Part One

BASIC CONCEPTS AND TOOLS

1

Stochastic processes

1.1 Introduction

The theme of this book is Bayesian Analysis of Stochastic Process Models. In this first chapter, we shall provide the basic concepts needed in defining and analyzing stochastic processes. In particular, we shall review what stochastic processes are, their most important characteristics, the important classes of processes that shall be analyzed in later chapters, and the main inference and decision-making tasks that we shall be facing. We also set up the basic notation that will be followed in the rest of the book. This treatment is necessarily brief, as we cover material which is well known from, for example, the texts that we provide in our final discussion.

1.2 Key concepts in stochastic processes

Stochastic processes model systems that evolve randomly in time, space or space-time. This evolution will be described through an index $t \in T$. Consider a random experiment with sample space Ω, endowed with a σ-algebra $\mathcal{F}$ and a base probability measure P. Associating numerical values with the elements of that space, we may define a family of random variables $\{X_t, t \in T\}$, which will be a stochastic process. This idea is formalized in our first definition that covers our object of interest in this book.

Definition 1.1: *A stochastic process $\{X_t, t \in T\}$ is a collection of random variables X_t, indexed by a set T, taking values in a common measurable space S endowed with an appropriate σ-algebra.*

T could be a set of times, when we have a temporal stochastic process; a set of spatial coordinates, when we have a spatial process; or a set of both time and spatial coordinates, when we deal with a spatio-temporal process. In this book, in general,

Bayesian Analysis of Stochastic Process Models, First Edition. David Rios Insua, Fabrizio Ruggeri and Michael P. Wiper.

we shall focus on stochastic processes indexed by time, and will call T the *space of times*. When T is discrete, we shall say that the process is *in discrete time* and will denote time through n and represent the process through $\{X_n,\ n = 0, 1, 2, \ldots\}$. When T is continuous, we shall say that the process is in *continuous time*. We shall usually assume that $T = [0, \infty)$ in this case. The values adopted by the process will be called the *states* of the process and will belong to the *state space* S. Again, S may be either discrete or continuous.

At least two visions of a stochastic process can be given. First, for each $\omega \in \Omega$, we may rewrite $X_t(\omega) = g_\omega(t)$ and we have a function of t which is a realization or a sample function of the stochastic process and describes a possible evolution of the process through time. Second, for any given t, X_t is a random variable. To completely describe the stochastic process, we need a joint description of the family of random variables $\{X_t, t \in T\}$, not just the individual random variables. To do this, we may provide a description based on the joint distribution of the random variables at any discrete subset of times, that is, for any $\{t_1, \ldots, t_n\}$ with $t_1 < \cdots < t_n$, and for any $\{x_1, \ldots, x_n\}$, we provide

$$P\left(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n\right).$$

Appropriate consistency conditions over these finite-dimensional families of distributions will ensure the definition of the stochastic process, via the Kolmogorov extension theorem, as in, for example, Øksendal (2003).

Theorem 1.1: *Let $T \subseteq [0, \infty)$. Suppose that, for any $\{t_1, \ldots, t_n\}$ with $t_1 < \cdots < t_n$, the random variables $X_{t_1}, \ldots, X_{t_n}$ satisfy the following consistency conditions:*

1. *For all permutations π of $1, \ldots, n$ and $x_1, \ldots, x_n$ we have that $P(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n) = P(X_{t_{\pi(1)}} \leq x_{\pi(1)}, \ldots, X_{t_{\pi(n)}} \leq x_{\pi(n)})$.*
2. *For all $x_1, \ldots, x_n$ and $t_{n+1}, \ldots, t_{n+m}$, we have $P(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n) = P(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n, X_{t_{n+1}} < \infty, \ldots, X_{t_{n+m}} < \infty)$.*

Then, there exists a probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and a stochastic process $X_t : T \times \Omega \to \mathbb{R}^n$ having the families $X_{t_1}, \ldots, X_{t_n}$ as finite-dimensional distributions.

Clearly, the simplest case will hold when these random variables are independent, but this is the territory of standard inference and decision analysis. Stochastic processes adopt their special characteristics when these variables are dependent.

Much as with moments for standard distributions, we shall use some tools to summarize a stochastic process. The most relevant are, assuming all the involved moments exist:

Definition 1.2: *For a given stochastic process $\{X_t, t \in T\}$ the mean function is*

$$\mu_X(t) = E[X_t].$$

The autocorrelation function of the process is the function

$$R_X(t_1, t_2) = E[X_{t_1} X_{t_2}].$$

Finally, the autocovariance function of the process is

$$C_X(t_1, t_2) = E[(X_{t_1} - \mu_X(t_1))(X_{t_2} - \mu_X(t_2))].$$

It should be noted that these moments are merely summaries of the stochastic process and do not characterize it, in general.

An important concept is that of a stationary process, that is a process whose characterization is independent of the time at which the observation of the process is initiated.

Definition 1.3: *We say that the stochastic process* $\{X_t, t \in T\}$ *is strictly stationary if for any* n, $t_1, t_2, \ldots, t_n$ *and* τ, $(X_{t_1}, \ldots, X_{t_n})$ *has the same distribution as* $(X_{t_1+\tau}, \ldots, X_{t_n+\tau})$.

A process which does not satisfy the conditions of Definition 1.3 will be called nonstationary. Stationarity is a typical feature of a system which has been running for a long time and has stabilized its behavior.

The required condition of equal joint distributions in Definition 1.3 has important parameterization implications when $n = 1, 2$. In the first case, we have that all X_t variables have the same common distribution, independent of time. In the second case, we have that the joint distribution depends on the time differences between the chosen times, but not on the particular times chosen, that is,

$$F_{X_{t_1}, X_{t_2}}(x_1, x_2) = F_{X_0, X_{t_2 - t_1}}(x_1, x_2).$$

Therefore, we easily see the following.

Proposition 1.1: *For a strictly stationary stochastic process* $\{X_t, t \in T\}$, *the mean function is constant, that is,*

$$\mu_X(t) = \mu_X, \forall t. \tag{1.1}$$

Also, the autocorrelation function of the process is a function of the time differences, that is,

$$R_X(t_1, t_2) = R(t_2 - t_1). \tag{1.2}$$

Finally, the autocovariance function is given by

$$C_X(t_1, t_2) = R(t_2 - t_1) - \mu_X^2,$$

assuming all relevant moments exist.

A process that fulfills conditions (1.1) and (1.2) is commonly known as a weakly stationary process. Such a process is not necessarily strictly stationary, whereas a strictly stationary process will be weakly stationary if first and second moments exist.

Example 1.1: A first-order autoregressive, or AR(1), process is defined through

$$X_n = \phi_0 + \phi_1 X_{n-1} + \epsilon_n,$$

where ϵ_n is a sequence of independent and identically distributed (IID) normal random variables with zero mean and variance σ^2. This process is weakly, but not strictly, stationary if $|\phi_1| < 1$. Then, we have $\mu_X = \phi_0 + \phi_1 \mu_X$, which implies that $\mu_X = \frac{\phi_0}{1-\phi_1}$. If $|\phi_1| \geq 1$, the process is not stationary. △

When dealing with a stochastic process, we shall sometimes be interested in its transition behavior, that is, given some observations of the process, we aim at forecasting some of its properties a certain time t ahead in the future. To do this, it is important to provide the so called *transition functions*. These are the conditional probability distributions based on the available information about the process, relative to a specific value of the parameter t_0.

Definition 1.4: *Let $t_0, t_1 \in T$ be such that $t_0 \leq t_1$. The* conditional transition distribution function *is defined by*

$$F(x_0, x_1; t_0, t_1) = P\left(X_{t_1} \leq x_1 \mid X_{t_0} \leq x_0\right).$$

When the process is discrete in time and space, we shall use the transition probabilities defined, for $m \leq n$, through

$$P_{ij}^{(m,n)} = P(X_n = j \mid X_m = i).$$

When the process is stationary, the transition distribution function will depend only on the time differences $t = t_1 - t_0$,

$$F(x_0, x; t_0, t_0 + t) = F(x_0, x; 0, t), \ \forall t_0 \in T.$$

For convenience, the previous expression will sometimes be written as $F(x_0, x; t)$. Analogously, for the discrete process $\{X_n\}_n$ we shall use the expression $P_{ij}^{(n)}$.

Letting $t \to \infty$, we may consider the long-term limiting behavior of the process, typically associated with the stationary distribution. When this distribution exists, computations are usually much simpler than doing short-term predictions based on the use of the transition functions. These limit distributions reflect a parallelism with

the laws of large numbers, for the case of IID observations, in that

$$\frac{1}{n}\sum_{i=1}^{n} X_{t_i} \to E[X_\infty]$$

when $t_n \to \infty$, for some limiting random variable X_∞. This is the terrain of ergodic theorems and ergodic processes, see, e.g., Walters (2000).

In particular, for a given stochastic process, we may be interested in studying the so-called time averages. For example, we may define the mean time average, which is the random variable defined by

$$\mu_X(T) = \frac{1}{T}\int_0^T X_t \, \mathrm{d}t.$$

If the process is stationary, interchanging expectation with integration, we have

$$E[\mu_X(T)] = \frac{1}{T}E\left[\int_0^T X_t \mathrm{d}t\right] = \frac{1}{T}\int_0^T E[X_t]\mathrm{d}t = \frac{1}{T}\int_0^T \mu_X = \mu_X.$$

This motivates the following definition.

Definition 1.5: *The process X_t is said to be mean ergodic if:*

1. $\mu_X(T) \to \mu_X$, *for some* μ_X, *and*
2. $var(\mu_X(T)) \to 0$.

An autocovariance ergodic process can be defined in a similar way. Clearly, for a stochastic process to be ergodic, it has to be stationary. The converse is not true.

1.3 Main classes of stochastic processes

Here, we define the main types of stochastic processes that we shall study in this book. We start with Markov chains and Markov processes, which will serve as a model for many of the other processes analyzed in later chapters and are studied in detail in Chapters 3 and 4.

1.3.1 Markovian processes

Except for the case of independence, the simplest dependence form among the random variables in a stochastic process is the Markovian one.

Definition 1.6: *Consider a set of time instants* $\{t_0, t_1, \ldots, t_n, t\}$ *with* $t_0 < t_1 < \cdots < t_n < t$ *and* $t, t_i \in T$. *A stochastic process* $\{X_t, t \in T\}$ *is Markovian if the distribution*

of X_t conditional on the values of $X_{t_1}, \ldots, X_{t_n}$ depends only on X_{t_n}, that is, the most recent known value of the process

$$\begin{aligned} &P\left(X_t \le x \mid X_{t_n} \le x_n, X_{t_{n-1}} \le x_{n-1}, \ldots, X_{t_0} \le x_0\right) \\ &\quad = P\left(X_t \le x \mid X_{t_n} \le x_n\right) = F\left(x_n, x; t_n, t\right). \end{aligned} \tag{1.3}$$

As a consequence of the previous relation, we have

$$F\left(x_0, x; t_0, t_0 + t\right) = \int_{y \in S} F\left(y, x; \tau, t\right) \mathrm{d}F\left(x_0, y; t_0, \tau\right) \tag{1.4}$$

with $t_0 < \tau < t$.

If the stochastic process is discrete in both time and space, then (1.3) and (1.4) adopt the following form: For $n > n_1 > \cdots > n_k$, we have

$$P\left(X_n = j \mid X_{n_1} = i_1, X_{n_2} = i_2, \ldots, X_{n_k} = i_{n_k}\right) = P\left(X_n = j \mid X_{n_1} = i_1\right) = p_{i_1 j}^{(n_1, n)}.$$

Using this property and taking r such that $m < r < n$, we have

$$\begin{aligned} p_{ij}^{(m,n)} &= P\left(X_n = j \mid X_m = i\right) \\ &= \sum_{k \in S} P\left(X_n = j \mid X_r = k\right) P\left(X_r = k \mid X_m = i\right). \end{aligned} \tag{1.5}$$

Equations (1.4) and (1.5) are called the *Chapman–Kolmogorov equations* for the continuous and discrete cases, respectively. In this book we shall refer to discrete state space Markov processes as Markov chains and will use the term Markov process to refer to processes with continuous state spaces and the Markovian property.

Discrete time Markov chains

Markov chains with discrete time space are an important class of stochastic processes whose analysis serves as a guide to the study of other more complex processes. The main features of such chains are outlined in the following text. Their full analysis is provided in Chapter 3.

Consider a discrete state space Markov chain, $\{X_n\}$. Let $p_{ij}^{(m,n)}$ be defined as in (1.5), being the probability that the process is at time n in j, when it was in i at time m. If $n = m + 1$, we have

$$p_{ij}^{(m,m+1)} = P\left(X_{m+1} = j \mid X_m = i\right),$$

which is known as the one-step *transition probability*. When $p_{ij}^{(m,m+1)}$ is independent of m, the process is stationary and the chain is called *time homogeneous*. Otherwise,

the process is called time inhomogeneous. Using the notation

$$p_{ij} = P(X_{m+1} = j \mid X_m = i)$$
$$p_{ij}^n = P(X_{n+m} = j \mid X_m = i)$$

for every m, the Chapman–Kolmogorov equations are now

$$p_{ij}^{n+m} = \sum_{k \in S} p_{ik}^n p_{kj}^m \tag{1.6}$$

for every $n, m \geq 0$ and i, j. The *n-step transition probability matrix* is defined as $\mathbf{P}^{(n)}$, with elements p_{ij}^n. Equation (1.6) is written $\mathbf{P}^{(n+m)} = \mathbf{P}^{(n)} \cdot \mathbf{P}^{(m)}$. These matrices fully characterize the transition behavior of an homogeneous Markov chain. When $n = 1$, we shall usually write $\mathbf{P}$ instead of $\mathbf{P}^{(1)}$ and shall refer to the *transition matrix* instead of the one-step transition matrix.

Example 1.2: A famous problem in stochastic processes is the gambler's ruin problem. A gambler with an initial stake, $x_0 \in \mathbb{N}$, plays a coin tossing game where at each turn, if the coin comes up heads, she wins a unit and if the coin comes up tails, she loses a unit. The gambler continues to play until she either is bankrupted or her current holdings reach some fixed amount m. Let X_n represent the amount of money held by the gambler after n steps. Assume that the coin tosses are IID with probability of heads p at each turn. Then, $\{X_n\}$ is a time homogeneous Markov chain with $p_{00} = p_{mm} = 1$, $p_{ii+1} = p$ and $p_{ii-1} = 1 - p$, for $i = 1, \ldots, m-1$ and $p_{ij} = 0$ for $i \in \{0, \ldots, m\}$ and $j \neq i$. △

The analysis of the stationary behavior of an homogeneous Markov chain requires studying the relations among states as follows.

Definition 1.7: *A state j is reachable from a state i if $p_{ij}^n > 0$, for some n. We say that two states that are mutually reachable, communicate, and belong to the same communication class.*

If all states in a chain communicate among themselves, so that there is just one communication class, we shall say that the Markov chain is irreducible. In the case of the gambler's ruin problem of Example 1.2, we can see that there are three communication classes: $\{0\}$, $\{1, \ldots, m-1\}$, and $\{m\}$.

Definition 1.8: *Given a state i, let p_i be the probability that, starting from state i, the process returns to such state. We say that state i is recurrent if $p_i = 1$ and transitory if $p_i < 1$.*

We may easily see that if state i is recurrent and communicates with another state j, then j is recurrent. In the case of gambler's ruin, only the states $\{0\}$ and $\{m\}$ are recurrent.

Definition 1.9: *A state i has* period k *if $p_{ii}^n = 0$ whenever n is not divisible by k and k is the biggest integer with this property. A state with period one is aperiodic.*

We may also see easily that if i has period k and states i and j communicate, then state j has period k. In the gambler's ruin problem, states $\{0, m\}$ are aperiodic and the remaining states have period two.

Definition 1.10: *A state i is positive recurrent if, starting at i, the expected time until return to i is finite.*

Positive recurrence is also a class property in the sense that, if i is positively recurrent and states i and j communicate, then state j is also positively recurrent. We may also prove that in a Markov chain with a finite number of states all recurrent states are positive recurrent. The final key definition is the following.

Definition 1.11: *A positive recurrent, aperiodic state is called ergodic.*

We then have the following important limiting result for a Markov chain, whose proof may be seen in, for example, Ross (1995).

Theorem 1.2: *For an ergodic and irreducible Markov chain, then $\pi_j = \lim_{n\to\infty} p_{ij}^n$, which is independent of i π_j is the unique nonnegative solution of $\pi_j = \sum_i \pi_i p_{ij}$, $j \geq 0$, with $\sum_{i=0}^{\infty} \pi_i = 1$.*

Continuous time Markov chains

Here, we describe only the homogeneous case. Continuous time Markov chains are stochastic processes with discrete-state space and continuous space time such that whenever a system enters in state i, it remains there for an exponentially distributed time with mean $1/\lambda_i$, and when it abandons this state, it goes to state $j \neq i$ with probability p_{ij}, where $\sum_{j\neq i} p_{ij} = 1$.

The required transition and limited behavior of these processes and some generalizations are presented in Chapter 4.

1.3.2 Poisson process

Poisson processes are continuous time, discrete space processes that we shall analyze in detail in Chapter 5. Here, we shall distinguish between homogeneous and nonhomogeneous Poisson processes.

Definition 1.12: *Suppose that the stochastic process $\{X_t\}_{t\in T}$ describes the number of events of a certain type produced until time t and has the following properties:*

1. *The number of events in nonoverlapping intervals are independent.*
2. *There is a constant λ such that the probabilities of occurrence of events over 'small' intervals of duration Δt are:*
   - $P(\text{number of events in } (t, t+\Delta t] = 1) = \lambda \Delta t + o(\Delta t)$.
   - $P(\text{number of events in } (t, t+\Delta t] > 1) = o(\Delta t)$, *where* $o(\Delta t)$ *is such that* $o(\Delta t)/\Delta t \to 0$ *when* $\Delta t \to 0$.

Then, we say that $\{X_t\}$ is an homogeneous Poisson process with parameter λ, characterized by the fact that $X_t \sim Po(\lambda t)$.

For such a process, it can be proved that the times between successive events are IID random variables with distribution $\text{Ex}(\lambda)$.

The Poisson process is a particular case of many important generic types of processes. Among others, it is an example of a renewal process, that is, a process describing the number of events of a phenomenon of interest occurring until a certain time such that the times between events are IID random variables (exponential in the case of the Poisson process). Poisson processes are also a special case of continuous time Markov chains, with transition probabilities $p_{i,i+1} = 1, \forall i$ and $\lambda_i = \lambda$.

Nonhomogeneous Poisson processes

Nonhomogeneous Poisson processes are characterized by the intensity function $\lambda(t)$ or the mean function $m(t) = \int_0^t \lambda(s)\text{d}s$; we consider, in general, a time-dependent intensity function but it could be space and space-time dependent as well. Note that, when $\lambda(t) = \lambda$, we have an homogeneous Poisson process. For a nonhomogeneous Poisson process, the number of events occurring in the interval $(t, t+s]$ will have a $\text{Po}(m(t+s) - m(t))$ distribution.

1.3.3 Gaussian processes

The Gaussian process is continuous in both time and state spaces. Let $\{X_t\}$ be a stochastic process such that for any n times $\{t_1, t_2, \ldots, t_n\}$ the joint distribution of $X_{t_i}, i = 1, 2, .., n$, is n-variate normal. Then, the process is *Gaussian*. Moreover, if for any finite set of time instants $\{t_i\}$, $i = 1, 2, \ldots$ the random variables are mutually independent and X_t is normally distributed for every t, we call it a *purely random Gaussian process*.

Because of the specific properties of the normal distribution, we may easily specify many properties of a Gaussian process. For example, if a Gaussian process is weakly stationary, then it is strictly stationary.

1.3.4 Brownian motion

This continuous time and state-space process has the following properties:

1. The process $\{X_t, t \geq 0\}$ has independent, stationary increments: for $t_1, t_2 \in T$ and $t_1 < t_2$, the distribution of $X_{t_2} - X_{t_1}$ is the same of $X_{t_2+h} - X_{t_1+h}$ for every $h > 0$

and, for nonoverlapping intervals (t_1, t_2) and (t_3, t_4), with $t_1 < t_2 < t_3 < t_4$, the random variables $X_{t_2} - X_{t_1}$ and $X_{t_4} - X_{t_3}$ are independent.

2. For any time interval (t_1, t_2), the random variable $X_{t_2} - X_{t_1}$ has distribution $N(0, \sigma^2(t_2 - t_1))$.

1.3.5 Diffusion processes

Diffusion processes are Markov processes with certain continuous path properties which emerge as solution of stochastic differential equations. Specifically,

Definition 1.13: *A continuous time and state process is a diffusion process if it is a Markov process* $\{X_t\}$ *with transition density* $p(s, t; x, y)$ *such that there are two functions* $\mu(t, x)$ *and* $\beta^2(t, x)$*, known as the drift and the diffusion coefficients, such that*

$$\int_{|x-y|\leq\epsilon} p(t, t + \Delta t; x, y)dy = o(\Delta t),$$

$$\int_{|x-y|\leq\epsilon} (y - x)p(t, t + \Delta t; x, y)dy = \mu(t, x) + o(\Delta t),$$

$$\int_{|x-y|\leq\epsilon} (y - x)^2 p(t, t + \Delta t; x, y)dy = \beta^2(t, x) + o(\Delta t).$$

The previous three types of processes are dealt with in Chapter 6.

1.4 Inference, prediction, and decision-making

Given the key definitions and results concerning stochastic processes, we can now informally set up the statistical and decision-making problems that we shall deal with in the following chapters.

Clearly, stochastic processes will be characterized by their initial value and the values of their parameters, which may be finite or infinite dimensional.

Example 1.3: In the case of the gambler's ruin problem of Example 1.2 the process is parameterized by p, the probability of heads. More generally, for a stationary finite Markov chain model with states $1, 2, \ldots, k$, the parameters will be the transition probabilities $(p_{11}, \ldots., p_{k,k})$, where p_{ij} satisfy that $p_{ij} \geq 0$ and $\sum_j p_{ij} = 1$.

The AR(1) process of Example 1.1 is parameterized through the parameters ϕ_0 and ϕ_1.

A nonhomogeneous Poisson process with intensity function $\lambda(t) = M\beta t^{\beta-1}$, corresponding to a Power Law model, is a finite parametric model with parameters (M, β).

A normal dynamic linear model (DLM) with univariate observations X_n, is described by

$$\theta_0|D_0 \sim \mathrm{N}(m_0, C_0)$$
$$\theta_n|\theta_{n-1} \sim \mathrm{N}(\boldsymbol{G}_n\theta_{n-1}, \boldsymbol{W}_n)$$
$$X_n|\theta_n \sim \mathrm{N}(F_n'\theta_n, V_n)$$

where, for each n, F_n is a known vector of dimension $m \times 1$, $\boldsymbol{G}_n$ is a known $m \times m$ matrix, V_n is a known variance, and $\boldsymbol{W}_n$ is a known $m \times m$ variance matrix. The parameters are now $\{\theta_0, \theta_1, \ldots\}$. △

Inference problems for stochastic processes are stated as follows. Assume we have observations of the stochastic process, which will typically be observations $X_{t_1}, \ldots, X_{t_n}$ at time points $t_1, \ldots, t_n$. Sometimes we could have continuous observations in terms of one, or more, trajectories within a given interval. Our aim in inference is then to summarize the available information about these parameters so as to provide point or set estimates or test hypotheses about them. It is important to emphasize that this available information comes from both the observed data and any available prior information.

More important in the context of stochastic processes is the task of forecasting the future behavior of the process, in both the transitory and limiting cases, that is, at a fixed future time and in the long term, respectively.

We shall also be interested in several decision-making problems in relation with stochastic processes. Typically, they will imply making a decision from a set of available ones, once we have taken the process observations. A reward will be obtained depending on the decision made and the future behavior of the process. We aim at obtaining the optimal solution in some sense.

This book explores how the problems of inference, forecasting, and decision-making with underlying stochastic processes may be dealt with using Bayesian techniques. In the following chapter, we review the most important features of the Bayesian approach, concentrating on the standard IID paradigm while in the later chapters, we concentrate on the analysis of some of the specific stochastic processes outlined earlier in Section 1.3.

1.5 Discussion

In this chapter, we have provided the key results and definitions for stochastic processes that will be needed in the rest of this book. Most of these results are of a probabilistic nature, as is usual in the majority of books in this field. Many texts provide very complete outlines of the probabilistic aspects of stochastic processes. For examples, see Karlin and Taylor (1975, 1981), Ross (1995), and Lawler (2006), to name a few.

There are also texts focusing on some of the specific processes that we have mentioned. For example, Norris (1998) or Ching and Ng (2010) are full-length books on Markov chains; Stroock (2005) deals with Markov processes; Poisson processes are studied in Kingman (1993); Rasmussen and Williams (2005) study Gaussian processes, whereas diffusions are studied by Rogers and Williams (2000a, 2000b).

As we have observed previously, there is less literature dedicated to inference for stochastic processes. A quick introduction may be seen in Lehoczky (1990) and both Bosq and Nguyen (1996), and Bhat and Miller (2002) provide applied approaches very much in the spirit of this book, although from a frequentist point of view. Prabhu and Basawa (1991), Prakasa Rao (1996), and Rao (2000) are much more theoretical.

Finally, we noted earlier that the index T of a stochastic process need not always be a set of times. Rue and Held (2005) illustrate the case of spatial processes, when T is a spatial set.

References

Bhat, U.N. and Miller, G. (2002). *Elements of Applied Stochastic Processes* (3rd edn.). New York: John Wiley & Sons, Inc.

Bosq, D. and Nguyen, H.T. (1996) *A Course in Stochastic Processes: Stochastic Models and Statistical Inference*. Dordrecht: Kluwer.

Ching, W. and Ng, N.K. (2010) *Markov Chains: Models, Algorithms and Applications*. Berlin: Springer.

Karlin, S. and Taylor, H.M. (1975) *A First Course in Stochastic Processes* (2nd edn.). New York: Academic Press.

Karlin, S. and Taylor, H.M. (1981) *A Second Course in Stochastic Processes*. New York: Academic Press.

