$ ; and
- density  $\frac{1}{2}f^{(c)}(x)$  for  $x > 0$ , where  $f^{(c)}(x)$  is the density of the lognormal distribution  $\mathcal{LN}(\mu, \sigma)$  with  $\mu = 0$  and  $\sigma = 1$ .

Compare the results with the case when  $X \sim \mathcal{LN}(0, 1)$ .

**2.8 (★)** Calculate mean, variance, skewness, mode, median and 0.9 quantile of a random variable  $X \sim \text{Pareto}(\xi = 3, x_0 = 1)$ .

**2.9 (★)** Suppose  $X \sim \text{Pareto}(\xi, x_0)$ . Given two quantiles  $q_1$  and  $q_2$  of random variable  $X$  at the confidence levels  $\alpha_1$  and  $\alpha_2$  respectively ( $\alpha_1 \neq \alpha_2$ ), find the distribution parameters  $\xi$  and  $x_0$ .



<http://www.springer.com/978-3-642-15922-0>

Modelling Operational Risk Using Bayesian Inference

Shevchenko, P.V.

2011, XVII, 302 p., Hardcover

ISBN: 978-3-642-15922-0