Kingman, J.F.C. (1993) *Poisson Processes*. Oxford: Oxford University Press.

Lawler, G.F. (2006) *Introduction to Stochastic Processes* (2nd edn.). New York: Chapman and Hall.

Lehoczky, J. (1990) Statistical methods. In *Stochastic Models*, D.P. Heyman and M.J. Sobel (Eds.). Amsterdam: North-Holland.

Norris, J.R. (1998) *Markov Chains*. Cambridge: Cambridge University Press.

Øksendal, B. (2003) *Stochastic Differential Equations: An Introduction with Applications*. Berlin: Springer.

Prabhu, N.U. and Basawa, I.V. (1991) *Statistical Inference in Stochastic Processes*. New York: Marcel Dekker.

Prakasa Rao, B.L.S. (1996) *Stochastic Processes and Statistical Inference*. New Delhi: New Age International.

Rao, M.M. (2000) *Stochastic Processes: Inference Theory*. Dordrecht: Kluwer.

Rasmussen, C.E. and Williams, C.K.I. (2005) *Gaussian Processes for Machine Learning*. Cambridge, MA: The MIT Press.

Rogers, L.C.G. and Williams, D. (2000a) *Diffusions, Markov Processes and Martingales: Volume 1 Foundations*. Cambridge: Cambridge University Press.

Rogers, L.C.G. and Williams, D. (2000b) *Diffusions, Markov Processes and Martingales: Volume 2 Ito Calculus*. Cambridge: Cambridge University Press.

Ross, S. (1995) *Stochastic Processes*. New York: John Wiley & Sons, Inc.

Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton: Chapman and Hall.

Stroock, D.W. (2005) *An Introduction to Markov Processes*. Berlin: Springer.

Walters, P. (2000) *Introduction to Ergodic Theory*. Berlin: Springer.

2

Bayesian analysis

2.1 Introduction

In this chapter, we briefly address the first part of this book's title, that is, Bayesian Analysis, providing a summary of the key results, methods and tools that are used throughout the rest of the book. Most of the ideas are illustrated through several worked examples showcasing the relevant models. The chapter also sets up the basic notation that we shall follow later on.

In the last few years numerous books dealing with various aspects of Bayesian analysis have been published. Some of the most relevant literature is referenced in the discussion at the end of this chapter. However, in contrast to the majority of these books, and given the emphasis of our later treatment of stochastic processes, we shall here stress two issues that are central to our book, that is, decision-making and computational issues.

The chapter is organized as follows. First, in Section 2.2 we outline the basics of the Bayesian approach to inference, estimation, hypothesis testing, and prediction. We also consider briefly problems of sensitivity to the prior distribution and the use of noninformative prior distributions. In Section 2.3, we outline Bayesian decision analysis. Then, in Section 2.4, we briefly review Bayesian computational methods. We finish with a discussion in Section 2.5.

2.2 Bayesian statistics

The Bayesian framework for inference and prediction is easily described. Indeed, at a conceptual level, one of the major advantages of the Bayesian approach is the ease with which the basic ideas are put into place.

In particular, one of the typical goals in statistics is to learn about one (or more) parameters, say $\boldsymbol{\theta}$, which describe a stochastic phenomenon of interest. To learn about $\boldsymbol{\theta}$, we will observe the phenomenon, collect a sample of data, say $\mathbf{x} = (x_1, x_2, \ldots, x_n)$

Bayesian Analysis of Stochastic Process Models, First Edition. David Rios Insua, Fabrizio Ruggeri and Michael P. Wiper.

and calculate the conditional density or probability function of the data given $\boldsymbol{\theta}$, which we denote as $f(\mathbf{x}|\boldsymbol{\theta})$. This joint density, when thought of as a function of $\boldsymbol{\theta}$, is usually referred to as the likelihood function and will be, in general, denoted as $l(\theta|\mathbf{x})$, or $l(\theta|\text{data})$ when notation gets cumbersome. Although this will not always be the case in this book, due to the inherent dependence in data generated from stochastic processes, in order to illustrate the main ideas of Bayesian statistics, in this chapter we shall generally assume $\mathbf{X} = (X_1, \ldots, X_n)$ to be (conditionally) independent and identically distributed (CIID) given $\boldsymbol{\theta}$.

As well as the likelihood function, the Bayesian approach takes into account another source of information about the parameters $\boldsymbol{\theta}$. Often, an analyst will have access to external sources of information such as expert information, possibly based on past experience or previous related studies. This external information is incorporated into a Bayesian analysis as the prior distribution, $f(\boldsymbol{\theta})$.

The prior and the likelihood can be combined via Bayes' theorem which provides the posterior distribution $f(\boldsymbol{\theta}|\mathbf{x})$, that is the distribution of the parameter $\boldsymbol{\theta}$ given the observed data $\mathbf{x}$,

$$f(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})}{\int f(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})d\boldsymbol{\theta}} \propto f(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta}). \tag{2.1}$$

The posterior distribution summarizes all the information available about the parameters and can be used to solve all standard statistical problems, like point and interval estimation, hypothesis testing or prediction. Throughout this chapter, we shall use the following two examples to illustrate these problems.

Example 2.1: Following Ríos Insua *et al.* (1997), we are interested in modeling the logarithm, X_i, of inflows to a reservoir in a given month. Suppose that $X_1, \ldots, X_n$ are CIID $\mathrm{N}(\theta, \sigma^2)$, given θ, σ^2, where σ^2 is assumed known. In the absence of prior information, we might use an improper, uniform prior for θ, that is $f(\theta) \propto 1$. Simple computations show that the posterior distribution is

$$f(\theta \mid \mathbf{x}) \propto \exp\left(-\frac{n}{2}\left(\frac{\theta^2}{\sigma^2} - 2\frac{\theta\bar{x}}{\sigma^2}\right)\right)$$

and, therefore,

$$\theta \mid \mathbf{x} \sim \mathrm{N}\left(\bar{x}, \frac{\sigma^2}{n}\right), \tag{2.2}$$

where $\bar{x} = \sum x_i / n$. Assume now that prior information was available and could be modeled as $\theta \sim \mathrm{N}(\mu_0, \sigma_0^2)$. Then, it can be easily shown that

$$f(\theta \mid \mathbf{x}) \propto \exp\left(-\frac{1}{2}\theta^2\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right) - 2\theta\left(\frac{\sum x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)\right)$$

and, consequently,

$$\theta \mid \mathbf{x} \sim \mathrm{N}\left(\frac{n\bar{x}/\sigma^2 + \mu_0/\sigma_0^2}{n/\sigma^2 + 1/\sigma_0^2}, \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1}\right).$$

Note that when we let the prior variance σ_0^2 approach infinity, then the prior distribution approaches a uniform distribution and the posterior distribution then approaches distribution that of (2.2). △

Example 2.2: Consider the gambler in Example 1.2. Suppose that she does not know the probability, $(\boldsymbol{\theta} =)p$, that the coin comes up heads. Although she believes that the coin is probably unbiased, she has some uncertainty about this. She represents such uncertainty by setting a symmetric, beta prior distribution, centered at 0.5, for example $p \sim \mathrm{Be}(5, 5)$. Assume also that before she plays the game seriously, she is offered the chance to observe 12 tosses of the coin without cost. Suppose that in these 12 tosses, the gambler observes nine heads and three tails. Then, her posterior distribution is

$$\begin{aligned} f(p|\mathbf{x}) &\propto f(\mathbf{x}|p)f(p) \quad \text{from (2.1)} \\ &\propto \binom{12}{9} p^9(1-p)^3 \frac{1}{B(5,5)} p^{5-1}(1-p)^{5-1} \\ &\propto p^{14-1}(1-p)^{8-1}. \end{aligned}$$

Therefore,

$$p|\mathbf{x} \sim \mathrm{Be}(14, 8).$$

Figure 2.1 illustrates the relative influence of the prior distribution and the likelihood function on the gambler's posterior density. In particular, it shows the prior density, the likelihood function scaled to integrate to 1, that is,

$$\frac{l(p|\mathbf{x})}{\int_0^1 l(p|\mathbf{x})\,\mathrm{d}p} = \frac{1}{B(10,4)} p^{10-1}(1-p)^{4-1},$$

which is a Be(10, 4) density function, and the posterior density. It can be observed that the posterior distribution is a combination of the prior and the likelihood. △

2.2.1 Parameter estimation

As an example of usage of the posterior distribution, we may be interested in point estimation. This is typically addressed by summarizing the distribution through, either

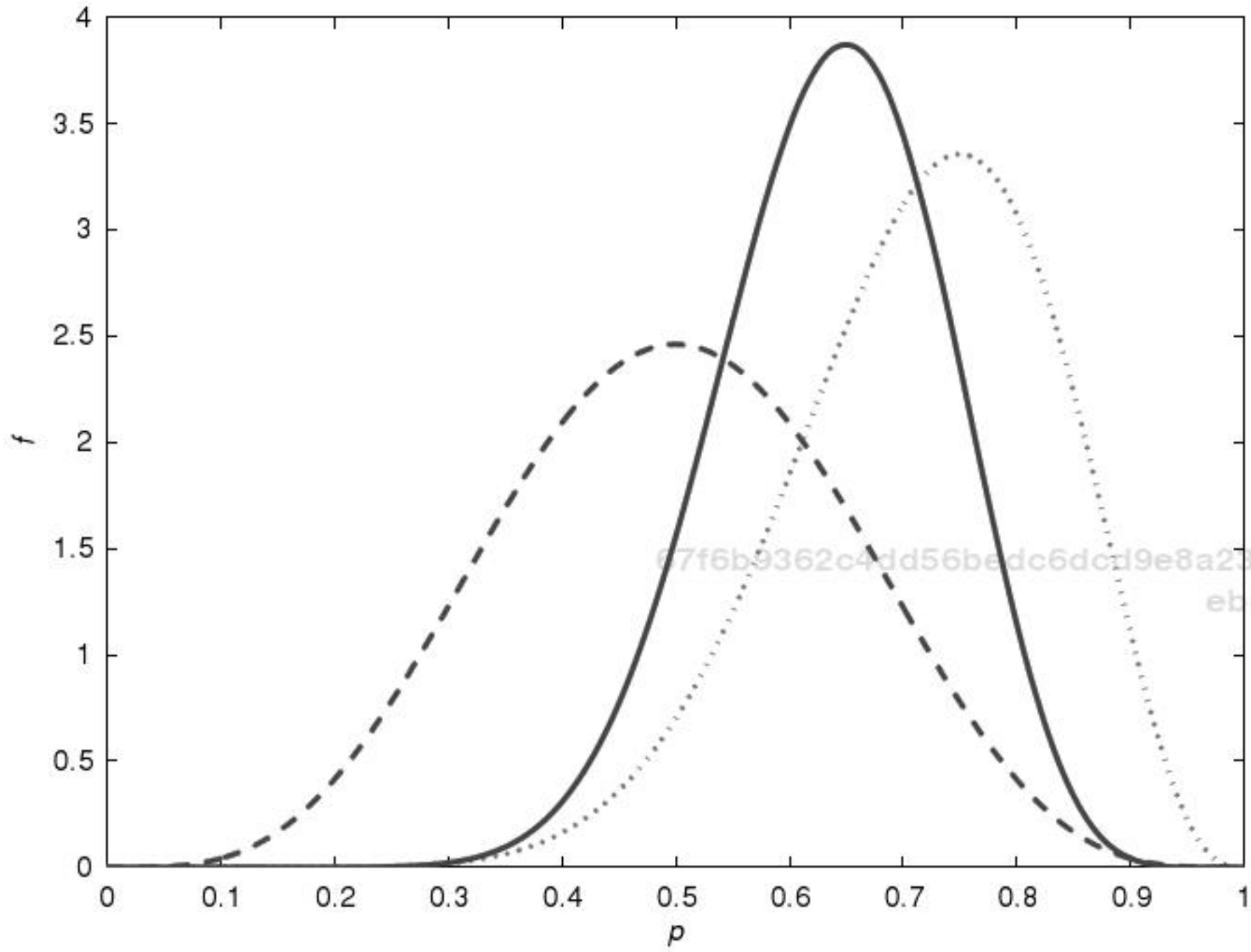

Figure 2.1 Prior (dashed line), scaled likelihood (dotted line), and posterior distribution (solid line) for the gambler's ruin problem.

the posterior mean, that is,

$$E[\boldsymbol{\theta}|\mathbf{x}] = \int \boldsymbol{\theta} f(\boldsymbol{\theta}|\mathbf{x}) \mathrm{d}\boldsymbol{\theta},$$

or, in the univariate case, through a posterior median, that is,

$$\theta_{\text{med}} \in \{y : P(\theta \leq y|x) = 1/2; P(\theta \geq y|x) = 1/2\}$$

or through a posterior mode, that is

$$\theta_{\text{mode}} = \arg\max f(\theta|\mathbf{x}).$$

Example 2.3: In the normal–normal Example 2.1, the posterior is normal and, hence, symmetric and unimodal so the mean, median, and mode are all equal. In particular, in the case when a uniform prior was used, these point estimates are all equal to $\bar{x}$, which is the maximum likelihood estimate (MLE). When the informative prior is

applied, they are all equal to

$$\frac{n\bar{x}/\sigma^2 + \mu_0/\sigma_0^2}{n/\sigma^2 + 1/\sigma_0^2},$$

which is a weighted average of the MLE and the prior mean with weights proportional to the precision of the MLE and the prior precision, respectively.

In the case of Example 2.2, the posterior distribution is asymmetric and so the mean, median, and mode are different. In particular, the gambler's posterior mean estimate of p is 0.6364, the posterior median is (approximately) 0.6406, and the posterior mode is (approximately) 0.65. △

Set estimation, that is, summarizing the posterior distribution through a set that includes $\boldsymbol{\theta}$ with high posterior probability, is also straightforward. When θ is univariate, one of the standard solutions is to fix the probability content of the set to $1-\alpha$, where typically used values of α, as in classical statistics, are 0.01, 0.05, or 0.1. An interval with this probability content is called a $100(1-\alpha)\%$ credible interval. Usually, there are (infinitely) many such credible intervals. One particular case that is often applied in practice is to use a *central posterior interval.* To calculate such an interval, the $\alpha/2$ and $1-\alpha/2$ posterior quantiles, say $q_{\frac{\alpha}{2}}$ and $q_{1-\frac{\alpha}{2}}$, are computed so that $P(\theta \in [q_1, q_2]|x) \geq 1-\alpha$, and $[q_{\frac{\alpha}{2}}, q_{(1-\frac{\alpha}{2})}]$ is the central interval. Another possibility is to use the shortest possible interval of probability $1-\alpha$, that is the *highest posterior density* (HPD) *interval.*

Example 2.4: In Example 2.1, if μ_1 and σ_1, respectively, designate the posterior mean and standard deviation, a posterior 95% central credible interval will be $[\mu_1 - 1.96\sigma_1, \mu_1 + 1.96\sigma_1]$, where 1.96 designates the 0.975 quantile of the standard normal distribution. Given the symmetry and unimodality of the normal distribution, this interval is also a HPD interval.

In Example 2.2, a posterior, 95% central credible interval can be shown numerically to be (0.4303, 0.8189). However, this interval is not an HPD interval. △

2.2.2 Hypothesis testing

In principle, hypothesis testing problems are straightforward. Consider the case in which we have to decide between two hypotheses with positive probability content, that is, $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ and $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1$. Then, theoretically, the choice of which hypothesis to accept can be treated as a simple decision problem (see Section 2.3). If we accept H_0 (H_1) when it is true, then we lose nothing. Otherwise, if we accept H_0, when H_1 is true, we lose a quantity l_{01} and if we accept H_1, when H_0 is true, we lose a quantity l_{10}. Then, given the posterior probabilities, $P(H_0|\mathbf{x})$ and $P(H_1|\mathbf{x})$, the expected loss if we accept H_0 is given by $P(H_1|\mathbf{x})l_{01}$ and the expected loss if we accept H_1 is $P(H_0|\mathbf{x})l_{10}$. The supported hypothesis is that which minimizes the expected loss. In particular, if $l_{01} = l_{10}$ we should simply select the hypothesis which is most likely a posteriori.

In many cases, such as model selection problems or multiple hypothesis testing problems, the specification of the prior probabilities in favor of each model or hypothesis may be very complicated and an alternative procedure that is not dependent on these prior probabilities may be preferred. The standard tool for such contexts is the *Bayes factor*.

Definition 2.1: *Suppose that H_0 and H_1 are two hypotheses with prior probabilities $P(H_0)$ and $P(H_1)$ and posterior probabilities $P(H_0|\mathbf{x})$ and $P(H_1|\mathbf{x})$, respectively. Then, the Bayes factor in favor of H_0 is*

$$B_1^0 = \frac{P(H_1)P(H_0|\mathbf{x})}{P(H_0)P(H_1|\mathbf{x})}.$$

It is easily shown that the Bayes factor reduces to the marginal likelihood ratio, that is,

$$B_1^0 = \frac{f(\mathbf{x}|H_0)}{f(\mathbf{x}|H_1)},$$

which is independent of the values of $P(H_0)$ and $P(H_1)$ and is, therefore, a measure of the evidence in favor of H_0 provided by the data. Note, however, that, in general, it is not totally independent of prior information as

$$f(\mathbf{x}|H_0) = \int_{\boldsymbol{\Theta}_0} f(\mathbf{x}|H_0, \boldsymbol{\theta}_0) f(\boldsymbol{\theta}_0|H_0) d\boldsymbol{\theta}_0,$$

which depends on the prior density under H_0, and similarly for $f(\mathbf{x}|H_1)$. Kass and Raftery (1995) presented the following Table 2.1, which indicates the strength of evidence in favor of H_0 provided by the Bayes factor.

Table 2.1 Strength of evidence in favor of H_0 provided by the Bayes factor.

B_1^0	$2 \log_{10} B_1^0$	Evidence against H_1
1–3	0–2	Hardly worth commenting
3–20	2–6	Positive
20–150	6–10	Strong
>150	> 10	Very strong

Example 2.5: Continuing with the gambler's ruin example, suppose that the gambler wishes to test whether or not the coin was biased in favor of heads, that is $H_0 : p > 0.5$ as against the alternative $H_1 : p \leq 0.5$. Given that the gambler's prior distribution was symmetric, we have $P(H_0) = P(H_1) = 0.5$ and, for example,

$$f(p|H_0) = \frac{2}{B(5,5)} p^{5-1}(1-p)^{5-1} \quad \text{for} \quad 0.5 < p < 1,$$

whereas the density $f(p|H_1)$ has the same expression, but for $0 \leq p \leq 0.5$. Then, the gambler's posterior probability that H_0 is true is $P(p > 0.5|\mathbf{x}) = 0.9054$ and the Bayes factor in favor of H_0 is $B_1^0 = 9.5682$. Therefore, there is positive evidence in favor of the hypothesis H_0 that the probability of heads is greater than 0.5. △

In computationally complex problems, the Bayes factor may often be difficult to evaluate and, in such cases, a simpler alternative is to use a model selection criterion. The most popular criterion in the Bayesian context, particularly in the situation of having to select between nested models, is the deviance information criterion, or DIC, developed in Spiegelhalter *et al.* (2002), which we define in the following text.

Definition 2.2: *Suppose that a model $\mathcal{M}$ is parameterized by $\boldsymbol{\theta}$. Then, given a sample of data, $\mathbf{x}$, the DIC for $\mathcal{M}$ is given by*

$$DIC_{\mathcal{M}} = -4E\left[\log f(\mathbf{x}|\boldsymbol{\theta})|\mathbf{x}\right] + 2\log f(\mathbf{x}|E[\boldsymbol{\theta}|\mathbf{x}]).$$

Lower values of the DIC indicate more plausible models.

2.2.3 Prediction

In many applications, rather than being interested in the parameters, we shall be more concerned with the prediction of future observations of the variable of interest. This is especially true in the case of stochastic processes, when we will typically be interested in predicting both the short- and long-term behavior of the process.

For prediction of future values, say $\mathbf{Y}$, of the phenomenon, we use the predictive distribution. To do this, given the current data $\mathbf{x}$, if we knew the value of $\boldsymbol{\theta}$, we would use the conditional predictive distribution $f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$. However, since there is uncertainty about $\boldsymbol{\theta}$, modeled through the posterior distribution, $f(\boldsymbol{\theta}|\mathbf{x})$, we can integrate this out to calculate the predictive density

$$f(\mathbf{y}|\mathbf{x}) = \int f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{x}) f(\boldsymbol{\theta}|\mathbf{x}) \mathrm{d}\boldsymbol{\theta}. \tag{2.3}$$

Note that in the case that the sampled values of the phenomenon are conditionally IID, the formula (2.3) simplifies to

$$f(\mathbf{y}|\mathbf{x}) = \int f(\mathbf{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{x}) \mathrm{d}\boldsymbol{\theta},$$

although, in general, to predict the future values of stochastic processes, this simplification will not be available. The predictive density may be used to provide point or set forecasts and test hypotheses about future observations, much as we did earlier.

Example 2.6: In the normal-normal example, to predict the next observation $Y = X_{n+1}$, we have that in the case when a uniform prior was applied, then

$X_{n+1} \mid x \sim \mathrm{N}(\bar{x}, \frac{n+1}{n}\sigma^2)$. Then a predictive $100(1-\alpha)\%$ probability interval is $\left[\bar{x} - z_{\alpha/2}\sigma\sqrt{(n+1)/n}, \bar{x} + z_{\alpha/2}\sigma\sqrt{(n+1)/n}\right]$. △

Example 2.7: In the gambler's ruin example, suppose that the gambler formally starts to play the game with an initial stake $x_0 = 2$ and that she wins the game if she reaches a total of $m = 10$ euros. Then, the gambler might be interested in the probability that she is ruined in the next few turns. Let x_t be her stake after t more tosses of the coin. Then, clearly, $P(x_1 = 0|\mathbf{x}) = 0$ and

$$P(x_2 = 0|\mathbf{x}) = \int_0^1 (1-p)^2 f(p|\mathbf{x})\,\mathrm{d}p = \frac{B(14, 10)}{B(14, 8)} = 0.1423.$$

In a similar way, we can see that $P(x_3 = 0|\mathbf{x}) = 0.1423$ and

$$P(x_4 = 0|\mathbf{x}) = \int_0^1 (1-p)^2\,[1 + 2p(1-p)]\,f(p|\mathbf{x})\,\mathrm{d}x = 0.2087.$$

Usually, however, the gambler will be more interested in predicting the probability that either she eventually wins the game, or the probability that she is eventually ruined, rather than the earnings she will have after a given number of plays.

Assume that the gambler has an initial stake x_0 and that she wins the game if she increases her stake to $m \geq x_0$, where $x_0, m \in \mathbb{N}$. Then, it is well known that, for a given p, the probability that she wins the game is

$$P(\text{wins}|p) = \frac{1 - \left(\frac{q}{p}\right)^{m-x_0}}{1 - \left(\frac{q}{p}\right)^{m}},$$

where $q = 1 - p$ and we assume that $q \neq p$. Otherwise, the winning probability can be shown to be equal to x_0/m.

Therefore, her predictive probability of winning, given her current posterior for p, is

$$\begin{aligned} P(\text{wins}|\mathbf{x}) &= \int_0^1 P(\text{wins}|p) f(p|\mathbf{x})\,\mathrm{d}p \\ &= \int_0^1 \frac{1}{B(14, 8)} p^{14-1}(1-p)^{8-1} \frac{1 - \left(\frac{1-p}{p}\right)^2}{1 - \left(\frac{1-p}{p}\right)^{10}}\,\mathrm{d}p \simeq 0.622. \end{aligned}$$

Her predictive probability of eventually being ruined is 0.378. △

2.2.4 Sensitivity analysis and objective Bayesian methods

As mentioned earlier, prior information may often be elicited from one or more experts. In such cases, the postulated prior distribution will often be an approximation to the expert's beliefs. In case that different experts disagree, there may be considerable uncertainty about the appropriate prior distribution to apply. In such cases, it is important to assess the sensitivity of any posterior results to changes in the prior distribution. This is typically done by considering appropriate classes of prior distributions, close to the postulated expert prior distribution, and then assessing how the posterior results vary over such classes.

Example 2.8: Assume that the gambler in the gambler's ruin problem is not certain about her Be(5, 5) prior and wishes to consider the sensitivity of the posterior predictive ruin probability over a reasonable class of alternatives. One possible class of priors that generalizes the gambler's original prior is

$$G = \{f : f \sim \text{Be}(c, c), c > 0\},$$

the class of symmetric beta priors. Then, over this class of priors, it can be shown that the gambler's posterior predictive ruin probability varies between 0.231, when $c \to 0$ and 0.8, when $c \to \infty$. This shows that there is a large degree of variation of this predictive probability over this class of priors. △

When little prior information is available, or in order to promote a more objective analysis, we may try to apply a prior distribution that provides little information and 'lets the data speak for themselves'. In such cases, we may use a noninformative prior. When Θ is discrete, a sensible noninformative prior is a uniform distribution. However, when Θ is continuous, a uniform distribution is not necessarily the best choice. In the univariate case, the most common approach is to use the Jeffreys prior.

Definition 2.3: *Suppose that* $X|\theta \sim f(\cdot|\theta)$. *The Jeffreys prior for* θ *is given by*

$$f(\theta) \propto \sqrt{I(\theta)},$$

where $I(\theta) = -E_X\left[\frac{d^2}{d\theta^2} \log f(X|\theta)\right]$ *is the expected Fisher information.*

Example 2.9: Consider the case of Example 2.1. We have

$$\log f(X|\theta) = -\frac{1}{2}\log \pi\sigma^2 - \frac{1}{2\sigma^2}(X-\theta)^2$$
$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X|\theta) = \frac{X-\theta}{\sigma^2}$$
$$\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\log f(X|\theta) = -\frac{1}{\sigma^2},$$

which is constant. Therefore, the Jeffreys prior is a uniform distribution, $f(\theta) \propto 1$, the distribution that was previously applied. Note that given this prior, the posterior mean and $100(1-\alpha)\%$ central credible interval coincide with the classical MLE and confidence interval, respectively.

In the gambler's ruin problem, however, given that we assume that we are going to observe the number of heads in 12 tosses of the coin, the Jeffreys prior is easily shown to be $p \sim \text{Be}(1/2, 1/2)$. When this prior is used, the gambler's posterior mean is $E[p|\mathbf{x}] = 0.76$ that is close, but not equal, to the MLE, $\hat{p} = 0.75$. △

The Jeffreys prior is not always appropriate when $\boldsymbol{\Theta}$ is multivariate. In this context, the most popular approach is the use of the so-called reference priors, as developed in Bernardo (1979).

2.3 Bayesian decision analysis

Often, the ultimate aim of statistical research will be to support decision-making. As an example, the gambler might have to decide whether or not to play the game and what initial stake to put. An important strength of the Bayesian approach is its natural inclusion into a coherent framework for decision-making, which, in practical terms, leads to Bayesian decision analysis.

If the consequences of the decisions, or actions of a decision maker (DM), depend upon the future values of observations, the general description of a decision problem is as follows. For each feasible action $a \in \mathcal{A}$, with $\mathcal{A}$ the action space, and each future result $\mathbf{y}$, we associate a consequence $c(a, \mathbf{y})$. For example, in the case of the gambler's ruin problem, if the gambler stakes a quantity x_0 (the action a) and wins the game after a sequence $\mathbf{y}$ of results, the consequence is that she wins a quantity $m - x_0$. This consequence will be evaluated through its utility $u(c(a, \mathbf{y}))$, which encodes the DM's preferences and risk attitudes. The DM should choose the action maximizing her predictive expected utility

$$\max_{a \in \mathcal{A}} \int u(c(a, \mathbf{y})) f(\mathbf{y}|\mathbf{x}) \mathrm{d}y,$$

where $f(\mathbf{y}|\mathbf{x})$ represents the DM's predictive density for $\mathbf{y}$ given her current knowledge and data, $\mathbf{x}$, described in (2.3).

In other instances, the consequences will actually depend on the parameter $\boldsymbol{\theta}$, rather than on the observable $\mathbf{y}$. In these cases, we shall be interested in maximizing the posterior expected utility

$$\max_{a \in \mathcal{A}} \int u(c(a, \boldsymbol{\theta})) f(\boldsymbol{\theta}|\mathbf{x}) \mathrm{d}\boldsymbol{\theta}. \tag{2.4}$$

In most statistical contexts, we normally talk about losses, rather than utilities, and we aim at minimizing the posterior (or predictive) expected loss. We just need to consider that utility is the negative of the loss. Note also that all the standard statistical

Table 2.2 Winning probabilities and expected utility gains for different initial stakes.

x_0	$P(\text{wins}\|\mathbf{x}, x_0)$	$E[u(x_0)\|\mathbf{x}]$
0	0.0000	0.0000
1	0.4237	0.3896
2	0.6218	**0.9746**
3	0.7260	0.8083
4	0.7863	0.2904
5	0.8239	−0.4087

approaches mentioned earlier may be justified within this framework. As an example, if we are interested in point estimation through the posterior mean, we may easily see that this estimate is optimal, in terms of minimizing posterior expected loss, when we use the quadratic loss function (see, e.g., French and Ríos Insua, 2000). We would like to stress, however, that we should not always appeal to such canonical utility/loss functions, but rather try to model whatever relevant consequential aspects we may deem appropriate in the problem at hand.

Example 2.10: Assume that the bank always starts with 8 euros and that the gambler has to pay a 2 euro fee plus her initial stake to play the game. Thus, if she starts with an initial stake x_0, then, if she plays and loses, she loses $x_0 + 2$ euro, whereas if she plays and wins, she gains $8 - x_0$ euro. The gambler has to select the optimal initial stake to play with.

Assume that the gambler has a linear utility function for money, thus, being risk neutral. Clearly, it is illogical for the gambler to play the game with an initial stake of more than or equal to 6 euros in this case. Table 2.2 gives the gambler's probability of winning and expected utility values for stakes between 0, when she does not play, and 5.

The optimal strategy for the gambler is to play the game with an initial stake of 2 euros, when her expected monetary gain is equal to 0.9746 euros. △

2.4 Bayesian computation

Even if Bayesian analysis is conceptually simple, leaving aside modeling complexities affecting real applications, very frequently we shall have to face considerable computational complexities not amenable to standard analytic solutions. Therefore, we now provide a brief review of some of the most important computational procedures that facilitate the implementation of Bayesian analysis, with special emphasis on simulation methods.

2.4.1 Computational Bayesian statistics

The key operation in the practical implementation of Bayesian methods is integration. In the examples we have seen so far in this chapter, most integrations are standard

and may be done analytically. This is a typical consequence of the use of conjugate prior distributions: a class of priors is conjugate to a given model, if the resulting posterior belongs to the same class of distributions. When the properties of the conjugate family of distributions are known, the use of conjugate prior distributions greatly simplifies Bayesian analysis procedures since, given observed data, the calculation of the posterior distribution reduces to simply modifying the parameters of the prior distribution. However, it is important to note that conjugate prior distributions are associated with (generalized) exponential family sampling distributions, and, therefore, that conjugate prior distributions do not always exist. For example, if we consider data generated from a Cauchy distribution, then it is well known that no conjugate prior exists.

However, more complex, nonconjugate models will generally not allow for such neat computations. Various techniques for approximating Bayesian integrals can be considered.

When the sample size is sufficiently large, central limit type theorems can sometimes be applied so that the posterior distribution is approximated by a normal distribution, when integrals may often be estimated in a straightforward way. Otherwise, in low-dimensional problems such as in Example 2.7, we can often apply numerical integration techniques like Gaussian quadrature. However, in higher dimensional problems, the number of function evaluations necessary to accurately evaluate the relevant integrals increases rapidly and such methods become inaccurate. Therefore, approaches based on simulation are typically preferred. Given their increasing importance in Bayesian statistical computation, we outline such methods.

The key idea is that of Monte Carlo integration, which substitutes an integral by a sample mean of a sufficiently large number, say N, of values simulated from the relevant posterior distribution. If $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^N$ is a sample from $f(\boldsymbol{\theta}|\mathbf{x})$, then we have that for some function, $g(\boldsymbol{\theta})$, with finite posterior mean and variance, then

$$\frac{1}{N}\sum_{i=1}^{N} g(\boldsymbol{\theta}^{(i)}) \cong E[g(\boldsymbol{\theta}) \mid \mathbf{x}].$$

This result follows from the strong law of large numbers, which provides almost sure convergence of the Monte Carlo approximation to the integral. The variance of the Monte Carlo approximation provides guidance on the precision of the estimate.

Example 2.11: In the gambler's ruin problem, a simple way of estimating the predictive mean ruin probability is to generate a sample from the posterior beta distribution, calculate the ruin probability for each sampled value, and then use the mean of the ruin probabilities to estimate the predictive ruin probability. Table 2.3 shows the predictive ruin probabilities estimated from samples of various different sizes.

For a sample of size 10 000, the estimated probability has converged to the same value calculated previously in Example 2.7 via numerical integration. △

In many problems, the posterior distribution will typically only be known up to a constant, and sampling directly from this distribution is not straightforward. In some

Table 2.3 Predictive ruin probabilities estimated from Monte Carlo samples of size N.

N	P(ruin$\|\mathbf{x}$)
10	0.397
100	0.384
1000	0.380
10000	0.378

cases, generalizations of the basic Monte Carlo method may be used. One possibility that can sometimes be applied is to use approaches such as independence samplers or rejection samplers that sample from distributions that are similar to the posterior distribution and then weigh the sampled elements appropriately.

However, the most intensely used techniques in modern Bayesian inference are Markov chain Monte Carlo (MCMC) methods. These methods are based on the assumption that we can find a Markov chain $\boldsymbol{\theta}^{(n)}$ with states $\boldsymbol{\theta}$ and stationary distribution equal to the (posterior) distribution of interest. The strategy is then to start from arbitrary values of $\boldsymbol{\theta}$, let the Markov chain run until practical convergence is judged, and then use the next N observed values from the chain to approximate a Monte Carlo sample from the distribution of interest. Given that successive values generated from a Markov chain are correlated, to make the sampled values approximately independent, some of the sampled values are often omitted in order to mitigate serial correlation.

The key issue in MCMC methods is how to construct Markov chains with the desired stationary distribution. Several generic strategies for designing such chains are available. Since these methods are generic, in what follows, we shall drop dependence from the data in their description, and we shall assume throughout that the objective is to sample from a density $f(\boldsymbol{\theta})$.

The most well-known MCMC approach is the Gibbs sampler. Suppose that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$. Suppose that the conditional distributions,

$$f(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}), \quad \text{where} \quad \boldsymbol{\theta}_{-i} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_k)$$

for $i = 1, \ldots, k$ can all be easily sampled from. Then, starting from arbitrary values, the Gibbs sampler simply iterates through these conditionals until convergence:

1. Choose initial values $\left(\boldsymbol{\theta}_2^0, \ldots, \boldsymbol{\theta}_k^0\right)$. $i = 1$.
2. Until convergence is detected, iterate through
   Generate $\boldsymbol{\theta}_1^i \sim \boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{i-1}, \ldots, \boldsymbol{\theta}_k^{i-1}$
   Generate $\boldsymbol{\theta}_2^i \sim \boldsymbol{\theta}_2|\boldsymbol{\theta}_1^i, \boldsymbol{\theta}_3^{i-1}, \cdots, \boldsymbol{\theta}_k^{i-1}$
   ...
   Generate $\boldsymbol{\theta}_k^i \sim \boldsymbol{\theta}_k|\boldsymbol{\theta}_1^i, \ldots, \boldsymbol{\theta}_{k-1}^i$.
   $i = i + 1$

Under sufficiently general conditions (see, e.g., Tierney, 1994), the sampler defines a Markov chain with the desired posterior as its stationary distribution. The Gibbs sampler is particularly attractive in many scenarios, such as the analysis of hierarchical models, because the conditional posterior density of one parameter given the others is often relatively simple, perhaps after the introduction of some auxiliary variables. A simple example from Berger and Ríos Insua (1998) is provided in the following text.

Example 2.12: Suppose $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and that the posterior density is

$$f(\theta_1, \theta_2 \mid \mathbf{x}) = \frac{1}{\pi} \exp\left(-\theta_1(1+\theta_2^2)\right).$$

defined over the set $\theta_1 > 0$ and $-\infty < \theta_2 < \infty$. Many posterior expectations associated with this distribution cannot be computed analytically. As an alternative, it is straightforward to see that $\theta_2|\theta_1 \sim \mathrm{N}\left(0, \frac{1}{2\theta_1}\right)$ and that, $\theta_1|\theta_2 \sim \mathrm{Ex}\left(1+\theta_2^2\right)$ and therefore, a Gibbs sampler can be applied. △

A general approach to designing a Markov chain with a desired stationary distribution is the Metropolis–Hastings (MH) algorithm. To implement MH, we only need to be able to evaluate the target distribution pointwise, up to a constant, and generate observations from a probing distribution $q(\cdot|\cdot)$ under certain technical conditions. At each step, the generated state is accepted according to certain accept–reject probabilities. Specifically, we proceed as follows:

1. Choose initial values $\boldsymbol{\theta}^{(0)}$. $i = 0$
2. Until convergence is detected, iterate through
   Generate a candidate $\boldsymbol{\theta}^* \sim q\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(i)}\right)$.
   If $p\left(\boldsymbol{\theta}^{(i)}\right)q\left(\boldsymbol{\theta}^{(i)} \mid \boldsymbol{\theta}^*\right) > 0$, $\alpha\left(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^*\right) = \min\left(\frac{p\left(\boldsymbol{\theta}^*\right)q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(i)}\right)}{p\left(\boldsymbol{\theta}^{(i)}\right)q\left(\boldsymbol{\theta}^{(i)} \mid \boldsymbol{\theta}^*\right)}, 1\right)$;
   else, $\alpha\left(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^*\right) = 1$.
   Do

$$\theta^{(i+1)} = \begin{cases} \theta^* \text{ with prob } \alpha\left(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^*\right), \\ \theta^{(i)} \text{ with prob } 1 - \alpha\left(\boldsymbol{\theta}^{(i)}, \boldsymbol{\theta}^*\right) \end{cases}$$

$i = i + 1$.

When the probing distribution is symmetric in its arguments, the acceptance probability, α, simplifies to $\min\left(p\left(\boldsymbol{\theta}^*\right)/p\left(\boldsymbol{\theta}^{(i)}\right), 1\right)$, corresponding to the variant introduced by Metropolis *et al.* (1953). Also, when the probing distribution $q(\cdot|\boldsymbol{\theta})$ is independent of $\boldsymbol{\theta}$, the sampler is known as an independence sampler. Finally, note that the Gibbs sampler is a particular case of a MH sampler in blocks, where the probing distribution is equal to the conditional distribution, $f(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i})$ and the acceptance probability is equal to 1.

Complex problems will typically require a mixture of various MCMC algorithms or, as they are known, hybrid methods. As an example, Müller (1991) suggests using Gibbs sampler steps when conditionals are available for efficient sampling and Metropolis steps, otherwise. We illustrate these ideas with a software reliability example based on a nonhomogeneous Poisson process.

Example 2.13: Consider a nonhomogeneous Poisson process, with rate function $\lambda(t) = m'(t) = ae^{-bt}$, which corresponds to Schneidewind's software reliability growth model (see, e.g., Singpurwalla and Wilson, 1999). Now $\boldsymbol{\theta} = (a, b)$. Assume we test until we observe n failures, and we observe $\mathbf{t} = (t_1, t_2, \cdots, t_n)$ as times between failures, that is, the first failure occurs at time $s_1 = t_1$; the second one occurs at time $s_2 = t_1 + t_2$, and so on, until $s_n = t_1 + t_2 + \ldots + t_n$. Assume gamma prior distributions for a and b

$$a \sim \text{Ga}(\alpha_1, \beta_1), \quad b \sim \text{Ga}(\alpha_2, \beta_2).$$

After some computations, we find

$$f\,(a, b|\mathbf{t}) \propto a^{\alpha_1+n-1}\mathrm{e}^{-\beta_1 a}\; b^{\alpha_2-1}\mathrm{e}^{-b\left(\beta_2+\sum_{i=1}^{n} s_i\right)}\mathrm{e}^{-\frac{a}{b}\left(1-\mathrm{e}^{-bs_n}\right)},$$

from which we obtain

$$f\,(a|b, \mathbf{t}) \propto a^{\alpha_1+n-1}\mathrm{e}^{-a\left[\beta_1+\frac{1}{b}\left(1-\mathrm{e}^{-bs_n}\right)\right]},$$

which is a gamma distribution with parameters $\alpha_1 + n$ and $\beta_1 + \frac{1}{b}\left(1 - \mathrm{e}^{-bs_n}\right)$, and

$$f\,(b|a, \mathbf{t}) \propto b^{\alpha_2-1}\mathrm{e}^{-b\left(\beta_2+\sum_{i=1}^{n} s_i\right)}\mathrm{e}^{-\frac{a}{b}\left(1-\mathrm{e}^{-bs_n}\right)},$$

which is a nonstandard distribution. However, the distribution can be sampled using the following Metropolis step:

1. `Sample` $\tilde{b} \sim \text{N}(b^{(i)}, \sigma^2)$.
2. `Set` $b^{(i+1)} = \tilde{b}$ `with probability` $\alpha = \min\left(1, \frac{f\left(\tilde{b}|a_i, \mathbf{t}\right)}{f\left(b^i|a_i, \mathbf{t}\right)}\right)$. `Otherwise,` `set` $b^{(i+1)} = b^{(i)}$.

Here, the standard deviation σ can be chosen to approximately optimize the acceptance rate of the Metropolis step to be between 20% and 50%; see, for example, Gamerman and Lopes (2006) for details. After some calculations, we obtain that:

$$\frac{f\,(b_{\text{cand}}|\,a_i\,, \text{data})}{f\,\left(b^i\,|\,a_i\,, \text{data}\right)} = \left(\frac{b_{\text{cand}}}{b^i}\right)^{(\alpha_2-1)} \mathrm{e}^{-\left(b_{\text{cand}}-b^i\right)\left(\beta_2+\sum_{i=1}^{n} s_i\right)}\; \mathrm{e}^{-\frac{a}{b_{\text{cand}}}\left(1-\mathrm{e}^{-b_{\text{cand}}s_n}\right)+\frac{a}{b^i}\left(1-\mathrm{e}^{-b^i s_n}\right)}$$

Then, we easily set up a Markov chain with Gibbs gamma steps when sampling from the conditional posterior of a and Metropolis steps when sampling from the conditional posterior of b. △

In the context of stochastic process models, sequential MCMC or particle filtering methods are particularly relevant. We briefly describe a specific case of particle filter. As a contrast between analytic and simulation methods, we first illustrate inference and forecasting with dynamic linear models, which can be performed analytically.

Consider the general, normal, dynamic linear model (DLM) with univariate observations X_n, which is characterized by the quadruple $\{\boldsymbol{F}_n, \boldsymbol{G}_n, V_n, \boldsymbol{W}_n\}$, where, for each n, $\boldsymbol{F}_n$ is a known vector of dimension $m \times 1$, $\boldsymbol{G}_n$ is a known $m \times m$ matrix, V_n is a known variance, and $\boldsymbol{W}_n$ is a known $m \times m$ variance matrix. The model is then succinctly written as

$$\begin{aligned} \theta_0|D_0 &\sim \mathrm{N}(m_0, C_0) \\ \theta_n|\theta_{n-1} &\sim \mathrm{N}(\boldsymbol{G}_n\theta_{n-1}, \boldsymbol{W}_n) \\ X_n|\theta_n &\sim \mathrm{N}(\boldsymbol{F}_n'\theta_n, V_n). \end{aligned} \tag{2.5}$$

The information, D_n, is defined recursively as $D_n = D_{n-1} \cup \{x_n\}$. Note that we may also write

$$\begin{aligned} \theta_n &= \boldsymbol{G}_n\theta_{n-1} + w_n, w_n \sim \mathrm{N}(0, \boldsymbol{W}_n) \\ X_n &= \boldsymbol{F}_n\theta_n + v_n, v_n \sim \mathrm{N}(0, V_n). \end{aligned}$$

The following result (see, e.g., West and Harrison, 1997) summarizes the basic features of DLMs for forecasting and inference.

Theorem 2.1: *For the general univariate DLM, the posterior and one-step ahead predictive distributions are, for each n:*

- *Posterior at* $n-1$,

$$\theta_{n-1}|D_{n-1} \sim N(m_{n-1}, C_{n-1}).$$

- *Prior at* n,

$$\theta_n|D_{n-1} \sim N(a_n, R_n).$$

- *One-step ahead forecast*,

$$X_n|D_{n-1} \sim N(f_n, Q_n).$$

Where

$$
\begin{aligned}
a_n &= \boldsymbol{G}_n m_{n-1}, \\
R_n &= \boldsymbol{G}_n C_{n-1} \boldsymbol{G}_n' + \boldsymbol{W}_n, \\
f_n &= \boldsymbol{F}_n' a_n, \\
Q_n &= \boldsymbol{F}_n' R_n \boldsymbol{F}_n + V_n, \\
A_n &= R_n \boldsymbol{F}_n Q_n^{-1}, \\
C_n &= R_n - A_n A_n' Q_n, \\
m_n &= a_n + A_n (x_n - f_n)
\end{aligned}
$$

Consider now the following generalization of (2.5):

$$
\begin{aligned}
\theta_0 | D_0 &\sim f_0 \\
\theta_n &= g_{n-1}(\theta_{n-1}, \boldsymbol{W}_n) \\
X_n &= f_n(\theta_n, V_n),
\end{aligned}
$$

where $\boldsymbol{W}_n$ and V_n are error terms with given distributions, jointly independent at every instant time. As earlier, we are interested in computing the posterior distributions of the parameter $\theta_n | D_n$ and, more importantly, the one-step ahead forecast distributions $X_{n+1} | D_n$. We may not do this now analytically and we shall propose an approximation based on simulation.

For that, we shall use a particle filter which, *at time n*, consists of a set of random values $\{\boldsymbol{\theta}_n^{(i)}\}_{i=1}^N$ with associated weights $m_n^{(i)}$, such that

$$
\sum_{i=1}^{N} h(\boldsymbol{\theta}_n^{(i)}) m_n^{(i)} \simeq \int h(\boldsymbol{\theta}_n) f(\boldsymbol{\theta}_n) \mathrm{d}\boldsymbol{\theta}_n,
$$

for typical functions h on the state space, the convergence being in probability as N grows. To evolve the particle filter and make the forecasts, we shall use the following recursions:

$$
\begin{aligned}
f(\boldsymbol{\theta}_n | D_{n-1}) &= \int f(\boldsymbol{\theta}_n | \boldsymbol{\theta}_{n-1}) f(\boldsymbol{\theta}_{n-1} | D_{n-1}) \mathrm{d}\boldsymbol{\theta}_{n-1}, \\
f(\boldsymbol{\theta}_n | D_n) &= \frac{f(x_n | \boldsymbol{\theta}_n) f(\boldsymbol{\theta}_n | D_{n-1})}{f(x_n | D_{n-1})}, \\
f(x_n | D_{n-1}) &= \int f(x_n | \boldsymbol{\theta}_n) f(\boldsymbol{\theta}_n | D_{n-1}) \mathrm{d}\boldsymbol{\theta}_n
\end{aligned}
$$

The particle filter we describe is based on the sampling importance resampling algorithm of Gordon *et al.* (1993) and is described as follows:

1. Generate $\theta_0^{(i)} \sim f(\theta_0|D_0), i = 1, \ldots, N$, with $m_1^{(i)} = 1/N$, $n = 1$.
2. While $n \leq M$:
   Approximate $f(\boldsymbol{\theta}_n|D_n)$, up to a normalizing constant K, with the mixture

$$K \sum_{i=1}^{N} f(\boldsymbol{\theta}_n|\boldsymbol{\theta}_{n-1}^{(i)}) f(x_n|\boldsymbol{\theta}_n).$$

   To do so, generate $\boldsymbol{\theta}_n^{(i)} \sim f_{n-1}(\boldsymbol{\theta}_{n-1}^{(i)}, W_{n-1}^{(i)})$, $i = 1, \ldots, N$ with importance weights $m_n^i = f(x_n|\boldsymbol{\theta}_n^{(i)})(\sum_j p(x_n|\boldsymbol{\theta}_n^{(i)}))$
   Sample N times independently with replacement from $\{\boldsymbol{\theta}_n^{(i)}, m_n^{(i)}\}$ to produce the random measure $\{\boldsymbol{\theta}_n^{(i)}, 1/N\}$.
   $n = n + 1$.

Carpenter *et al.* (1999), Pitt and Shepard (1999), and Del Moral *et al.* (2006) provide additional information and improvements over the above basic filter.

We end up with a discussion of a method that may be very useful in relation with predictive computations with stochastic process models. We consider the predictive computation of an event A dependent on $\boldsymbol{\theta}$. The posterior predictive probability would be

$$P(A|\mathbf{x}) = \int P(A|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathbf{x}) \mathrm{d}\boldsymbol{\theta}.$$

On the basis of a large sample $\{\theta_i\}_{i=1}^N$ from the posterior distribution, we could approximate it by Monte Carlo through

$$P_{\mathrm{MC}}(A|\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} P(A|\boldsymbol{\theta}_i).$$

This approach may be infeasible when finding each $P(A|\boldsymbol{\theta}_i)$ is computationally demanding, as happens with stochastic process based models, which we shall illustrate in later chapters. This problem entails that we are unable to use large enough samples so that standard MC approximations can be applied.

Assume instead that we are able to approximate the relevant posterior distribution $f(\boldsymbol{\theta}|\mathbf{x})$ by a simple probability distribution $\tilde{\boldsymbol{\Theta}}$ with m support points $\{\boldsymbol{\theta}_i\}_{i=1}^m$ each with probability p_i, $i = 1, \ldots, m$, where m is small enough so that m predictive computations $P(A|\boldsymbol{\theta}_i)$ are actually amenable. Then, we could aim at approximating the quantity of interest through

$$\tilde{P}(A) = \sum_{i=1}^{m} P(A|\boldsymbol{\theta}_i) p_i,$$

satisfactorily under appropriate conditions. We first determine the order $m \geq 1$ of this reduced order model (ROM), based on purely computational reasons: our computational budget allows only for m $P(A|\boldsymbol{\theta})$ computations. Then, we determine the range $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$ of $\tilde{\boldsymbol{\Theta}}$ for the selected m. Finally, we calculate the probabilities $\{p_1, \ldots, p_m\}$ of $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$, and use the ROM approximation. Grigoriu (2009) describes procedures for such an approximation.

2.4.2 Computational Bayesian decision analysis

We now briefly address computational issues in relation with Bayesian decision analysis problems. In principle, this involves two operations: (1) integration to obtain expected utilities of alternatives and (2) optimization to determine the alternative with maximum expected utility. To fix ideas, we shall assume that we aim at solving problem (2.4), that is finding the alternative of maximum posterior expected utility. If the posterior distribution is independent of the action chosen, then we may drop the denominator $\int f(\mathbf{x}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$, solving the possibly simpler problem

$$\max_a \int u(a, \boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.$$

Also recall that for standard statistical decision theoretical problems, the solution of the optimization problem is well known. For example, in an estimation problem with absolute value loss, the optimal estimate will be the posterior median. We shall refer here to problems with general utility functions. We first describe two simulation-based methods and then present a key optimization principle in sequential problems, Bellman's dynamic programming principle, which will be relevant when dealing with stochastic processes.

The first approach we describe is called *sample path optimization* in the simulation literature and was introduced in statistical decision theory in Shao (1989). To be most effective, it requires that the posterior does not depend on the action chosen. In such cases, we may use the following strategy:

1. Select a sample $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^N \sim p(\boldsymbol{\theta}|\mathbf{x})$.
2. Solve the optimization problem

$$\max_{a \in \mathcal{A}} \frac{1}{N} \sum_{i=1}^{N} u(a, \boldsymbol{\theta}^i)$$

yielding a_N^*.

If the maximum expected utility alternative a^* is unique, we may prove that $a_N^* \to a^*$, almost surely. Note that the auxiliary problem used to find a_N^* is a standard mathematical programming problem, see Nemhauser *et al.* (1990) for ample information.

Suppose now that the posterior actually depends on the chosen action. Assume that the posterior is $f(\boldsymbol{\theta}|\mathbf{x}, a) > 0$, for each pair $(a, \boldsymbol{\theta})$. If the utility function is

positive and integrable, we may define an artificial distribution on the augmented product space $\mathcal{A} \times \boldsymbol{\Theta}$ with density $h(a, \boldsymbol{\theta})$ proportional to the product of the utility function and the posterior probability density

$$h(a, \boldsymbol{\theta}) \quad \propto \quad u(a, \boldsymbol{\theta}) \times p(\boldsymbol{\theta} \mid \mathbf{x}, a).$$

If we compute the marginal distribution *on* a, we have

$$h(a) = \int h(a, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \quad \propto \quad \int u(a, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x}, a)\, \mathrm{d}\boldsymbol{\theta}.$$

Hence, the marginal of the artificial distribution is proportional to the posterior expected utility. Therefore, the maximum expected utility alternative coincides with the mode of the marginal of the artificial distribution $h(a, \boldsymbol{\theta})$ in the space of alternatives. This suggests that we may solve our problem approximately with the following simulation-based procedure:

1. Generate a sample $\left((\boldsymbol{\theta}^1, a^1), \ldots, (\boldsymbol{\theta}^m, a^m)\right)$ from the density $h(a, \boldsymbol{\theta})$.
2. Convert this to a sample $(a^1, \ldots, a^m)$ from the marginal density $h(a)$.
3. Find the sample mode of the marginal sample.

For Step 3 of the above algorithm, we can use tools from exploratory data analysis (see, e.g., Scott, 1992). To implement Step 1, we can appeal to MCMC methods, as illustrated in Bielza *et al*. (1999).

Decisions are seldom taken singly, but are more often nested within a series of related decisions, especially when stochastic processes are involved. This is the field of sequential decision analysis. Clearly, we cannot review this field in totality in a single section, and therefore, here we only provide the key ideas, including Bellman's principle of optimality, with a relevant statistical problem referring to sequential sampling. Other ideas are illustrated in some of the later chapters.

Assume that a DM faces a decision problem with state space $\boldsymbol{\Theta}$. The problem is structured into a number of stages, say $n = 1, 2, \ldots, N$. At each stage, the DM may choose to make an observation X_n or stop sampling and take an action from a set $\mathcal{A}$. We shall assume that the X_i are conditionally IID. If she makes an observation at stage n she incurs a cost γ. If she stops sampling and takes an action $a \in \mathcal{A}$, she incurs a consequence $c(a, \boldsymbol{\theta})$ if $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is the current state. Therefore, if the DM samples to stage n and then takes action a, she will receive a timestream of outcomes: $(-\gamma, -\gamma, \ldots, -\gamma, c(a, \boldsymbol{\theta}))$. We assume that her utility function over these timestreams is additive:

$$u(\gamma, \gamma, \ldots, \gamma, c(a, \boldsymbol{\theta})) = -(n-1)\gamma + u(a, \boldsymbol{\theta}),$$

where $u(\cdot, \cdot)$ is her terminal utility resulting from taking action a given $\boldsymbol{\theta}$, and we assume that γ and the terminal utility are measured on the same scale. A decision in this problem will consist of two components: (1) a sampling plan, which determines whether she should keep sampling having seen a series of observations, $X_1 = x_1$, $X_2 = x_2, \ldots, X_n = x_n$; and (2) a sequence of decision rules, which specify, in the light of the observations, which action she should take when she stops sampling.

Bellman (1957) presented a principle of optimality that simplifies the formulation and also points to a characterization of the solution in the above problem. Notice first that the utility function used is *monotonic* and *separable*: if the first few terms are fixed, then maximizing the whole sum is equivalent to maximizing the remaining terms, and this property will hold when expectations are taken. This observation leads to Bellman's principle of optimality:

> The optimal sequential decision policy for the problem that begins with $P(\cdot)$ as the DM's prior for $\boldsymbol{\theta}$ and has R stages to run must have the property that if, at any stage $n < N$ the observations $X_1 = x_1$, $X_2 = x_2, \ldots, X_n = x_n$ have been made, then the continuation of the optimal policy must be the optimal sequential policy for the problem beginning with $P(\cdot \mid x_1, x_2, \ldots, x_n)$ as her prior and having $(N - n)$ stages to run.

Let

$$f = P(\cdot)$$

and, generally, for $n = 1, 2, \ldots, N$

$$f(x_1, x_2, \ldots, x_n) = P(\cdot \mid x_1, x_2, \ldots, x_n).$$

We shall also use $f(X_1, X_2, \ldots, X_n)$ in expectations over future observations as indicating her expected future state of knowledge given the yet unknown observations. We then let $r_n(f)$ be the expected utility of the optimal policy with at most n stages left to run when she begins with knowledge f. We can now write down some simple recursions.

First, in circumstances when she has no option to take an observation and she must choose an action $a \in \mathcal{A}$,

$$r_0(f) = \max_{a \in \mathcal{A}} E[u(a, \boldsymbol{\theta})],$$

where the expectation over $\boldsymbol{\theta}$ is taken with respect to f. Next imagine that she has the opportunity to take at most one observation and that currently her knowledge of $\boldsymbol{\theta}$ is given by f. Then:

- either she takes an action $a \in \mathcal{A}$ immediately with expected utility of $r_0(f)$, since she would be using the optimal action;

- or she makes a single observation X at a cost γ and then chooses an action $a \in \mathcal{A}$ in the light of her current knowledge $f(X)$ at an expected utility of $E_X[r_0(f(X))]$, since she would be using the optimal action.

Note that Bellman's principle of optimality is used in the assumption that, after the observation, she takes the optimal action for the state of knowledge in which she then finds herself. It follows that she will choose whichever of these leads to the bigger expected utility. So,

$$r_1(f) = \max\{r_0(f), -\gamma + E_X[r_0(f(X))]\}.$$

A similar argument shows that if she has the opportunity to take at most n further observations and her current state of knowledge is f,

$$r_n(f) = \max\{r_0(f), -\gamma + E_X[r_{n-1}(f(X))]\}$$

for $n = 1, 2, \ldots, N$. The recursion together with the initial condition defines an iteration known as *backward induction* or (*backward*) dynamic programming, as it first determines what the DM should do at stage N, when no further observations may be made, then at stage $(N-1)$, and so on back to stage 1.

In principle at least, dynamic programming both allows the calculation of $r_n(f)$ and also defines the optimal policy. In practice, however, the computational complexity of the calculations may make the scheme intractable. Dynamic programming algorithms suffer from the '*curse of dimensionality*': the optimization involved requires $r_{n-1}(f(X))$ for all conceivable posterior distributions $f(X)$ for θ. This can be an enormous computational demand, except in particular cases in which either the dimensionality of Θ is heavily restricted and/or in which $r_{n-1}(\cdot)$ can be expressed analytically. Fortunately, the characterization of the expected utility is sometimes sufficient to enable iterative schemes, such as *value iteration* and *policy iteration*, which allow an optimal – or approximately optimal – policy to be identified.

2.5 Discussion

This has been a brief introduction to Bayesian concepts and methods that we shall be using in dealing with stochastic process models. We have placed emphasis in computational and decision analytic aspects that are key in applied settings for stochastic processes and focus our book.

There is now quite a large literature on Bayesian analysis that details the material presented here. One of the early works on conjugate Bayesian inference is Box and Tiao (1973). More modern approaches emphasizing the inferential aspects of the Bayesian approach are, for example, Gelman *et al*. (2003), Lee (2004), or Carlin and Louis (2008). Statistical decision theory is well described in, for example, Berger (1985), Robert (1994), Bernardo and Smith (1994), and French and Ríos Insua (2000) and decision analytic aspects are covered in, for example, Clemen and Reilly (2004). We have only considered the case of single experiments. When observations are taken from related variables, the Bayesian way of connecting these variables and borrowing

strength is via the use of hierarchical or empirical Bayes models. Good sources to the literature on such models are, for example, Gelman *et al*. (2003), Congdon (2010), and Efron (2010).

An important problem in subjective Bayesian analysis that we have not commented about in this chapter is how to elicit probabilities or utilities from experts. The problems of prior elicitation will sometimes be considered in later chapters and a good source to the elicitation literature is, for example, O'Hagan *et al*. (2006). Also, we have only briefly commented on robustness and sensitivity issues here. A much fuller review on the Bayesian robustness literature is given in, for example, Ríos Insua and Ruggeri (2000) and the many references therein. As we have noted, when little information is available, the alternative to the use of expert priors is the use of noninformative priors. Good reviews of the various objective priors available and the advantages of the objective Bayesian philosophy are given in, for example, Berger (1985, 2006). In contrast, the subjective Bayesian approach is championed in Goldstein (2006) and good comparisons of the subjective and objective Bayesian approaches are given in, for example, Press (2002) and the discussion of Goldstein (2006). Utility elicitation is described in Farquhar (1984).

We have mentioned point estimation, interval estimation, hypothesis testing, and prediction as the key relevant inferential problems. Related problems such as model selection or experimental design are also important and, in particular, model selection can be thought of as a generalization of hypothesis testing. Jeffreys (1961) is a seminal work on both topics and Kass and Raftery (1996) provide a comprehensive survey on the use of Bayes factors as a key tool for model selection; see also Bernardo and Smith (1994). The related problem of experimental design is also covered in, for example, Chaloner and Verdinelli (1995). Good reviews of Bayesian predictive inference are given by Aitchison and Dunsmore (1975) and Geisser (1993).

The literature on modern integration methods is vast. Some reviews of Bayesian Monte Carlo or MCMC methods are given in, for example, Gamerman and Lopes (2006) or Casella and Robert (2010). Sequential Monte Carlo methods are also covered in great detail by, for example, Liu (2001) or Doucet *et al*. (2010). Variable dimension MCMC methods are also very important, including reversible jump samplers, as in Green (1995) or Richardson and Green (1997). Also, the class of dynamic models, briefly mentioned here, see, for example, West and Harrison (1996) and Petris *et al*. (2009), provide powerful tools for forecasting large classes of time series models. ROMs are presented in Grigoriu (2009).

In low-dimensional parameter problems, methods like quadrature (Naylor and Smith, 1982) or asymptotic approximations (Lindley 1980, Tierney and Kadane 1986) provide good results. This last class of methods are based on asymptotic properties such as large sample normality of the posterior distribution under certain technical conditions (see, e.g., Le Cam, 1953).

General ideas on simulation may be seen in, for example, Ripley (1987). Their application in statistical contexts is well outlined in French and Ríos Insua (2000). The augmented simulation method described in Section 2.4.2 is based on Bielza *et al*. (1999) and Müller (1999), where several applications are illustrated.

A key reference on sequential statistical decision theory is DeGroot (1970); see, also, Berger (1985). Bellman (1957) is still essential reading on dynamic programming and an early Bayesian view of dynamic programming is given in Lindley (1961). Dynamic programming pervades many decision analytic algorithms including the evaluation of decision trees and influence diagrams (see Clemen and Reilly, 2004).

All the problems we have mentioned are parametric. By now, there is a plethora of Bayesian nonparametric and semiparametric methods. They may be seen as parametric problems in which the parameter space is infinite dimensional, with key roles for tools such as Dirichlet processes and their mixtures, and Polya tress. Some useful references are Dey *et al*. (1998) and, more recently, Ghosh and Ramamoorthi (2010) and Hjort *et al*. (2010).

Finally, we should note that the gambler's ruin problem has been used to introduce Bayesian analysis of stochastic processes in this chapter. A good introduction to the probabilistic theory of this problem is Edwards (1983). The Bayesian statistical approach to this problem and Bayesian robustness is analyzed in Tsay and Tsao (2003) and Bayesian asymptotics are studied in, for example, Ganesh and O'Connell (1999, 2000).

References

Aitchison, J. and Dunsmore, I.R. (1975) *Statistical Prediction Analysis*. Cambridge: Cambridge University Press.

Bellman (1957) *Dynamic Programming*. Princeton: Princeton University Press.

Berger, J.O. (1985) *Statistical Decision Theory and Bayesian Analysis*. Berlin: Springer.

Berger, J.O. (2006) The case for objective Bayesian analysis. *Bayesian Analysis*, **1**, 385–402.

Berger, J.O. and Ríos Insua, D. (1998) Recent developments in Bayesian inference with applications in hydrology. In *Statistical and Bayesian Methods in Hydrology*, Parent, Hubert, Miquel, and Bobee (Eds.). Paris: UNESCO Press, pp. 56–80.

Bernardo, J.M. and Smith, A.F.M. (1994) *Bayesian Theory*. New York: John Wiley & Sons, Inc.

Bielza, C., Müller, P., and Ríos Insua, D. (1999) Decision analysis by augmented probability simulation. *Management Science*, **45**, 995–1007.

Box, G.E. and Tiao, G.C. (1973) *Bayesian Inference in Statistical Analysis*. New York: John Wiley & Sons, inc.

Carlin, B.P. and Louis, T.A. (2008) *Bayesian Methods for Data Analysis*. Boca Raton: Chapman and Hall.

Carpenter, J., Clifford, P., and Fearnhead, P. (1999) Improved particle filters for nonlinear problems. *IEE Proceedings. Radar, Sonar and Navigation*, **146**, 2–7.

Casella, G. and Robert, C. (2010) *Monte Carlo Statistical Methods* (2nd edn.). Berlin: Springer.

Chaloner, K. and Verdinelli, I. (1995) Bayesian experimental design: a review. *Statistical Science*, **10**, 273–304.

Clemen, R.T. and Reilly, T. (2004) *Making Hard Decisions with Decision Tools*. Belmont, CA: Duxbury.

Congdon, P. (2010) *Applied Bayesian Hierarchical Methods*. London: Chapman and Hall.

DeGroot, M. (1970) *Optimal Statistical Decisions*. New York: McGraw-Hill.

Del Moral, P., Doucet, A., and Jasra, A. (2006) Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society B*, **68**, 411–436.

Dey, D.D., Müller, P., and Sinha, D. (1998) *Practical Nonparametric and Semiparametric Bayesian Statistics*. New York: Springer.

Doucet, A., de Freitas, N., and Gordon, N. (2010) *Sequential Monte Carlo Methods in Practice*. New York: Springer.

Edwards, A.W.F. (1983) Pascal's problem: the 'gambler's ruin'. *International Statistical Review*, **51**, 73–79.

Efron, B. (2010) *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge: Cambridge University Press.

Farquhar, P. (1984) Utility assessment methods. *Management Science*, **30**, 1283–1300.

French, S. and Ríos Insua, D. (2000) *Statistical Decision Theory*. London: Arnold.

Gamerman, D. and Lopes, H.F. (2006) *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Boca Raton: Chapman and Hall.

Ganesh, A.J. and O'Connell, N. (1999) An inverse of Sanov's theorem. *Statistics and Probability Letters*, **42**, 201–206.

Ganesh, A.J. and O'Connell, N. (2000) A large-deviation principle for Dirichlet posteriors. *Bernoulli*, **6**, 1021–1034.

Geisser, S. (1993) *Predictive Inference: An Introduction*. New York: Chapman and Hall.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003) *Bayesian Data Analysis* (2nd edn.). New York: Chapman and Hall.

Ghosh, J.K. and Ramamoorthi, R.V. (2010) *Bayesian Nonparametrics*. New York: Springer.

Goldstein, M. (2006) Subjective Bayesian analysis: principles and practice (with discussion). *Bayesian Analysis*, **1**, 403–472.

Gordon, N.J., Salmond, D.J., and Smith, A.F.M. (1993) A novel approach to non-linear and non-Gaussian state estimation. *IEE Proceedings F*, **140**, 107–133.

Green, P. (1995) Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Grigoriu, M. (2009) Reduced order models for random functions. Applications to stochastic problems. *Applied Mathematical Modelling*, **33**, 161–175.

Hjort, N.L., Holmes, C., Müller, P., and Walker, S.G. (2010). *Bayesian Nonparametrics*. Cambridge: Cambridge University Press.

Jeffreys, H. (1961) *Theory of Probability* (3rd edn.). Oxford: Oxford University Press.

Kass, R. and Raftery, A. (1990) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Le Cam, L. (1953) On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. *University of California Publications in Statistics*, **1**, 277–328.

Lee, P. (2004) *Bayesian Statistics: An Introduction* (3rd edn.). Chichester: John Wiley & Sons, Ltd.

Lindley, D.V. (1961) Dynamic programming and decision theory. *Applied Statistics*, **10**, 39–51.

Lindley, D.V. (1980) Approximate Bayesian methods. In *Bayesian Statistics*, J.M. Bernardo, M.H. De Groot, D.V. Lindley, and A.F.M. Smith (Eds.). Valencia: University Press.

Liu, J.S. (2001) *Monte Carlo Strategies in Scientific Computing*. New York: Springer.

Müller, P. (1991) A generic approach to posterior integration and Gibbs sampling. *Technical Report, Department of Statistics, Purdue University.*

Müller, P. (1999) Simulation based optimal design. In *Bayesian Statistics 6*, J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith (Eds.). Oxford: Oxford University Press, pp. 459–474.

Naylor, J.C. and Smith, A.F.M. (1982) Application of a method for the efficient computation of posterior distributions. *Applied Statistics*, **31**, 214–225.

Nemhauser, G., Rinooy Kan, A., and Todt, J. (1990) *Optimization*. Amsterdam: North Holland.

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., and Rakow, T. (2006) *Uncertain Judgements: Eliciting Experts' Probabilities*. Chichester: John Wiley & Sons, Ltd.

Petris, G., Petrone, S., and Campagnoli, P. (2009) *Dynamic Linear Models with R*. New York: Springer.

Pitt, M. and Shephard, N. (1999) Filtering via simulation: Auxiliary particle filtering *Journal of the American Statistical Association*, **94**, 590–599.

Press, S.J. (2002) *Subjective and Objective Bayesian Statistics: Principles, Models and Applications*. New York: John Wiley & Sons, Inc.

Richardson, S. and Green, P. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society B*, **59**, 731–792.

Ríos Insua, D., Bielza, C., Muller, P., and Salewicz, K. (1997) Bayesian methods in reservoir operations. In *The Practice of Bayesian Analysis*, S. French and J.Q. Smith (Eds.). London: Arnold, pp. 107–130.

Ríos Insua, D. and Ruggeri, F. (2000) *Robust Bayesian Analysis*. New York: Springer.

Ripley, B.D. (1987) *Stochastic Simulation*. New York: John Wiley & Sons, Inc.

Robert, C. (2001) *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation* (2nd edn.). New York: Springer.

Singpurwalla, N.D. and Wilson, S. (1999) *Statistical Methods in Software Engineering*. New York: Springer.

Scott, D. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley & Sons, Inc.

Shao, J. (1989) Monte Carlo approximations in Bayesian decision theory. *Journal of the American Statistical Association*, **84**, 727–732.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and Van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B*, **64**, 583–639.

Tierney, L. (1994) Markov chains for exploring posterior distributions, *Annals of Statistics*, **22**, 1701–1728.

Tierney, L. and Kadane, J. (1986) Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.

Tsay, J.J. and Tsao, C.A. (2003) Statistical gambler's ruin problem. *Communications in Statistics: Theory and Methods*, **32**, 1377–1359.

West, M. and Harrison, P.J. (1997) *Bayesian Forecasting and Dynamic Models* (2nd edn.). Berlin: Springer.

Part Two

MODELS

3

Discrete time Markov chains and extensions

3.1 Introduction

As we mentioned in Chapter 1, Markov chains are one of the simplest stochastic processes to study and are characterized by a lack of memory property, so that future observations depend only on the current state and not on the whole of the past history of the process. Despite their simplicity, Markov chains can be and have been applied to many real problems in areas as diverse as web-browsing behavior, language modeling, and persistence of surnames over generations. Furthermore, as illustrated in Chapter 2, with the development of Markov chain Monte Carlo (MCMC) methods, Markov chains have become a basic tool for Bayesian analysis.

In this chapter, we shall study the Bayesian analysis of discrete time Markov chains, focusing on homogeneous chains with a finite state space. We shall also analyze many important subclasses and extensions of this basic model such as reversible chains, branching processes, higher order Markov chains, and discrete time Markov processes with continuous state spaces. The properties of the basic Markov chain model and these variants are outlined from a probabilistic viewpoint in Section 3.2.

In Section 3.3, inference for time homogeneous, discrete state space, first-order chains is considered. Then, Section 3.4 provides inference for various extensions and particular classes of chains. A case study on the analysis of wind directions is presented in Section 3.5 and Markov decision processes are studied in Section 3.6. The chapter concludes with a brief discussion.

Bayesian Analysis of Stochastic Process Models, First Edition. David Rios Insua, Fabrizio Ruggeri and Michael P. Wiper.

3.2 Important Markov chain models

This chapter analyzes Bayesian inference and prediction for data generated from discrete time Markov chains. In Chapter 1, we defined these as discrete time discrete space stochastic processes $\{X_n\}$, which possess the Markov property. In this chapter, we shall focus on finite, time homogeneous Markov chains in detail. For such a chain, with states $\{1, \ldots, K\}$, we shall write the transition matrix as $\boldsymbol{P} = (p_{ij})$, where $p_{ij} = \boldsymbol{P}(X_n = j | X_{n-1} = i)$, for $i, j \in \{1, \ldots, K\}$. Should it exist, the stationary distribution $\boldsymbol{\pi}$ is the unique solution of $\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}$, $\pi_i \geq 0$, $\sum \pi_i = 1$. There are many extensions of this basic model that are analyzed in the following text.

3.2.1 Reversible chains

Most Markov chains considered in the context of MCMC have the property of reversibility.

Definition 3.1: *A Markov chain with transition probabilities p_{ij} for $i, j = 1, \ldots, K$ is reversible if there exists a probability distribution π that satisfies the detailed balance equation for any i, j*

$$p_{ij}\pi(j) = p_{ji}\pi(i).$$

For a reversible Markov chain, it can be immediately demonstrated that π is its stationary distribution. Inference for reversible Markov chains is examined in Section 3.4.1.

3.2.2 Higher order chains and mixtures

Generalizing from Definition 1.6, a discrete time stochastic process, $\{X_n\}$ is a Markov chain of order r if $P(X_n = x_n | X_0 = x_0, \ldots, X_{n-1} = x_{n-1}) = P(X_n = x_n | X_{n-r} = x_{n-r}, \ldots, X_{n-1} = x_{n-1})$ so that the state of the chain is determined by the previous r states. It is possible to represent such a chain as first-order chain by simply combining states.

Example 3.1: Consider a second-order, homogeneous Markov chain $\{X_n\}$ with two possible states (1 and 2) and write $p_{ijl} = P(X_n = l | X_{n-1} = j, X_{n-2} = i)$ for i, j, $l = 1, 2$. Then the first-order transition matrix is

$$\begin{array}{c} \\ (1,1) \\ (1,2) \\ (2,1) \\ (2,2) \end{array} \begin{array}{c} \begin{array}{cccc} (1,1) & (1,2) & (2,1) & (2,2) \end{array} \\ \begin{pmatrix} p_{111} & p_{112} & 0 & 0 \\ 0 & 0 & p_{121} & p_{122} \\ p_{211} & p_{212} & 0 & 0 \\ 0 & 0 & p_{221} & p_{222} \end{pmatrix} \end{array}$$

△

The disadvantage of modeling higher order Markov chain models in such a way is that the number of states necessary to reduce such models to a first-order Markov chain is large. For example, if X_n can take values in $\{1, \ldots, K\}$, then K^r states are needed to define an rth order chain. Therefore, various alternative approaches to modeling rth order dependence have been suggested. One of the most popular ones is the mixture transition distribution (MTD) model of Raftery (1985). In this case, it is assumed that

$$P(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_{n-r} = x_{n-r}) = \sum_{i=1}^{r} w_i \, p_{x_{n-i} x_n}, \tag{3.1}$$

where $\sum_{i=1}^{r} w_i = 1$ and $\boldsymbol{P} = (p_{ij})$ is a transition matrix. This approach leads to more parsimonious modeling than through the full rth order chain. In particular, in Example 3.1, four free parameters are necessary to model the full second-order chain, whereas using the MTD model only three free parameters are necessary. Inference for higher order Markov chains and for the MTD model is examined in Section 3.4.2.

3.2.3 Discrete time Markov processes with continuous state space

As noted in Chapter 1, Markov processes can be defined with both discrete and continuous state spaces. We have seen that for a Markov chain with discrete state space, the condition for the chain to have an equilibrium distribution is that the chain is aperiodic and that all states are positive recurrent. Although the condition of positive recurrence cannot be sensibly applied to chains with continuous state space, a similar condition known as Harris recurrence applies to chains with continuous state space, which essentially means that the chain can get close to any point in the future. It is known that Harris recurrent, aperiodic chains also possess an equilibrium distribution, so that if the conditional probability distribution of the chain is $P(X_n|X_{n-1})$, then the equilibrium density π satisfies

$$\pi(x) = \int P(x|y)\pi(y)\,\mathrm{d}y.$$

As with Markov chains with discrete state space, a sufficient condition for a process to possess an equilibrium distribution is to be reversible.

Example 3.2: Simple examples of continuous space Markov chain models are the autoregressive (AR) models. The first-order AR process was outlined in Example 1.1. Higher order dependence can also be incorporated. An AR(k) model is defined by

$$X_n = \phi_0 + \sum_{i=1}^{k} \phi_i X_{n-i} + \epsilon_n.$$

The condition for this process to be (weakly) stationary is the well-known unit roots condition that all roots of the polynomial

$$\phi_0 z^k - \sum_{i=1}^{k} \phi_i z^{k-i}$$

must lie within the unit circle, that is, each root z_i must satisfy $|z_i| < 1$. △

Inference for AR processes and other continuous state space processes is briefly reviewed in Section 3.4.3.

3.2.4 Branching processes

The Bienaymé–Galton–Watson branching process was originally introduced as a model for the survival of family surnames over generations and has later been applied in areas such as survival of genes. The process is defined as follows. Assume that at time 0, a population consists of a single individual who lives for a single time unit and then dies and is replaced by his offspring. These offspring all survive for a further single time unit and are then replaced by their offspring, and so on.

Formally, define Z_n to be the population after time n. Then, $Z_0 = 1$. Also let X_{ij} be the number of offspring born to the jth individual in generation i. Assume that the X_{ij} are all independent and identically distributed variables, $X_{ij} \sim X$, with some distribution $P(X = x) = p_x$ for $x = 0, 1, 2, \ldots$ where we assume that $p_0 > 0$. Then,

$$Z_{n+1} = \sum_{j=1}^{Z_n} X_{nj}.$$

Interest is usually focused on the probability γ of extinction,

$$\gamma = P(Z_n = 0, \text{ for some } n = 1, 2, \ldots). \tag{3.2}$$

It is well known that extinction is certain if $\theta = E[X] \leq 1$. Otherwise, γ is the smallest root of the equation $G(s) = s$, where $G(s)$ is the probability generating function of X (see Appendix B). Obviously, if the initial population is of size $k > 1$, then the probability of eventual extinction is γ^k.

Inference for branching processes is provided in Section 3.4.4.

3.2.5 Hidden Markov models

Hidden Markov models (HMMs) have been widely applied to the analysis of weakly dependent data in diverse areas such as econometrics, ecology, and signal processing. A hidden Markov model is defined as follows. Observations Y_n for $n = 0, 1, 2, \ldots$ are generated from a conditional distribution $f(y_n|X_n)$ with parameters that depend on an unobserved or hidden state, $X_n \in \{1, 2, \ldots K\}$. The hidden states follow a Markov

chain with transition matrix $\boldsymbol{P}$ and an initial distribution, usually assumed to be the equilibrium distribution, $\pi(\cdot|\boldsymbol{P})$, of the underlying Markov chain.

The architecture of this process can be represented by an influence diagram as in Figure 3.1, with arrows denoting conditional dependencies.

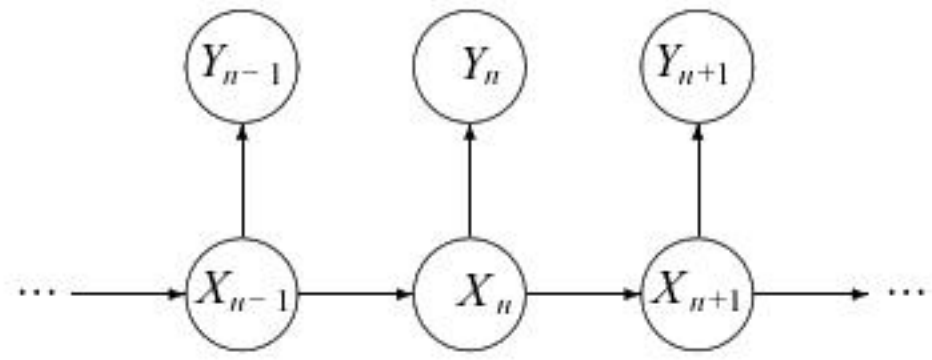

Figure 3.1 Influence diagram representing the dependence structure of a HMM.

In the preceding text, we are assuming that the hidden state space of the HMM is discrete. However, it is straightforward to extend the definition to HMMs with a continuous state space. A simple example is the dynamic linear model described in Section 2.4.1. Inference for HMMs is overviewed in Section 3.4.5.

3.3 Inference for first-order, time homogeneous, Markov chains

In this section, we study inference for a first-order, time homogeneous, Markov chain, $\{X_n\}$, with state space $\{1, 2, \ldots, K\}$ and (unknown) transition matrix $\boldsymbol{P}$.

Initially, we consider the simple experiment of observing m successive transitions of the Markov chain, say $X_1 = x_1, \ldots, X_m = x_m$, given a known initial state $X_0 = x_0$. In this case, the likelihood function is

$$l(\boldsymbol{P}|\mathbf{x}) = \prod_{i=1}^{K}\prod_{j=1}^{K} p_{ij}^{n_{ij}}, \tag{3.3}$$

where $n_{ij} \geq 0$ is the number of observed transitions from state i to state j and $\sum_{i=1}^{K}\sum_{j=1}^{K} n_{ij} = m$.

Given the likelihood function (3.3), it is easy to show that the classical, maximum likelihood estimate for $\boldsymbol{P}$ is $\hat{\boldsymbol{P}}$ with i, jth element equal to the proportion of transitions from state i that go to state j, that is,

$$\hat{p}_{ij} = \frac{n_{ij}}{n_{i\cdot}}, \quad \text{where} \quad n_{i\cdot} = \sum_{j=1}^{K} n_{ij}.$$

However, especially in chains where the number K of states is large and, therefore, a very large number K^2 of transitions are possible, it will often be the case that there are no observed transitions between various pairs, (i, j), of states and thus $\hat{p}_{ij} = 0$.

3.3.1 Advantages of the Bayesian approach

Obviously, a Bayesian approach using a prior distribution for **P** with mass on irreducible, aperiodic chains eliminates the possible problems associated with classical inference. Another, more theoretical justification of the use of a Bayesian approach to inference for Markov chains can be based on de Finetti type theorems.

The well-known de Finetti (1937) theorem states that for an infinitely exchangeable sequence, $X_1, X_2, \ldots$ of zero-one random variables with probability measure P, there exists a distribution function F such that the joint mass function is

$$p(x_1, \ldots x_n) = \int_{\boldsymbol{\theta}} \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i} \mathrm{d}F(\boldsymbol{\theta}).$$

Obviously, observations from a Markov chain cannot generally be regarded as exchangeable and so the basic de Finetti theorem cannot be applied. However, an appropriate definition of exchangeability is to say that a probability measure P defined on recurrent Markov chains is partially exchangeable if it gives equal probability to all sequences $X_1, \ldots, X_n$ (assuming some fixed x_0) with the same transition count matrix. Given this definition of exchangeability, it can be shown that for a finite sequence, say $\mathbf{x} = (x_1, \ldots, x_n)$, there exists a distribution function F so that

$$p(\mathbf{x}|x_0) = \int_{\boldsymbol{P}} p_{ij}^{n_{ij}} \mathrm{d}F(\boldsymbol{P}),$$

where n_{ij} are the transition counts. Similar to the standard de Finetti theorem, the distribution F may be interpreted as a Bayesian prior distribution for $\boldsymbol{P}$.

3.3.2 Conjugate prior distribution and modifications

Given the experiment of this Section, a natural conjugate prior for $\boldsymbol{P}$ is defined by letting $\mathbf{p}_i = (p_{i1}, \ldots, p_{iK})$ have a Dirichlet distribution, say

$$\mathbf{p}_i \sim \mathrm{Dir}(\boldsymbol{\alpha}_i), \quad \text{where } \boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iK}) \text{ for } i = 1, \ldots, K.$$

This defines a matrix beta prior distribution. Given this prior distribution and the likelihood function of (3.3), the posterior distribution is also of the same form, so that

$$\mathbf{p}_i|\mathbf{x} \sim \mathrm{Dir}(\boldsymbol{\alpha}_i') \quad \text{where } \alpha_{ij}' = \alpha_{ij} + n_{ij} \text{ for } i, j = 1, \ldots, K. \tag{3.4}$$

When little prior information is available, a natural possibility is to use the Jeffreys prior, which is a matrix beta prior with $\boldsymbol{\alpha}_{ij} = 1/2$ for all $i, j = 1, \ldots, K$. An

alternative, improper prior distribution along the lines of the Haldane (1948) prior for binomial data is to set

$$f(\mathbf{p}_i) \propto \prod_{j=1}^{K} \frac{1}{p_{ij}},$$

which can be thought of as the limit of a matrix beta prior, setting $\alpha_{ij} \to 0$ for all $i, j = 1, \ldots, K$. In this case, the posterior distribution is $\mathbf{p}_i|\mathbf{x} \sim \text{Dir}(n_{i1}, \ldots, n_{ik})$ so that, for example, the posterior mean of the ijth element of the transition matrix is $E[p_{ij}|\mathbf{x}] = n_{ij}/n_{i\cdot}$, equal to the maximum likelihood estimate. However, this approach cannot be recommended, as if any $n_{ij} = 0$, which may often be the case for chains with a relatively large number of states, then the posterior distribution is improper.

Example 3.3: Rainfall levels at the Sydney Botanic Gardens weather center in Australia have been recorded for some time. The following data taken from weatherzone.com.au illustrate the occurrence (2) or nonoccurrence (1) of rain between February 1 and March 20, 2008. The data are to be read consecutively from left to right. Thus, it rained on February 1st and did not rain on March 20th.

2	2	2	2	2	2	2	2	2	2
1	1	2	1	1	1	1	1	1	1
2	2	1	1	1	1	2	2	2	1
2	1	1	1	1	1	2	1	1	1
1	1	1	1	1	1	1	1	1	

Assume that the daily occurrence of rainfall is modeled as a Markov chain with transition matrix

$$\boldsymbol{P} = \begin{pmatrix} p_{11} & 1 - p_{11} \\ 1 - p_{22} & p_{22} \end{pmatrix}.$$

Given a Jeffreys prior, $p_{ii} \sim \text{Be}(1/2, 1/2)$, for $i = 1, 2$, then conditioning on the occurrence of rainfall on February 1st, the posterior distribution is

$$p_{11}|\mathbf{x} \sim \text{Be}(25.5, 5.5) \quad p_{22}|\mathbf{x} \sim \text{Be}(12.5, 6.5).$$

The expectation of the transition matrix is

$$E[\boldsymbol{P}|\mathbf{x}] = \begin{pmatrix} 0.823 & 0.177 \\ 0.342 & 0.658 \end{pmatrix}.$$

△

In some cases, it may be known that certain transitions are impossible a priori. For example, it may be impossible to remain in a state, so that $p_{ii} = 0$ for $i = 1, \ldots, K$. Obviously, it is straightforward to include these types of constraints by simply restricting the matrix beta prior to the space of transitions with positive

probability and setting the remaining transition probabilities to zero, when inference remains fully conjugate as above. In other cases, the elements of $\boldsymbol{P}$ may all depend on some common probabilities as in Example 1.2. As we have seen in Example 2.2, this case is also easily dealt with.

A more interesting problem that has been little studied in the literature is the situation where, a priori, it is unknown which transitions are possible and which are impossible so that the chain may be periodic or transient. In this situation, one possibility is to define a hierarchical prior distribution by first setting the probabilities that different transitions are impossible as follows:

$$\begin{aligned} P(p_{ij} = 0|q) &\propto q \quad \text{for } i, j \in 1, \ldots, K \\ q &\sim \mathrm{U}(0, 1), \end{aligned}$$

where this prior is restricted so that $P(\mathbf{p}_i = \mathbf{0}|q) = 0$ so that, for example, the prior probability (conditional on q) that row i of the transition matrix contains exactly r_i zeros at locations $j_1, \ldots, j_{r_i}$ and $K - r_i$ ones at locations $j_{r_i+1}, \ldots, j_K$ is given by

$$\frac{q^{r_i}(1-q)^{K-r_i}}{1-q^K} \quad \text{for } r = 0, 1, \ldots, K-1.$$

Given the set of possible transitions from state i, to $j_1, \ldots, j_r$ say, a Dirichlet prior can be defined for the vector of transition probabilities, for example,

$$(p_{ij_1}, \ldots, p_{ij_r}) \sim \mathrm{Dir}\left(\underbrace{\frac{1}{2}, \ldots, \frac{1}{2}}_{r}\right).$$

Now, let $\boldsymbol{Z}$ be a random, $K \times K$ matrix such that $Z_{ij} = 0$ if $p_{ij} = 0$, and, otherwise, $Z_{ij} = 1$. Assume that $\mathbf{z}$ is a matrix where the ith row of $\mathbf{z}$ contains r_i zeros in positions $j_1, \ldots, j_r$ for $i = 1, \ldots, K$. Then, the posterior probability that $\boldsymbol{Z}$ is equal to $\mathbf{z}$ can be evaluated as

$$\begin{aligned} P(\boldsymbol{Z} = \mathbf{z}|\mathbf{x}) &\propto f(\mathbf{x}|\mathbf{z})P(\mathbf{z}) \\ &\propto \int f(\mathbf{x}|\mathbf{z}, \boldsymbol{P}) f(\boldsymbol{P}|\mathbf{z})\, \mathrm{d}\boldsymbol{P} \int_0^1 P(\boldsymbol{Z} = \mathbf{z}|q) f(q)\, \mathrm{d}q \\ &\propto \frac{1}{\Gamma\left(\frac{1}{2}\right)^{K^2 - K\bar{r}}} \prod_{i=1}^{K} \frac{\Gamma\left(\frac{K-r_i}{2}\right)}{\Gamma\left(\frac{K-r_i}{2} + \sum_{s=r_i+1}^{K} n_{ij_s}\right)} \times \\ &\quad \prod_{s=r_i+1}^{K} \Gamma\left(\frac{1}{2} + n_{ij_s}\right) \int_0^1 \frac{q^{K\bar{r}}(1-q)^{K(K-\bar{r})}}{(1-q^K)^K}\, \mathrm{d}q, \end{aligned}$$

where the probability is positive over the range $n_{ij_1}, \ldots, n_{ij_{r_i}} = 0$ for $i = 1, \ldots, K$.

For relatively small dimensional transition matrices, this probability may be evaluated directly, but for Markov chains with a large number of states and many values $n_{ij} = 0$, exact evaluation will be impossible. In such cases, it would be preferable to employ a sampling algorithm over values of $\boldsymbol{Z}$ with high probability. The posterior probability that the chain is periodic, or transient, could then be evaluated by simply summing those $P(\boldsymbol{Z} = \mathbf{z}|\mathbf{x})$, where $\mathbf{z}$ is equivalent to a periodic transition matrix.

3.3.3 Forecasting short-term behavior

Suppose that we wish to predict future values of the chain. For example, we can predict the next value of the chain, at time $n + 1$ using

$$P(X_{n+1} = j|\mathbf{x}) = \int P(X_{n+1} = j|\mathbf{x}, \boldsymbol{P}) f(\boldsymbol{P}|\mathbf{x}) \mathrm{d}\boldsymbol{P}$$
$$= \int p_{x_n j} f(\boldsymbol{P}|\mathbf{x}) \mathrm{d}\boldsymbol{P} = \frac{\alpha_{x_n j} + n_{x_n j}}{\alpha_{x_n \cdot} + n_{x_n \cdot}},$$

where $\alpha_{i\cdot} = \sum_{j=1}^{K} \alpha_{ij}$.

Prediction of the state at $t > 1$ steps is slightly more complex. For small t, we can use

$$P(X_{n+t} = j|\mathbf{x}) = \int \left(\boldsymbol{P}^t\right)_{x_n j} f(\boldsymbol{P}|\mathbf{x}) \mathrm{d}\boldsymbol{P},$$

which gives a sum of Dirichlet expectation terms. However, as t increases, the evaluation of this expression becomes computationally infeasible. A simple alternative is to use a Monte Carlo algorithm based on simulating future values of the chain as follows:

For $s = 1, \ldots, S$:

Generate $\boldsymbol{P}^{(s)}$ from $f(\boldsymbol{P}|\mathbf{x})$.
Generate $x_{n+1}^{(s)}, \ldots, x_{n+t}^{(s)}$ from the Markov chain with $\boldsymbol{P}^{(s)}$ and initial state x_n.

Then, $P(X_{n+t} = j|\mathbf{x}) \approx \frac{1}{S} \sum_{s=1}^{S} I_{x_{n+t}^{(s)} = j}$ where I is an indicator function and $E[X_{n+t}|\mathbf{x}] \approx \frac{1}{S} \sum_{s=1}^{S} x_{n+t}^{(s)}$.

Example 3.4: Assume that it is now wished to predict the Sydney weather on March 21 and 22. Given that it did not rain on March 20, then immediately, we have

$$P(\text{no rain on March 21}|\mathbf{x}) = E[p_{11}|\mathbf{x}] = 0.823,$$
$$P(\text{no rain on March 22}|\mathbf{x}) = E\left[p_{11}^2 + p_{12}p_{21}|\mathbf{x}\right] = 0.742,$$
$$P(\text{no rain on both}) = E[p_{11}^2|\mathbf{x}] = 0.681.$$

△

3.3.4 Forecasting stationary behavior

Often interest lies in the stationary distribution of the chain. For a low-dimensional chain where the exact formula for the equilibrium probability distribution can be derived, this is straightforward.

Example 3.5: Suppose that $K = 2$ and $\boldsymbol{P} = \begin{pmatrix} p_{11} & 1-p_{11} \\ 1-p_{22} & p_{22} \end{pmatrix}$. Then the equilibrium probability of being in state 1 can easily be shown to be

$$\pi_1 = \frac{1-p_{22}}{2-p_{11}-p_{22}}$$

and the predictive equilibrium distribution is

$$E[\pi_1|\mathbf{x}] = \int_0^1 \int_0^1 \frac{1-p_{22}}{2-p_{11}-p_{22}} f(p_{11}, p_{22}|\mathbf{x})\,\mathrm{d}x$$

which can be evaluated by simple numerical integration techniques. △

Example 3.6: In the Sydney rainfall example, we have

$$E[\pi_1|\mathbf{x}] = E\left[\frac{1-p_{22}}{2-p_{11}-p_{22}}\middle|\mathbf{x}\right] = 0.655$$

so that we predict that it does not rain on approximately 65% of the days at this weather center. △

For higher dimensional chains, it is simpler to use a Monte Carlo approach as earlier so that given a Monte Carlo sample $\boldsymbol{P}^{(1)}, \ldots, \boldsymbol{P}^{(S)}$ from the posterior distribution of $\boldsymbol{P}$, then the equilibrium distribution can be estimated as

$$E[\boldsymbol{\pi}|\mathbf{x}] \approx \frac{1}{S}\sum_{s=1}^{S} \boldsymbol{\pi}^{(s)},$$

where $\boldsymbol{\pi}^{(s)}$ is the stationary distribution associated with the transition matrix $\boldsymbol{P}^{(s)}$.

3.3.5 Model comparison

One may often wish to test whether the observed data are independent or generated from a first (or higher) order Markov chain. The standard method of doing this is via Bayes factors (see Section 2.2.2).

Example 3.7: Given the experiment proposed at the start of section 3.2, suppose that we wish to compare the Markov chain model ($\mathcal{M}_1$) with the assumption that the data

are independent and identically distributed with some distribution $\mathbf{q} = (q_1, \ldots, q_K)$, $(\mathcal{M}_2)$ where we shall assume a Dirichlet prior distribution,

$$\mathbf{q} \sim \text{Dir}(a_1, \ldots, a_K).$$

Then,

$$f(\mathbf{x}|\mathcal{M}_1) = \int f(\mathbf{x}|\boldsymbol{P}) f(\boldsymbol{P}|\mathcal{M}_1) \mathrm{d}\boldsymbol{P}$$
$$= \prod_{i=1}^{k} \frac{\Gamma(\alpha_{i\cdot})}{\Gamma(n_{i\cdot} + \alpha_{i\cdot})} \prod_{j=1}^{k} \frac{\Gamma(\alpha_{ij} + n_{ij}),}{\Gamma(\alpha_{ij})}$$

where $n_{i\cdot} = \sum_{j=1}^{k} n_{ij}$ and $\alpha_{i\cdot} = \sum_{j=1}^{K} \alpha_{ij}$. Also, under the independent model, we have

$$f(\mathbf{x}|\mathcal{M}_2) = \frac{\Gamma(a)}{\Gamma(a+n)} \prod_{i=1}^{K} \frac{\Gamma(a_i + n_{\cdot i})}{\Gamma(a_i)},$$

where $a = \sum_{i=1}^{K} a_i$ and $n_{\cdot i}$ is the number of times that event i occurs (discounting the initial state X_0). The Bayes factor can now be calculated as the ratio of the two marginal likelihood functions, as illustrated. △

Example 3.8: For the Australian rainfall data, assuming that the initial state is known and given the Jeffreys prior for the Markov chain model, the marginal likelihood is

$$f(\mathbf{x}|\mathcal{M}_1) = \left(\frac{\Gamma(1)}{\Gamma(1/2)^2}\right)^2 \frac{\Gamma(25.5)\Gamma(5.5)}{\Gamma(31)} \frac{\Gamma(6.5)\Gamma(12.5)}{\Gamma(19)}$$

and, taking logs, we have $\log f(\mathbf{x}|\mathcal{M}_1) \approx -28.60$.

For the independent model, $\mathcal{M}_2$, conditional on the initial state and assuming a beta prior, $q_1 \sim \text{Be}(1/2, 1/2)$, we have

$$f(\mathbf{x}|\mathcal{M}_2) = \frac{\Gamma(1)}{\Gamma(1/2)^2} \frac{\Gamma(31.5)\Gamma(17.5)}{\Gamma(49)}$$

so that $\log f(\mathbf{x}|\mathcal{M}_2) \approx -33.37$, which implies a strong preference for the Markovian model over the independent model. △

3.3.6 Unknown initial state

When the initial state, X_0, is not fixed in advance, to implement Bayesian inference, we need to define a suitable prior distribution for X_0. The standard approach is simply to assume a multinomial prior distribution, $P(X_0 = x_0|\boldsymbol{\theta}) = \theta_{x_0}$ where $0 < \theta_k < 1$ and

$\sum_{k=1}^{K} \theta_k = 1$. Then, we can define a Dirichlet prior for the multinomial parameters, say $\boldsymbol{\theta} \sim \text{Dir}(\boldsymbol{\gamma})$ so that, a posteriori, $\boldsymbol{\theta}|\mathbf{x} \sim \text{Dir}(\boldsymbol{\gamma}')$, with $\gamma'_{x_0} = \gamma_{x_0} + 1$ and, otherwise, $\gamma'_i = \gamma_i$ for $i \neq x_0$. Inference for $\boldsymbol{P}$ then proceeds as before.

An alternative approach, which may be reasonable if it is assumed that the chain has been running for some time before the start of the experiment, is to assume that the initial state is generated from the equilibrium distribution, $\boldsymbol{\pi}$, of the Markov chain. Then, making the dependence of $\boldsymbol{\pi}$ on $\boldsymbol{P}$ obvious, the likelihood function becomes

$$l(\boldsymbol{P}|\mathbf{x}) = \pi(x_0|\boldsymbol{P}) \prod_{i=1}^{K} \prod_{j=1}^{K} p_{ij}^{n_{ij}}.$$

In this case, simple conjugate inference is impossible but, given the same prior distribution for $\boldsymbol{P}$ as above, it is straightforward to generate a Monte Carlo sample of size S from the posterior distribution of $\boldsymbol{P}$ using, for example, a rejection sampling algorithm as follows:

For $s = 1, \ldots, S$:

> For $i = 1, \ldots, K$, generate $\tilde{\mathbf{p}}_i \sim \text{Dir}(\boldsymbol{\alpha}')$ with $\boldsymbol{\alpha}'$ as in (3.4).
> Set $\tilde{\boldsymbol{P}}$ to be the transition probability matrix with rows $\tilde{p}_1, \ldots, \tilde{p}_K$.
> Calculate the stationary probability function $\tilde{\boldsymbol{\pi}}$ satisfying $\tilde{\boldsymbol{\pi}} = \tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{P}}$.
> Generate $u \sim \text{U}(0, 1)$. If $u < \tilde{\pi}(x_0)$, set $\mathbf{P}^{(s)} = \tilde{\boldsymbol{P}}$. Otherwise repeat from Step 1.

Example 3.9: Returning to the Sydney rainfall example, assume now that the weather on February 1st was generated from the equilibrium distribution. Then, using a Monte Carlo sample of size 10000, we have

$$E[\boldsymbol{P}|\mathbf{x}] \approx \begin{pmatrix} 0.806 & 0.194 \\ 0.321 & 0.679 \end{pmatrix}$$

and $E[\pi_1|\mathbf{x}] \approx 0.618$, which are close to the results in Section 3.2.4. Also, recalculating the log likelihood for the Markovian model ($\mathcal{M}_1$) under this assumption, and assuming that the probability of rain on February 1st is the same as the other days under the independent model ($\mathcal{M}_2$), we now have that

$$\log f(\mathbf{x}|\mathcal{M}_1) \approx -29.66 \quad \log f(\mathbf{x}|\mathcal{M}_2) \approx -34.40$$

so that the conclusions remain the same as in Example 3.8. △

3.3.7 Partially observed data

Assume now that the Markov chain is only observed at a number of finite time points. Suppose, for example, that x_0 is a known initial state and that we observe $\mathbf{x}_o = (x_{n_1}, \ldots, x_{n_m})$, where $n_1 < \ldots < n_m \in N$. In this case, the likelihood function is

$$l(\boldsymbol{P}|\mathbf{x}_o) = \prod_{i=1}^{m} p_{n_{i-1}n_i}^{(t_i - t_{i-1})}$$

where $p_{ij}^{(t)}$ represents the (i, j)th element of the t step transition matrix, defined in Section 1.3.1. In many cases, the computation of this likelihood will be complex. Therefore, it is often preferable to consider inference based on the reconstruction of missing observations. Let $\mathbf{x}_m$ represent the unobserved states at times $1, \ldots, t_1 - 1, t_1 + 1, \ldots, t_{n-1} - 1, t_{n-1} + 1, \ldots, t_n$ and let $\mathbf{x}$ represent the full data sequence. Then, given a matrix beta prior, we have that $\boldsymbol{P}|\mathbf{x}$ is also matrix beta. Furthermore, it is immediate that

$$P(\mathbf{x}_m|\mathbf{x}_o, \boldsymbol{P}) = \frac{P(\mathbf{x}|\boldsymbol{P})}{P(\mathbf{x}_o|\boldsymbol{P})} \propto P(\mathbf{x}|\boldsymbol{P}), \tag{3.5}$$

which is easy to compute for given $\boldsymbol{P}, \mathbf{x}_m$. One possibility would be to set up a Metropolis within Gibbs sampling algorithm to sample from the posterior distribution of $\boldsymbol{P}$.

Such an approach is reasonable if the amount of missing data is relatively small. However, if there is much missing data, it will be very difficult to define an appropriate algorithm to generate data from $P(\mathbf{x}_m|\mathbf{x}_o, \boldsymbol{P})$ in (3.5). In such cases, one possibility is to generate the elements of $\mathbf{x}_m$ one by one, using individual Gibbs steps. Thus, if t is a time point amongst the times associated with the missing observations, then we can generate a state x_t using

$$P(x_t|\mathbf{x}_{-t}, \boldsymbol{P}) \propto p_{x_{t-1}x_t} p_{x_t x_{t+1}}$$

where $\mathbf{x}_{-t}$ represents the complete sequence of states except for the state at time t.

Example 3.10: For the Sydney rainfall example, total rainfall was observed for March 21 and 22. From these data, it can be assumed that it rained on at least one of these two days. In this case, the likelihood function, including this data, becomes

$$l(\boldsymbol{P}|\mathbf{x}) = p_{11}^{25} p_{12}^{5} p_{21}^{6} p_{22}^{12} (p_{11}p_{12} + p_{12}p_{21} + p_{12}p_{22}) = p_{11}^{25} p_{12}^{6} p_{21}^{6} p_{22}^{12} (p_{11} + 1)$$

so that $p_{22}|\mathbf{x} \sim \text{Be}(12.5, 6.5)$ as earlier and p_{11} has a mixture posterior distribution

$$p_{11}|\mathbf{x} \sim 0.44\ \text{Be}(26.5, 6.5) + 0.56\ \text{Be}(25.5, 6.5).$$

The posterior mean is

$$E[\boldsymbol{P}|\mathbf{x}] = \begin{pmatrix} 0.800 & 0.200 \\ 0.342 & 0.658 \end{pmatrix}.$$

The predictive equilibrium probability is $E[\pi_1|\mathbf{x}] = 0.627$. △

One disadvantage of such approaches is that with large amounts of missing data, the Gibbs algorithms are likely to converge slowly as they will depend on the reconstruction of large quantities of latent variables. Further ideas on data reconstruction for Markov chains are indicated in Section 3.4.5.

3.4 Special topics

3.4.1 Reversible Markov chains

Assume that we have a reversible Markov chain with unknown transition matrix $\boldsymbol{P}$ and equilibrium distribution π satisfying the conditions of Definition 3.1. Then, for the standard experiment of observing a sequence of observations, $x_0, \ldots, x_n$, from the chain, where the initial state x_0 is assumed known, a conjugate prior distribution can be derived as follows.

First, the chain is represented as a graph, G, with vertices V and edges E, so that two vertices i and j are connected by an edge, $e = \{i, j\}$, if and only if $p_{ij} > 0$ and the edges $e \in E$ are weighted so that, for $e = \{i, j\}$, $w_e \propto \pi(i)p_{ij} = \pi(j)p_{ji}$ and $\sum_{e\in E} w_e = 1$. Note that if $p_{ii} > 0$, then there is a corresponding edge, $e = \{i, i\}$ called a loop. The set of loops shall be denoted by E_{loop}.

A conjugate probability distribution of a reversible Markov chain can now be defined as a distribution over the weights, $\mathbf{w}$ as follows. For an edge $e \in E$, let $\bar{e}$ represent the endpoints of e; for a vertex $v \in V$, set $w_v = \sum_{e:v\in\bar{e}} w_e$. Also, define $\mathcal{T}$ to be the set of spanning trees of G, that is, the set of maximal subgraphs that contains all loops in G, but no cycles. For a spanning tree, $T \in \mathcal{T}$, let $E(T)$ represent the edge set of T. Then, a conjugate prior distribution for $\mathbf{w}$ is given by:

$$f(\mathbf{w}|v_0, \mathbf{a}) \propto \frac{\prod_{e\in E\setminus E_{\text{loop}}} w_e^{a_e - 1/2} \prod_{e\in E_{\text{loop}}} w_e^{a_e/2-1}}{w_{v_0}^{a_{v_0}/2} \prod_{v\in V\setminus v_0} w_v^{(a_v+1)/2}} \sqrt{\sum_{T\in\mathcal{T}} \prod_{e\notin E(T)} \frac{1}{w_e}},$$

where v_0 represents the node of the graph corresponding to the initial state, x_0, $\mathbf{a} = (a_e)_{e\in E}$ is a matrix of arbitrary, nonnegative constants and $a_v = \sum_{e:v\in\bar{e}} a_e$.

The posterior distribution is $f(\mathbf{w}|\mathbf{x}) = f(\mathbf{w}|v_n, \mathbf{a}')$, where $\mathbf{a}' = (a_e + k_e(\mathbf{x}))_{e\in E}$ and

$$k_e(\mathbf{x}) = \begin{cases} |\{i \in \{1, \ldots, n\} : \{x_{i-1}, x_i\} = e\}|, & \text{for } e \in E \setminus E_{\text{loop}} \\ 2|\{i \in \{1, \ldots, n\} : \{x_{i-1}, x_i\} = e\}|, & \text{for } e \in E_{\text{loop}}, \end{cases}$$

where $|\cdot|$ represents the cardinality of a set. Therefore, for an edge e which is not a loop, $k_e(\mathbf{x})$ represents the number of traversals of e by the path $\mathbf{x} = (x_0, x_1, \ldots, x_n)$ and for a loop, $k_e(\mathbf{x})$ is twice the number of traversals of e.

The integrating constant and moments of the distribution are known and it is straightforward to simulate from the posterior distribution; for more details, see Diaconis and Rolles (2006).

3.4.2 Higher order chains and mixtures of Markov chains

Bayesian inference for the full rth order Markov chain model can, in principle, be carried out in exactly the same way as inference for the first-order model, by expanding the number of states appropriately, as outlined in Section 3.2.2.

Example 3.11: In the Australian rainfall example, Markov chains of orders $r = 2$ and 3 were considered. In each case, Be(1/2, 1/2) priors were used for the first nonzero element of each row of the transition matrix and it was assumed that the initial r states were generated from the equilibrium distribution. Then, the predictive equilibrium probabilities of the different states under each model are as follows

		States							
r	2	(1, 1)	(1, 2)	(2, 1)	(2, 2)				
	π	0.5521	0.1198	0.1198	0.2084				
r	3	(111)	(112)	(121)	(122)	(211)	(212)	(221)	(222)
	π	0.4567	0.0964	0.0731	0.0550	0.0964	0.0317	0.0550	0.1357

The log marginal likelihoods are −30.7876 for the second-order model and −32.1915 for the third-order model, respectively, which suggest that the simple Markov chain model should be preferred. △

Bayesian inference for the MTD model of (3.1) is also straightforward. Assume first that the order r of the Markov chain mixture is known. Then, defining an indicator variable Z_n such that $P(Z_n = z|\mathbf{w}) = w_z$, observe that the mixture transition model can be represented as

$$P(X_n = x_n|X_{n-1} = x_{n-1}, \ldots, X_{n-r} = x_{n-r}, Z_n = z, \boldsymbol{P}) = p_{x_{n-z}x_n}.$$

Then, a posteriori,

$$P(Z_n = z|X_n = x_n, \ldots, X_{n-r} = x_{n-r}, Z_n = z, \boldsymbol{P}) = \frac{w_z p_{x_{n-z}x_n}}{\sum_{j=1}^{r} w_j p_{x_{n-j}x_n}}. \quad (3.6)$$

Now, define the usual matrix beta prior for $\boldsymbol{P}$, a Dirichlet prior for $\mathbf{w}$, say $\mathbf{w} \sim \text{Dir}(\beta_1, \ldots, \beta_r)$, and a probability model $P(x_0, \ldots, x_{r-1})$ for the initial states of the chain. Then given a sequence of data, $\mathbf{x} = (x_0, \ldots, x_n)$, if the indicator variables are $\mathbf{z} = (z_r, \ldots, z_n)$ then

$$f(\boldsymbol{P}|\mathbf{x}, \mathbf{z}, \mathbf{w}) = \prod_{t=r}^{n} p_{x_{t-z_t} x_t} f(\boldsymbol{P})$$
$$f(\mathbf{w}|\mathbf{z}) = \prod_{t=r}^{n} w_{z_t} f(\mathbf{w}), \tag{3.7}$$

which are matrix beta and Dirichlet distributions, respectively. Therefore, a simple Gibbs sampling algorithm can be set up to sample the posterior distribution of $\mathbf{w}$, $\boldsymbol{P}$ by successively sampling from (3.6), (3.7), and (3.7).

When the order of the chain is unknown, two approaches might be considered. First, models of different orders could be fitted and then Bayes factors could be used for model selection as in Section 3.3.5. Otherwise a prior distribution can be defined over the different orders and then a variable dimension MCMC algorithm such as reversible jump (Green, 1995, Richardson and Green 1997) could be used to evaluate the posterior distribution, as in the following example.

Example 3.12: For the Australian rainfall data, consider mixture transition models of orders up to five. In order to simplify calculations, assume that the first five data are known throughout. Setting a discrete uniform, $r \sim \text{DU}[1, 5]$, prior distribution on the order and Dirichlet prior distributions $\mathbf{w}|r \sim \text{Dir}\left(\underbrace{\frac{1}{2}, \ldots, \frac{1}{2}}_{r}\right)$, then the estimated posterior distribution of r based on 200 000 reversible jump MCMC iterations is given in Figure 3.2.

The most likely model is the simple Markov chain model, confirming the results of Example 3.11. △

3.4.3 AR processes and other continuous state space processes

Assume that we wish to undertake inference for an AR(k) process as in Example 3.2. Given a sample of n data, $X_{k+1} = x_{k+1}, \ldots, X_{k+n} = x_{k+n}$ and known initial values, say $X_1 = x_1, \ldots, X_k = x_k$, and assuming the following prior structure:

$$\frac{1}{\sigma^2} \sim \text{Ga}\left(\frac{a}{2}, \frac{b}{2}\right)$$
$$\boldsymbol{\beta} \sim N(\mathbf{m}, \mathbf{V}).$$

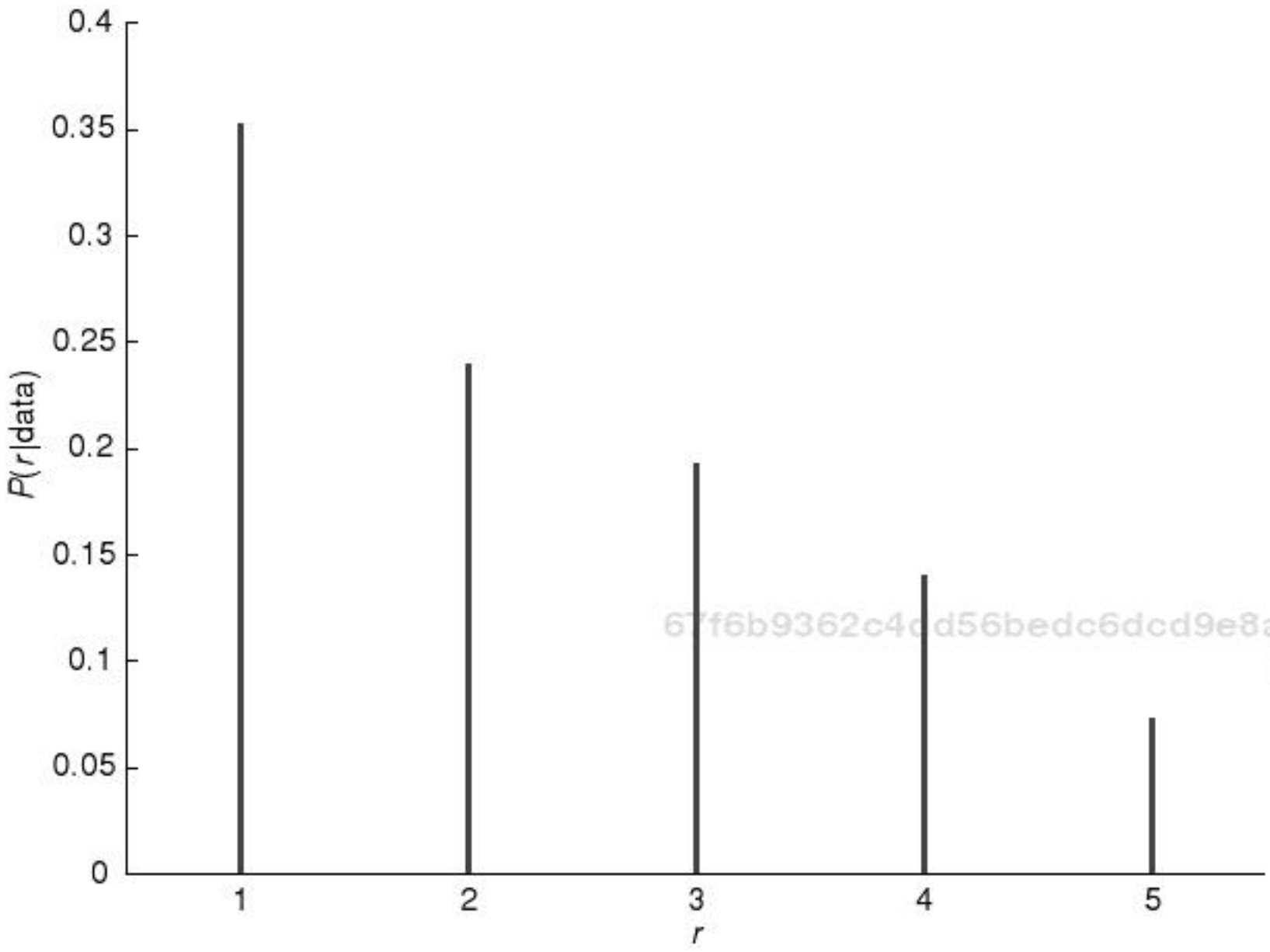

Figure 3.2 Posterior distribution of the number of terms in the mixture transition distribution model.

it is straightforward to calculate the full conditional posterior distributions as follows

$$\boldsymbol{\phi}|\sigma^2, \mathbf{x} \sim \mathrm{N}\left(\left(\mathbf{V}^{-1} + \frac{1}{\sigma^2}\mathbf{Z}\mathbf{Z}^T\right)^{-1}\left(\mathbf{V}^{-1}\mathbf{m} + \frac{1}{\sigma^2}\mathbf{Z}^T\mathbf{x}\right), \left(\mathbf{V}^{-1} + \frac{1}{\sigma^2}\mathbf{Z}\mathbf{Z}^T\right)^{-1}\right)$$

$$\frac{1}{\sigma^2}\bigg|\mathbf{x} \sim \mathrm{Ga}\left(\frac{a+n}{2}, \frac{b+(\mathbf{x}-\mathbf{Z}\boldsymbol{\phi})^T(\mathbf{x}-\mathbf{Z}\boldsymbol{\phi})}{2}\right),$$

where $\mathbf{x} = (x_{k+1}, \ldots, x_{k+n})^T$, $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T$ and $\mathbf{z}_t = (1, x_{t+k-1}, \ldots, x_t)^T$. Therefore, given suitable starting values, a simple Gibbs sampler can be implemented to iterate through these conditional distributions and approximate a sample from the posterior parameter distribution as in Section 2.4.1.

Note that it is straightforward to extend this model to incorporate the assumption of stationarity. Thus, if a Monte Carlo sample is generated from the posterior distribution of $\boldsymbol{\beta}$, σ^2, then by rejecting those sampled values with unit roots, then the sample can be reduced to a sample from the posterior distribution based on a normal gamma prior distribution truncated onto the region where the parameters satisfy the stationarity condition.

Second, the problem of model selection can be assessed either by defining a prior on k and using, for example, a reversible jump procedure (Green, 1995) to sample the posterior distribution, or by using Bayes factors or an information criterion such as DIC, as defined in Section 2.2.2, to select an appropriate value of k.

Example 3.13: Quarterly data on seasonally adjusted gross national product of the United States between 1947 and 1991 are analyzed in Tsay (2005). Using classical statistical methods, Tsay fits these data using an AR(3) model. Here, we consider AR models with 0 up to 4 lags and use the DIC to choose the appropriate model. We assume that the first four data are known and set independent prior distributions $\frac{1}{\sigma^2} \sim$ Ga(0.001, 0.001) and $\beta_i \sim$ N (0,0.0001) for $i = 0, \ldots, k$. The package WinBUGS (see, e.g., Lunn *et al.*, 2000) was used to run the Gibbs sampler with 100 000 iterations to burn in and 100 000 iterations in equilibrium in each case. Table 3.1 gives the values of the DIC for each model.

Table 3.1 DIC values for AR models with different lags.

Lags	DIC
0	−1065.2
1	−1090.3
2	−1099.1
3	**−1102.7**
4	−1092.3

The model suggested by deviance information criterion (DIC) is the AR(3) model. Furthermore, the model fitted in Tsay (2005) was

$$X_n = 0.0047 + 0.35X_{n-1} + 0.18X_{n-2} - 0.14X_{n-3} + \epsilon_n$$

where the standard deviation of the error term was estimated by $\hat{\sigma} = 0.0098$. In our case, the posterior mean predictor was

$$0.0047 + 0.3516X_{n-1} + 0.1798X_{n-2} - 0.1445X_{n-3},$$

and the posterior mean of σ was 0.0100. Finally, Figure 3.3 shows the fitted, in sample mean estimates and 95% predictive intervals using the AR(3) model for the series.

The AR(3) model appears to fit the series reasonably well. △

For more general models, inference may be somewhat more complicated and, typically, numerically intensive approaches will have to be used. Thus, if we assume that X_n is generated according to a Markov process, $X_n|X_{n-1}, \boldsymbol{\theta} \sim f(\cdot|X_{n-1}, \boldsymbol{\theta})$, given a sample of size n and a known initial value X_0 for the chain, a likelihood function can be constructed using the Markov property as $l(\boldsymbol{\theta}|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|x_{i-1}, \boldsymbol{\theta})$ and the posterior distribution must be estimated numerically.

A number of approaches are available. In some cases, MCMC methods may be employed. Another possibility is to approximate by assuming that $\boldsymbol{\theta}$ is time varying so that $\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + \epsilon_n$ where ϵ_n has a suitably small variance and this new model is a state space model which can be well fitted using filtering techniques as outlined

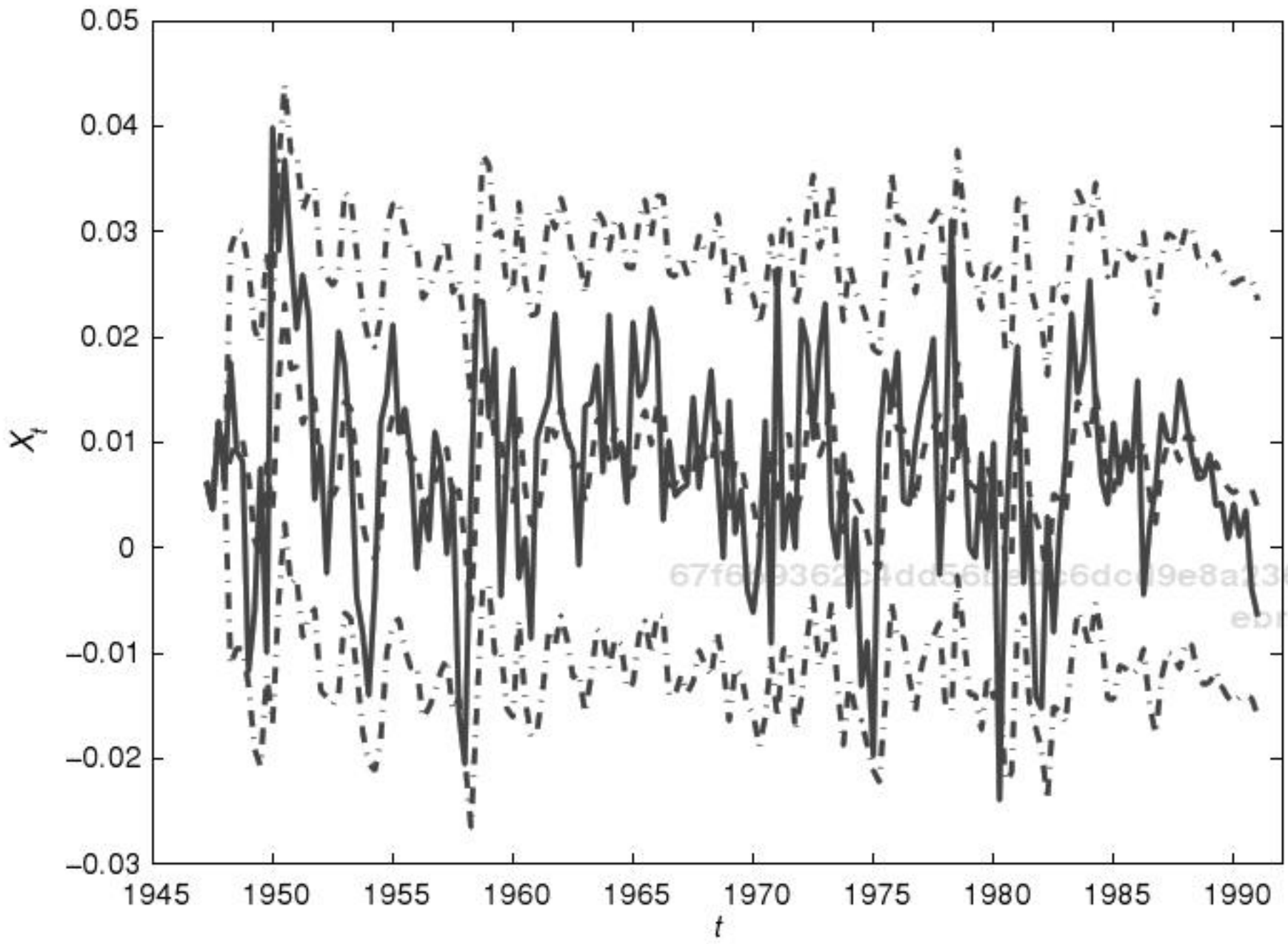

Figure 3.3 Gross National Product time series (solid line) with predictive mean (dashed line) and 95% interval (dot dash line).

in Section 2.4.1. Otherwise, Gaussian approximations or variational Bayes methods might be used (see, e.g., Roberts and Penny, 2002).

3.4.4 Branching processes

In inference for branching processes, the parameters of interest will usually be the mean of the offspring distribution, which determines whether or not extinction is certain, and the probability of eventual extinction.

Typically, the number of offspring born to each individual will not be observed. Instead, we will simply observe the size of the population at each generation. Suppose that given a fixed, initial population z_0, we observe the population sizes of n generations of a branching process, say $Z_1 = z_1, \ldots, Z_n = z_n$. Then, for certain parametric distributions of the number of offspring, Bayesian inference is straightforward.

Example 3.14: Assume that the number of offspring born to an individual has a geometric distribution

$$P(X = x|p) = p(1-p)^x, \quad \text{for } x = 0, 1, 2, \ldots$$

with $E[X|p] = \frac{1-p}{p}$. Now, for $p \geq 0.5$, $E[X|p] < 1$ and extinction is certain. Otherwise, from (3.2), the probability of extinction can be shown to be

$$\pi = \frac{p}{1-p}.$$

Given a beta prior distribution $p \sim \text{Be}(\alpha, \beta)$ and the sample data, and recalling that the sum of geometric distributed random variables has a negative binomial distribution, it is easy to see that

$$p|\mathbf{z} \sim \text{Be}(\alpha + z - z_n, \beta + z - z_0),$$

where $z = \sum_{i=0}^{n} z_i$. Therefore, the predictive mean of the offspring distribution is

$$E[X|\mathbf{z}] = E\left[\left.\frac{1-p}{p}\right|\mathbf{z}\right] = \frac{\beta + z - z_0}{\alpha + z - z_n - 1}.$$

The predictive probability that the population dies out in the next generation is

$$P(Z_{n+1} = 0|\mathbf{z}) = E\left[\left. p^{z_n}\right|\mathbf{z}\right] = \frac{B(\alpha + z, \beta + z - z_0)}{B(\alpha + z - z_n, \beta + z - z_0)}$$

and the predictive probability of eventual extinction is

$$\begin{aligned} E[\pi|\mathbf{z}] &= P(p > 0.5|\mathbf{z}) + \int_0^{0.5} \left(\frac{p}{1-p}\right)^{z_n} f(p|\mathbf{z})\,\mathrm{d}z \\ &= IB(0.5, \beta + z - z_0, \alpha + z - z_n) + \\ &\quad \frac{B(\alpha + z, \beta + z - z_0 - z_n)}{B(\alpha, \beta)} IB(0.5, \alpha + z, \beta + z - z_0 - z_n), \end{aligned}$$

where $IB(x, a, b) = \int_0^x \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}\,\mathrm{d}x$ is the incomplete beta function. △

Conjugate inference is also straightforward when, for example, binomial, negative binomial or Poisson distributions are assumed (see, e.g., Guttorp, 1991).

Example 3.15: The family trees of *Harry Potter* and other key characters in the famous series of books by J.K. Rowling is provided in

http://en.wikipedia.org/wiki/Harry_Potter_(character)

The number of male offspring born with the surname *Weasley* starting from a single ancestor (*Septimus Weasley*) in the 0th generation are as follows: $z_0 = 1, z_1 = 1, z_2 = 6, z_3 = 3$.

Two different parametric models for the offspring distribution were examined. First, a Poisson distribution and, second, a geometric distribution as described in Example 3.14. Assume a Poisson model with mean λ and setting an exponential prior $\lambda \sim \text{Ex}(\log 2)$, defined so that, a priori, the predictive probability that extinction is certain is equal to $1/2$. Then, the predictive mean of the offspring distribution is 1.265, the probability of extinction in a single generation is 0.038 and the probability of eventual extinction is 0.441. The log marginal likelihood for this model was -12.05.

Assume the geometric model with parameter p, and given a beta prior distribution, $p \sim \text{Be}(1/2, 1/2)$, which implies that the prior predictive probability that eventual extinction is certain is also $1/2$. Then, the posterior distribution of p is $p|\mathbf{z} \sim \text{Be}(8.5, 10.5)$ and the predictive mean of the offspring distribution is $E\left[\frac{1-p}{p}\middle|\mathbf{z}\right] \approx 1.4$. The probability that the *Weasleys* die out in a single further generation is $E\left[p^3\middle|\mathbf{z}\right] = 0.106$ and the probability that they eventually become extinct is $E[\pi|\mathbf{z}] = 0.562$. The marginal likelihood for the geometric model was -10.02.

Comparing both log marginal likelihoods gives a difference of, approximately, two, which conveys positive evidence in favor of the geometric model. $\triangle$

It is more interesting to consider the case where the offspring distribution is unknown but where the maximum number of offspring per individual is finite, say $K < \infty$. Given a Dirichlet prior distribution for the offspring distribution

$$\mathbf{p} = (p_0, p_1, \ldots, p_K) \sim \text{Dir}(\alpha_0, \alpha_1, \ldots, \alpha_K),$$

conjugate inference is impossible, but it is possible to use a normal approximation to estimate the offspring mean through

$$E[X|\mathbf{z}] \approx \frac{\alpha m + z - z_0}{\alpha + z - z_n} = \frac{\alpha}{\alpha + z - z_n} m + \frac{z - z_n}{\alpha + z - z_n} \hat{\mu},$$

where $\alpha = \sum_{k=1}^{K} \alpha_k$, $m = \frac{1}{K} \sum_{k=1}^{K} k\alpha_k$ is the prior mean and $\hat{\mu}$ is the maximum likelihood estimate of $\mu = E[X]$. In the case in which little prior information is available, a noninformative prior distribution is proposed by Mendoza and Gutiérrez Peña (2000), where it is shown that approximate posterior inference for μ can be undertaken.

Example 3.16: In the previous example, suppose that it is known that at maximum, a Weasley can have up to eight male offspring and that $P(X = x|\mathbf{p}) = p_x$ is the offspring distribution where $x = 0, 1, \ldots, 10$. Assume that a Dirichlet prior distribution is set for $\mathbf{p}$, that is $\mathbf{p} \sim \text{Dir}(\boldsymbol{\alpha})$ where $\alpha_0 = \alpha_1 = 0.5$ and $\alpha_i = 0.5/i$ for $i = 2, \ldots, 8$. This prior is set so that, a priori, $E[\mu] = 1$.

Using the normal approximation, the predictive posterior mean of the offspring distribution is 1.628. However, in this case, the posterior distribution can be explicitly

calculated by enumerating the possible numbers of births and deaths born to each individual in each generation. Thus,

$$\mathbf{p}|\mathbf{z} = \sum_{i=1}^{3} w_i \mathrm{Dir}(\boldsymbol{\alpha}_i'),$$

which is a mixture of three Dirichlets with weights $\mathbf{w} = (0.716, 0.188, 0.096)$ and parameters $\alpha_{ij}' = \alpha_j$ for 1, 2, 3 and $j = 0, \ldots, 8$, except $\alpha_{i6}' = \alpha_6 + 1$, for $i = 1, 2, 3$ and $\alpha_{10}' = \alpha_0 + 3$, $\alpha_{11}' = \alpha_1 + 4$, $\alpha_{20}' = \alpha_0 + 4$, $\alpha_{21}' = \alpha_1 + 2$, $\alpha_{22}' = \alpha_2 + 1$, $\alpha_{30}' = \alpha_0 + 5$, $\alpha_{31}' = \alpha_1 + 1$ and $\alpha_{33}' = \alpha_3 + 1$. Then, the exact posterior mean of the offspring distribution is 1.628, the predictive probability that the population dies out in a single generation is 0.33 and the probability that it eventually dies out is 0.659.

In this case, the log marginal likelihood of the model is -9.88 and, therefore, there is no real evidence to prefer this model over the geometric model of Example 3.15. Note also that there is a large amount of sensitivity to the choice of prior distribution. Small changes can produce relatively large changes both in the predictions and in the log likelihood. △

3.4.5 Hidden Markov models

Consider the HMM outlined in Section 3.2.5. Given the sample data, $\mathbf{y} = (y_0, \ldots, y_n)$, the likelihood function is

$$l(\boldsymbol{\theta}, \boldsymbol{P}|\mathbf{y}) = \sum_{x_0, \ldots, x_n} \pi(x_0) f(y_0|\boldsymbol{\theta}_{x_0}) \prod_{j=1}^{n} p_{x_{j-1}x_j} f(y_j|\boldsymbol{\theta}_{x_j}),$$

which contains K^{n+1} terms. In practice, this will usually be impossible to compute directly. A number of approaches can be taken in order to simplify the problem.

Suppose that the states, $\mathbf{x}$, of the hidden Markov chain were known. Then, the likelihood simplifies to

$$\begin{aligned} l(\boldsymbol{\theta}, \boldsymbol{P}|\mathbf{x}, \mathbf{y}) &= \pi(x_0) \prod_{j=1}^{n} p_{x_{j-1}x_j} \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}_{x_i}) \\ &= l_1(\boldsymbol{P}|\mathbf{x}) l_2(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}), \end{aligned}$$

where $l_1(\boldsymbol{P}|\mathbf{x}) = \pi(x_0|\boldsymbol{P}) \prod_{j=1}^{n} p_{x_{j-1}x_j}$ and $l_2(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}) = \prod_{i=0}^{n} f(y_i|\boldsymbol{\theta}_{x_i})$. Given the usual matrix beta prior distribution for $\boldsymbol{P}$, then a simple rejection algorithm could be used to sample from $f(\boldsymbol{P}|\mathbf{x})$ as in Section 3.3.6. Similarly, when $Y|\boldsymbol{\theta}$ is a standard exponential family distribution, then a conjugate prior for $\boldsymbol{\theta}$ will usually be available and, therefore, drawing a sample from each $\boldsymbol{\theta}_i|\mathbf{x}, \mathbf{y}$ will also be straightforward.

Also, letting $\mathbf{x}_{-t} = (x_0, x_1, \ldots, x_{t-1}, x_{t+1}, \ldots, x_n)$, it is straightforward to see that, for $i = 1, \ldots, K$,

$$\begin{aligned} P(x_0 = i|\mathbf{x}_{-0}, \mathbf{y}) &\propto \pi(i) p_{i x_1} f(y_1|\boldsymbol{\theta}_i) \\ P(x_t = i|\mathbf{x}_{-t}, \mathbf{y}) &\propto p_{x_{t-1} i} p_{i x_{t+1}} f(y_t|\theta_i) \quad \text{for } 1 < t < n \\ P(x_n = i|\mathbf{x}_{-n}, \mathbf{y}) &\propto p_{x_{n-1} i} f(y_n|\boldsymbol{\theta}_i) \end{aligned}$$

so that a full Gibbs sampling algorithm can be set up.

A disadvantage of this type of algorithm is that the generated sequences $\mathbf{x}^{(s)}$ can be highly autocorrelated, particularly when there is high dependence amongst the elements of $\mathbf{x}$ in their posterior distribution. In many cases, it is more efficient to sample directly from $P(\mathbf{x}|\mathbf{y})$. The standard approach for doing this is to use the forward-backward or Baum–Welch formulas (Baum *et al.*, 1970).

First, note that $P(x_n|x_{n-1}, \mathbf{y}) = P(x_n|x_{n-1}, y_n) \propto p_{x_{n-1} x_n} f(y_n|x_n) \equiv P'_n(x_n|x_{n-1}, \mathbf{y})$ which is the unnormalized conditional density of $x_n|x_{n-1}, \mathbf{y}$. Also, it is easy to show that we have a backward recurrence relation

$$P(x_t|x_{t-1}, \mathbf{y}) \propto p_{x_{t-1} x_t} f(y_t|x_t) \sum_{i=1}^{K} p'_{t+1}(i|x_t, \mathbf{y}) \equiv P'_t(x_t|x_{t-1}, \mathbf{y})$$

and, finally,

$$P(x_0|\mathbf{y}) \propto \pi(x_0) f(y_0|x_0) \sum_{i=1}^{K} P'_1(i|x_0, \mathbf{y}) \equiv P'_0(x_0|\mathbf{y}).$$

Given this system of equations, it is now possible to simulate a sample from $P(\mathbf{x}|\mathbf{y})$ by using forward simulation, so that x_0 is simulated from $P(x_0|\mathbf{y})$ and then, x_1 is simulated from $P(x_1|x_0, \mathbf{y})$, and so on.

In many cases, the order of the hidden Markov chain will be unknown. Two options are available. First, the hidden Markov model may be run over various different chain dimensions and the optimal dimension may be selected using, for example, Bayes factors. Alternatively, a prior distribution for the dimension of the HMM could be defined and a transdimensional MCMC algorithm such as reversible jump could be used to mix over chains of different sizes.

Finally, for HMMs with continuous state space, we should first note that such models can be expressed in state space form as

$$\begin{aligned} Y_n|X_n &\sim g(\cdot|X_n) \\ X_n|X_{n-1} &\sim f(\cdot|X_{n-1}) \end{aligned}$$

for functions f and g. Then, inference can proceed via the use of particle filters such as the sequential importance resampling algorithm described in Section 2.4.1.

3.4.6 Markov chains with covariate information and nonhomogeneous Markov chains

Markov chains are a common model for discrete, longitudinal study data, where the state of each subject changes over time according to a Markov chain. Usually, covariate information is available for each subject and two situations have been considered. First, when the evolution of the subjects can be modeled hierarchically by a set of related, homogeneous Markov chains and, second, when the parameters of the chains are allowed to vary (slowly) over time.

Consider the first case. Suppose that we have M subjects and let $\mathbf{x}_m = (x_{m,0}, \ldots, x_{m,n_m})$, where $x_{mj} \in \{1, \ldots, K\}$ is the sequence of observed states for subject m and the initial state is assumed known. Assume that covariate information $\mathbf{c}_m$ is available for individual m. Then, the transition probabilities $p_{mij} = P(X_{m,t+1} = j | X_{m,t} = i)$ are assumed to follow a polytomous regression model

$$\log \frac{p_{mij}}{p_{miK}} = \mathbf{c}_m \boldsymbol{\theta}_{ij} \tag{3.8}$$

so that $p_{mij} \propto \exp(\mathbf{c}_i \boldsymbol{\theta}_{ij})$, where $\boldsymbol{\theta}_{ij}$ are unknown regression parameters. The regression parameters may be modeled with, standard, hierarchical, normal-Wishart prior distributions. Given the observed data, the logistic regression structure implies that the relevant conditional posterior distributions are log concave so that standard Gibbs sampling techniques could be used to sample the posterior distributions. Often, the complete paths will not be observed for all subjects. In such cases, the basic algorithm can be modified by conditioning on the missing data. Given the complete data, the Gibbs sampler for $\boldsymbol{\theta}$ proceeds as above. Given $\boldsymbol{\theta}$, the transition matrices for each subject are known so that the missing data can be sampled as in Section 3.3.7.

In the second case, (3.8) can be extended so that

$$\log \frac{p_{mijn}}{p_{miKn}} = \mathbf{c}_m \boldsymbol{\theta}_{ijn},$$

where $\boldsymbol{\theta}_{in}$ develops over time according to a state space model, for example,

$$\boldsymbol{\theta}_{in} = \boldsymbol{\theta}_{i(n-1)} + \boldsymbol{\epsilon}_n$$

and ϵ_t is a noise term. Again, using the standard normal Wishart model, inference follows easily.

3.5 Case study: Wind directions at Gijón

Since November 1994, wind directions have been recorded daily at the Davis automatic weather observatory in Somió situated 4 km from the city of Gijón on the North coast of Spain. Figure 3.4 shows the average, daily wind direction recorded over 1000 days starting from November 6, 1994. The data have been discretized and classified into sixteen cardinal directions $N, NNE, NE, \ldots, NW, NNW$. There is

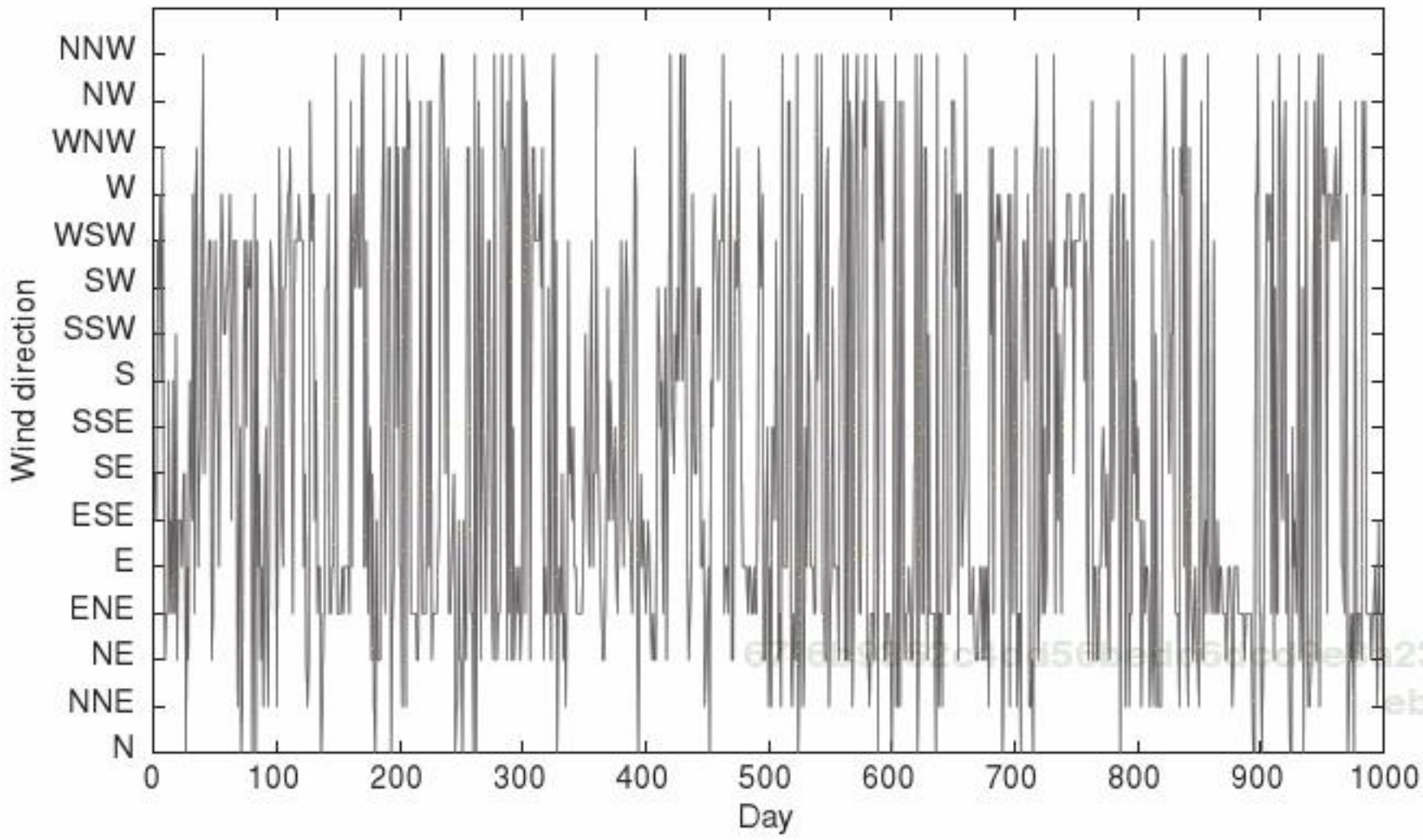

Figure 3.4 Time series plot of wind directions at Somió.

a very small number of days when observations have not been recorded which have not been taken into account for the purposes of this analysis. The original data may be obtained from http://infomet.am.ub.es/clima/gijon/.

Figure 3.5 shows a rose plot of this data. The most typical wind direction is northeasterly. Given the directional nature of the data, special techniques are necessary

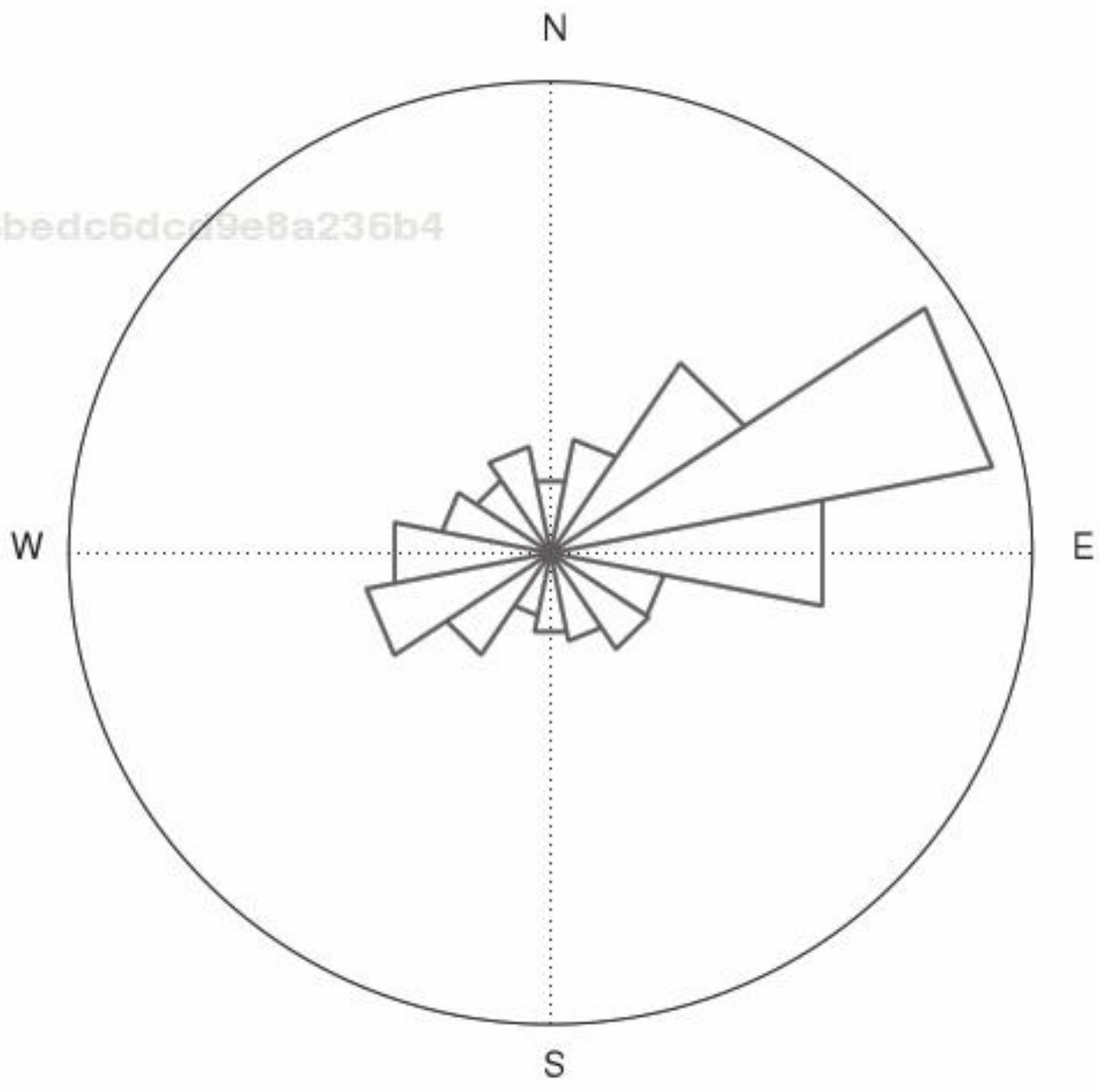

Figure 3.5 Rose plot of wind directions at Somió.

for its analysis. In particular, the mean of the data is approximately 165°, which is not a good measure of the average wind direction. Instead, we shall use the circular mean of a sample of directional data, $\theta_1, \ldots, \theta_n$, coded in radians, so that $0 \leq \theta_j < 2\pi$ for all j. This is defined by

$$\bar{\theta} = \arctan\left(\sum_{j=1}^{n} \sin\theta_i, \sum_{i=1}^{n} \cos\theta_j\right) = \arg\left(\frac{1}{n}\sum_{j=1}^{n} \exp(i\theta_j)\right),$$

where $i = \sqrt{-1}$. In this case, converting to degrees, the circular mean wind direction is approximately 65° North.

3.5.1 Modeling the time series of wind directions

For these data, there is no particular evidence of any seasonal effects, as rose plots for different months of the year show very similar forms. This suggests that stationary time series models might be considered. We consider the following four possibilities:

1. An independent multinomial model, assuming that the wind direction, θ_n on day n is independent of the wind directions on previous days so that $P(\theta_n = i|\boldsymbol{\pi}) = \pi_i$ for $i = 0, \ldots, 15$, where 0 represents North, 1 NNE, 2 NE, ..., and 15 NNW.
2. A first-order Markov chain $P(\theta_n = j|\theta_{n-1} = i, \mathbf{P}) = p_{ij}$ for $i, j = 0, \ldots, 15$
3. A parametric, wrapped Poisson HMM.
4. A semiparametric, multinomial HMM.

The first two models have been described previously. The following subsections outline the wrapped Poisson HMM and the multinomial HMM.

The wrapped Poisson HMM and its inference

Before considering the wrapped Poisson HMM, we shall first consider inference for the rate parameter of a single wrapped Poisson distribution, $\theta|\lambda \sim \text{WP}(k, \lambda)$, as defined in Appendix A. In our context, we shall assume $k = 16$ to represent the 16 wind directions.

Note first that this model is equivalent to assuming that $Y|\lambda, Z = z \sim \text{Po}(\lambda)$ where $Y = \theta + kZ$ and $Z = z \in \mathbb{Z}^+$ is an unwrapping coefficient with probability

$$P(Z = z|\lambda) = \sum_{j=0}^{k-1} \frac{\lambda^{kz+j} e^{-\lambda}}{(kz+j)!} \quad z = 0, 1, 2, \ldots$$

This implies that

$$P(Z = z|\lambda, \theta) \propto \frac{\lambda^{\theta+kz}}{(\theta+kz)!}. \tag{3.9}$$

Therefore, given a sample $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_m)$ from this wrapped Poisson distribution and assuming a Ga(a, b) prior distribution for λ, inference can be carried out via Gibbs sampling. Conditional on $\lambda, \boldsymbol{\theta}$, the unwrapping coefficients, $z_1, \ldots, z_m$ can be generated from (3.9) and then, the unwrapped observations, $y_t = \theta_t + kz_t$ can be evaluated for $t = 1, \ldots, m$. Also, we have

$$\lambda|\boldsymbol{\theta}, \mathbf{y} \sim \text{Ga}(a + m\bar{y}, b + m).$$

A wrapped Poisson HMM with s hidden states is defined as follows. First, we suppose that

$$\theta_n|\boldsymbol{\lambda}, x_n \sim \text{WP}(16, \lambda_{x_n}),$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_s)$ and $\{X_t\}$ is an unobserved Markov chain with transition matrix $\boldsymbol{P}$, so that

$$P(X_t = x_t|x_{t-1}, x_{t-2}, \ldots, \boldsymbol{P}) = p_{x_{t-1}, x_t}$$

for $x_t, x_{t-1} = 1, 2, \ldots, s$.

Given a set of observations, $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_n)$, generated from the wrapped Poisson HMM, then inference can be carried out using the general procedure outlined in Section 3.4.5. Conditional on the hidden states, then the likelihood reduces to the product of a set of individual likelihoods for $\lambda_1, \ldots, \lambda_s$ and inference for each λ_i can be carried out by conditioning on the hidden states, $\mathbf{x}$, and the unwrapping coefficients, $\mathbf{z}$, when inference is straightforward as outlined above. The joint conditional posterior distribution of $\mathbf{x}, \mathbf{z}$ is then expressed as

$$f(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}, \boldsymbol{P}, \boldsymbol{\lambda}) = f(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{P}, \boldsymbol{\lambda}) f(\mathbf{z}|\boldsymbol{\theta}, \mathbf{x}, \boldsymbol{\lambda}).$$

The generation of $\mathbf{z}$ from $f(\mathbf{z}|\boldsymbol{\theta}, \mathbf{x}, \boldsymbol{\lambda})$ can then be carried out by applying (3.9) and, in order to generate $\mathbf{x}$, a forward backward algorithm can be used. Thus, as in Section 3.4.5, we define the backward equations

$$P(x_n|x_{n-1}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{P}) \propto p_{x_{n-1}x_n} f_{WP}(\theta_n|k, \lambda_{x_n}) \equiv P'(x_n|x_{n-1}, \boldsymbol{\theta})$$

$$P(x_t|x_{t-1}, \boldsymbol{\theta}, \boldsymbol{\lambda}) \propto p_{x_{t-1}x_t} f_{WP}(\theta_t|k, \lambda_{x_t}) \sum_{i=1}^{s} P'_{t+1}(i|x_t, \boldsymbol{\theta}) \equiv P'(x_t|x_{t-1}, \boldsymbol{\theta})$$

$$P(x_0|\boldsymbol{\theta}, \boldsymbol{\lambda}) \propto \pi_{x_0} f(\theta_0|k, \lambda_{x_0}) \sum_{i=1}^{s} P'_1(i|x_0, \boldsymbol{\theta}) \equiv P'(x_0|\boldsymbol{\theta}),$$

where $f_{WP}(\theta|k, \lambda)$ is a wrapped Poisson probability function. Then $\mathbf{x}$ is generated by sampling successively from $P(x_0|\boldsymbol{\lambda}, \boldsymbol{\theta})$ and $P(x_t|x_{t-1}, \boldsymbol{\lambda}, \boldsymbol{\theta})$ for $t = 1, 2, \ldots, n$.

The multinomial HMM and its inference

Assume that $P(\theta_t = \theta | x_t, \mathbf{Q}) = q_{x_t}$, where $\{X_t\}$ is an unobserved Markov chain with transition matrix $\boldsymbol{P}$, as earlier, and where $\mathbf{Q} = (q_{ij})$ for $i = 1, \ldots, r$, $j = 0, \ldots, 15$ such that $q_{ij} \geq 0$ for all i, j and $\sum_{j=0}^{15} q_{ij} = 1$.

Define independent Dirichlet priors for the rows of $\mathbf{Q}$, that is,

$$\mathbf{q}_i = (q_{i0}, \ldots, q_{i15}) \sim \text{Dir}\left(\underbrace{\frac{1}{2}, \ldots, \frac{1}{2}}_{16}\right).$$

Given the usual matrix beta prior for $\boldsymbol{P}$ and defining the latent variables X_t to represent the unobserved states of the Markov chain as earlier, we have

$$\begin{aligned} P(X_1 = x_1 | \boldsymbol{\theta}, \boldsymbol{P}, \mathbf{Q}, \mathbf{x}_{-1}) &\propto \pi(x_1 | \boldsymbol{P}) q_{x_1 \theta_1} p_{x_1 x_2} \\ P(X_t = x_t | \boldsymbol{\theta}, \boldsymbol{P}, \mathbf{Q}, \mathbf{x}_{-t}) &\propto p_{x_{t-1} x_t} q_{x_t \theta_t} p_{x_t x_{t+1}} \\ P(X_n = x_n | \boldsymbol{\theta}, \boldsymbol{P}, \mathbf{Q}, \mathbf{x}_{-n}) &\propto p_{x_{n-1} x_n} q_{x_n \theta_n}. \end{aligned}$$

Therefore, simple Gibbs steps can be used to generate a sequence of hidden states given $\mathbf{Q}$. Furthermore,

$$\mathbf{q}_i | \boldsymbol{\theta}, \mathbf{x}, \boldsymbol{P} \sim \text{D}\left(\frac{1}{2} + n_{i0}, \ldots, \frac{1}{2} + n_{i15}\right),$$

where $n_{ij} = \sum_{t=1}^{n} I(x_t = i, \theta_t = j)$ and $I(\cdot, \cdot)$ is an indicator variable. Then, it is straightforward to generate values of $\mathbf{Q}$ conditional on the hidden states.

3.5.2 Results

Relatively uninformative priors were used for all parameters of all four models. HMMs of various orders were considered and here we show the results of fitting a wrapped Poisson HMM with five hidden states to an origin shifted version of the data, $\theta^I = \text{mod}(\theta - 3, 16)$, and a multinomial HMM with four hidden states.

Figure 3.6 shows the marginal frequencies of each wind direction and the predictive marginal probabilities under the independent and Markov chain models and for the two HMMs. The results are similar under all four models although the wrapped Poisson model smooths the predictive distribution slightly more than the alternative models. Note also that, in all cases, the predictive mean wind direction was around 65° North, very close to the empirical mean wind direction.

In order to assess the time dependence of these data, the circular autocorrelation function (CACF) can be considered. A number of alternative definitions for a CACF

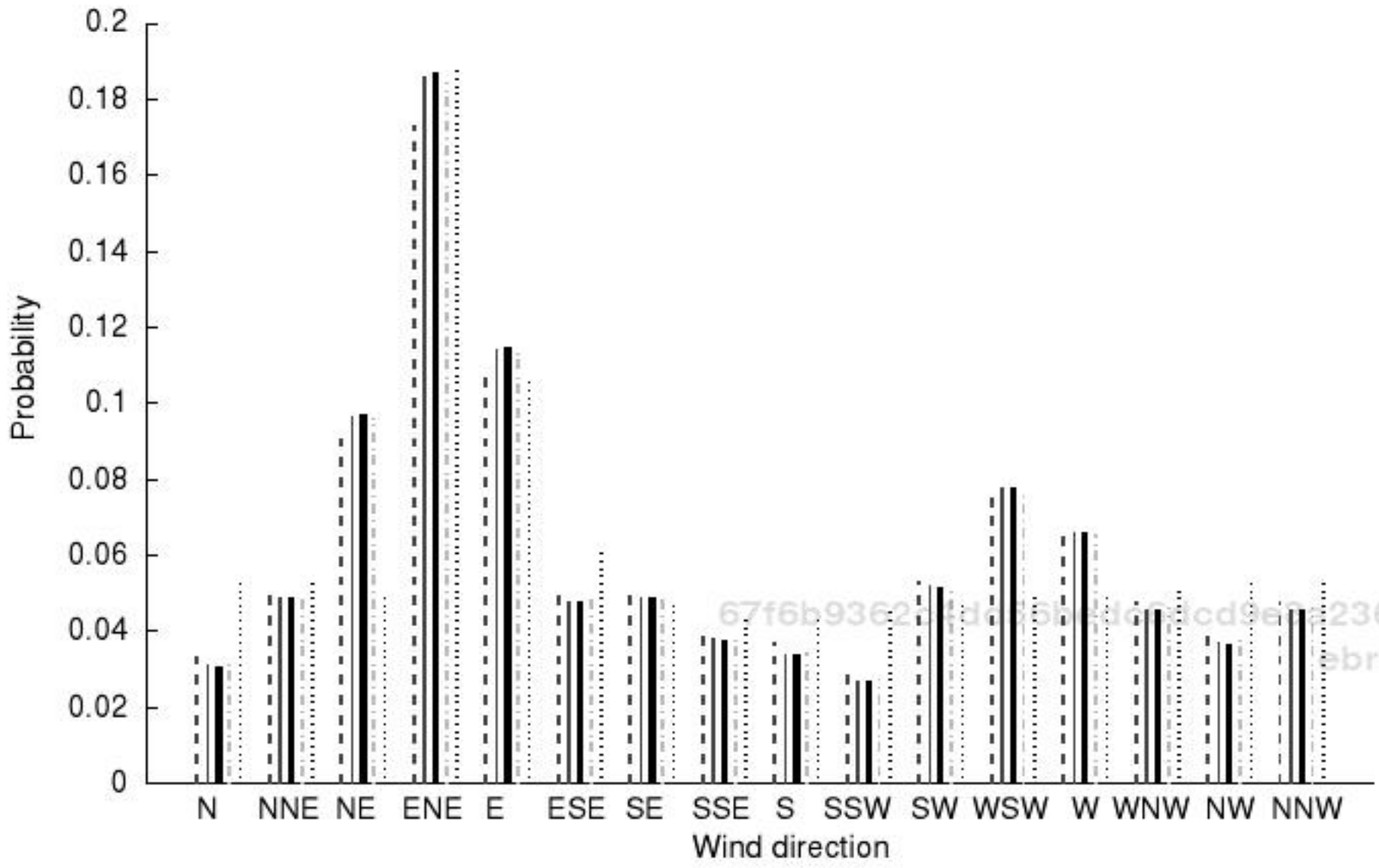

Figure 3.6 Marginal frequencies (solid thick line) and predictive probabilities under the independent (solid thin line), Markov chain (dashed line), multinomial (dot dash line), and wrapped Poisson (dotted line) HMMs.

have been proposed. Define the CACF of lag l for a sample $\theta_1, \ldots, \theta_n$ of data to be

$$\text{CACF}(l) = \frac{\sum_{t=1}^{n-l} \sin(\theta_t - \bar{\theta}) \sin(\theta_{t+l} - \bar{\theta})}{\sum_{t=1}^{n} \sin(\theta_t - \bar{\theta})^2}$$

and, similarly, for a variable $\{\theta\}_t$, then

$$\text{CACF}(l) = \frac{E\left[\sin(\theta_t - \mu) \sin(\theta_{t+l} - \mu)\right]}{E\left[\sin(\theta_t - \mu)^2\right]},$$

where μ represents the mean angular direction.

Figure 3.7 shows the empirical CACF and the predictive CACFs for all models except the independent model (where the CACFs are equal to zero). It can be observed that none of the proposed models estimate the empirical CACF very well. This is a feature that has been noted elsewhere when Markovian models and HMMs have been fitted to circular data (see, e.g., Holzmann, *et al*. 2007).

As might be expected, comparing the independent and Markovian models via Bayes factors showed strong evidence in favor of the Markovian model. Formal comparisons with the other models were not carried out here, but it can be noted that the wrapped Poisson HMM is less heavily parameterized than the multinomial HMM, which is also less parameterized than the simple Markov chain model which suggests that one of these two approaches should be preferred.

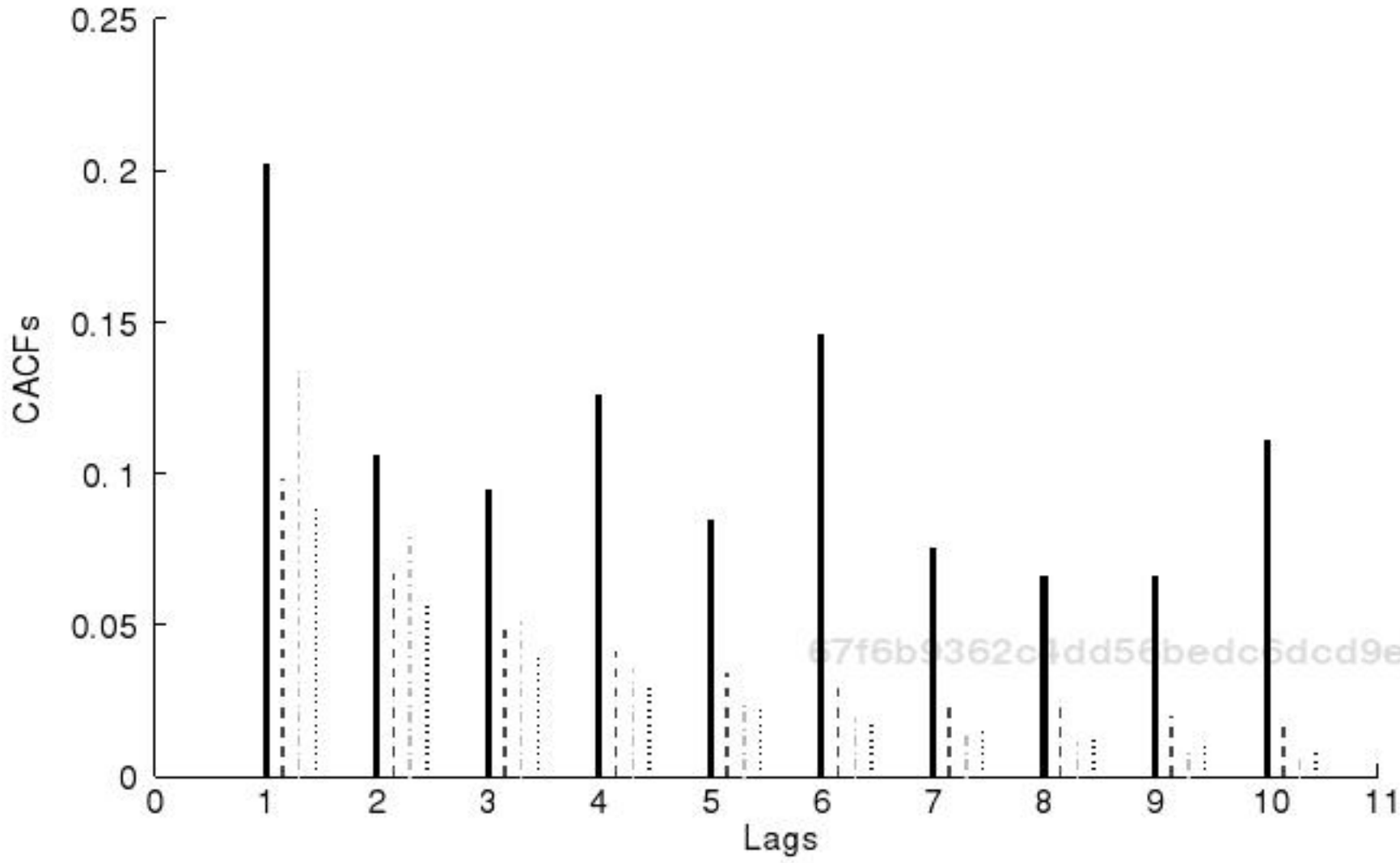

Figure 3.7 Empirical (solid thick line) and predictive CACFs under the Markov chain (dashed line), multinomial (dot dash line), and wrapped Poisson (dotted line) HMMs.

Finally, in order to compare the predictive capacities of the different models, one step ahead predictions were calculated for a period of 20 days ahead, with the predictive mean angular direction being used to predict the daily wind direction. The cumulative mean absolute predictive errors, $\frac{1}{t}\sum_{i=1}^{t}\epsilon_i$, were calculated for each model, where the error, $\epsilon = \epsilon(\hat{\theta}, \theta)$ of a prediction $\hat{\theta}$ of θ is calculated as

$$\epsilon = \min\left\{|\theta - \hat{\theta}|, 16 - |\theta - \hat{\theta}|\right\}.$$

These are plotted in Figure 3.8.

It can be seen that the independent model does somewhat worse than the Markovian and HMMs and that the best predictions over these 20 days in terms of this error function are given by the wrapped Poisson model.

3.6 Markov decision processes

Assume that a system can be in one of a finite number, K, of observable states, say $X \in \mathcal{X} = \{1, \ldots, K\}$ and that transitions between states occur at discrete stages, $n = 0, 1, 2, \ldots$. At each time step, a decision maker (DM) can select one of a finite set of actions, say $a_n \in \mathcal{A} = \{a_1, \ldots, a_m\}$. Then the transition probabilities, which describe the evolution of the system are given by $p_{ija} = P(X_{n+1} = j | X_n = i, a_n = a)$ for $i, j \in \mathcal{X}$ and $a \in \mathcal{A}$, and depend on both the current state of the chain and upon the action taken by the DM. At stage n, the DM receives a reward or utility $r_n = r(x_{n-1}, a_{n-1}, x_n)$. A Markov decision process (MDP) is defined by the

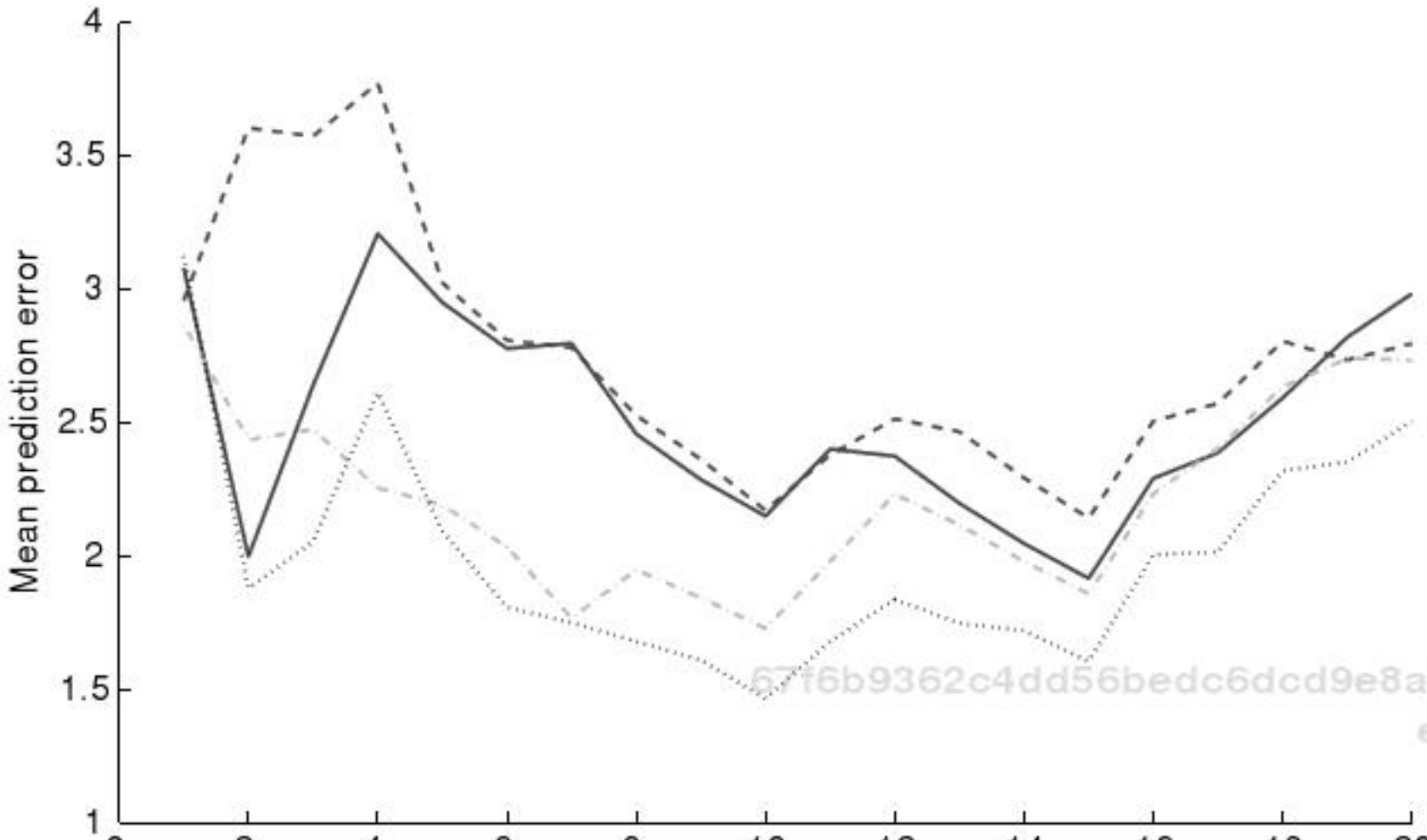

Figure 3.8 Cumulative mean prediction errors for the independent (solid line) and Markov chain (dashed line) models and the multinomial (dot dash line) and wrapped Poisson (dotted line) HMMs.

four-tuple $< \mathcal{X}, \mathcal{A}, \mathcal{T}, \mathcal{R} >$ where $\mathcal{T}$ is the set of transition probabilities and $\mathcal{R}$ is the set of rewards.

The behavior of the DM in an MDP can be modeled using the idea of a policy. A deterministic stationary policy $\pi : \mathcal{X} \rightarrow \mathcal{A}$ prescribes the action to be taken given the current state. A stochastic policy chooses the actions in a given state according to a probability distribution. The objective of the DM is to choose a policy π which maximizes the expected, discounted reward defined by

$$E\left[\sum_{n=0}^{\infty} \gamma^n r_n^{\pi}\right], \tag{3.10}$$

where $\gamma < 1$ is a discount factor and r_n^{π} is the expected reward received at time n under policy π.

If the value of performing action a in a state x is defined as

$$Q(x, a) = \sum_{x' \in \mathcal{X}} R(x, a) + \gamma V(x'),$$

where $R(x, a) = \sum_{x' \in \mathcal{X}} p_{xx'a} r(x, a, x')$ is the expected reward or utility achieved by taking action a in state x and $V(x)$ is the overall value of state x defined by Bellman's

equations

$$V(x) = \max_{a \in \mathcal{A}} \left\{ R(x,a) + \gamma \sum_{x' \in \mathcal{X}} p_{xx'a} V(x'), \right\} \tag{3.11}$$

then the optimal stationary policy π may be derived through

$$\pi(x) = \arg\max_{a \in \mathcal{A}} \left\{ R(x,a) + \gamma \sum_{x' \in \mathcal{X}} p_{xx'a} V(x') \right\}.$$

For known transition probabilities and utility functions, various algorithms to estimate the optimal decision policy are available (see, e.g., Puterman, 1994).

When rewards and transition probabilities are unknown, then the analysis is much more complex. Reinforcement learning (RL) comprises finding the optimal policy in such situations. Here, we shall only examine the case where the transition probabilities are unknown but the rewards are known. Then, given independent, matrix beta prior distributions for the transition matrices $\boldsymbol{P}_a = (p_{ija})$ clearly, given the observation of a sequence of states, the posterior distribution has the same form. In a similar way, it is often assumed that rewards are randomly distributed, typically according to a normal distribution, when inference for the reward parameters is straightforward given conjugate prior distributions. Most of the following analyses can be extended to these cases.

Assume that the DM's distribution for the transition matrices is parameterized by a set of parameters $\boldsymbol{\alpha}$. Then, Bellman's equations (3.11) can be modified as follows

$$V(x,\boldsymbol{\alpha}) = \max_{a} \left\{ R(x,a,\boldsymbol{\alpha}) + \gamma \sum_{x' \in \mathcal{X}} E[p_{xx'a}|\boldsymbol{\alpha}] V(x', \boldsymbol{\alpha}^{xx'a}) \right\},$$

where $R(x,a,\boldsymbol{\alpha}) = \sum_{x'} E[p_{xx'a}|\boldsymbol{\alpha}] r(x,a,x')$. $V(x,\boldsymbol{\alpha})$ is the expected value function given $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^{xx'a}$ represents the updated set of parameter values conditional on the transition from state x to x' when action a is taken.

Martin (1967) demonstrates that a set of solutions to this problem do exist. However, it is obvious that this problem has an infinite set of states $(x, \boldsymbol{\alpha})$ which makes its direct solution infeasible in practice. Many alternative approaches have been suggested.

An early approach to policy optimization was based on Thompson sampling, introduced in Thompson (1933). Given the current state x and model parameters $\boldsymbol{\alpha}$, a set of transition probabilities are drawn from $f(\cdot|\boldsymbol{\alpha})$ and the optimal policy π given these probabilities is calculated. Note that this is not Bayes optimal as it is a myopic strategy, and does not take into account the effects of actions on the DMs future belief states.

A second approach is based on sparse sampling (see Kearns, *et al*. 2002). Instead of considering an infinite horizon problem, a finite, effective horizon is assumed so

that rewards up to only n time steps in the future are considered. Then, if n is relatively small, it would, in theory, be possible to enumerate all possible future actions, states and rewards and then calculate the sequence of actions with the overall expected rewards. Of course, as n increases, the number of possible futures increases very rapidly, and therefore, approaches based on simulating futures up to the effective horizon can be considered. Thus, given the current state, x_0 say, for each possible decision a_0, states $x_1 = x_1(a_0)$ are simulated. Then, for each possible decision a_1, states $x_2 = x_2(a_1)$ are simulated and so on, up to the horizon. Thus, a sparse, lookahead tree is grown. The optimal policy is then estimated by maximizing over the expected rewards as for any other decision tree. A disadvantage of such approaches is the obviously high computational cost.

A third alternative is based on percentile optimization (see, e.g., Delage and Mannor, 2010). Under this framework, it is assumed that the DM wishes to select a policy π to maximize $y \in R$ subject to

$$P\left(E\left[\sum_{n=0}^{\infty} \gamma^n r_n^{\pi}\right] \geq y\right) \geq 1 - \epsilon$$

for some small ϵ. In the case where rewards are random and normally distributed and the transition matrices are known, then Delage and Manner (2010) show that an optimal solution to this problem exists and can be found in polynomial time. However, in the case of unknown transition matrices, they demonstrate that this problem is NP hard, although they provide some heuristic algorithms which can find approximate solutions.

Finally, a fourth approach uses policy gradients (see Williams, 1992). Here, it is assumed that the stationary policy defines a parameterized distribution, $P(\cdot|x, \theta)$ over actions conditioned on the current state. Then, a class of smoothly parameterized stochastic policies is defined and the gradient of the expected return (3.10) is evaluated with respect to the current policy parameters θ and the policy is improved by adjusting the parameters in the direction of the gradient.

3.7 Discussion

Bayesian inference for discrete time, finite Markov chains developed from the initial papers of Silver (1963) and Martin (1967). Other early works of interest are Lee and Judge (1968), Dubes and Donoghue (1970), and Bartholomew (1975). More recent studies are Assodou and Essebbar (2003) and Welton and Ades (2005). Empirical Bayes approaches have also been developed by, for example, Meshkani and Billard (1992) and Billard and Meshkani (1995). From a theoretical viewpoint, the de Finetti theorem for Markov chains was developed in Diaconis and Freedman (1980), where extensions and similar result for transient chains is also given. Further details and techniques for comparing multinomial, Markov and HMMs are given in Johansson and Olofsson (2007).

One point that we have not considered here is the consistency and convergence rate of the posterior distribution. For exchangeable data, there is a large amount of literature on this topic, but for Markov chains, there are fewer results. However, large and moderate deviation principles for the convergence of a sequence of Bayes posteriors have been established by Papangelou (1996) and Eichelsbacher and Ganesh (2002), which demonstrate the exponential convergence of Bayesian posterior distributions for Markov chains; see also Ganesh and O'Connell (2000) for further results.

Inference for reversible Markov chains is considered in Diaconis and Rolles (2006) and extensions to variable order chains are examined in Bacallado (2010). The Markov chain mixtures considered here were developed in Raftery (1985). Markov models with covariate information are examined in Deltour *et al*. (1999) and nonhomogeneous Markov chains are analyzed in Soyer and Sung (2007) and Hung (2000).

Bayesian inference for AR models via the Gibbs sampler is introduced in McCulloch and Tsay (1994). Extensions are developed in, for example, Barnett *et al*. (1996). Many other related models such as AR moving average models or vector AR models are also well analyzed in the Bayesian time series literature; see, for example, Prado and West (2010).

Useful references to Bayesian inference for hidden Markov chains are Gharamani (2001), Scott (2002), Cappé *et al*. (2005), McGrory and Titterington (2009). Particle filtering approaches for continuous state space chains are considered in, for example, Fearnhead and Clifford (2003) or Cappé *et al*. (2005).

Bayesian inference for branching processes using parametric models and normal approximations has been considered by, amongst others, Dion (1972, 1974), Heyde (1979), Scott (1987), and Guttorp (1991). Nonparametric approaches are examined by Guttorp (1991) and Mendoza and Guttiérez Peña (2000) and power series prior asymptotic results are analyzed in Scott and Heyde (1979). There has also been much literature extending the basic Galton–Watson process, for example, bisexual branching processes are analyzed in, for example, Molina *et al*. (1998). Finally there is a large literature on the related problem of phylogenetic inference, that is the study of the evolutionary tree of an organism (see, e.g., Huelsenbeck and Ronquist, 2001).

Finally, the theory of MDPs dates from the work of Howard (1960). Bayesian inference was considered by Silver (1963) and more recent approaches have been developed in, for example, Strens (2000), Kearns *et al*. (2002), Wang *et al*. (2005), Ghavamzada and Engel (2007), and Delage and Mannor (2010).

References

Assodou, S. and Essebbar, B. (2003) A Bayesian model for Markov chains via Jeffreys prior. *Communications in Statistics: Theory and Methods*, **32**, 2163–2184.

Bacallado, S. (2010) Bayesian analysis of variable-order, reversible Markov chains. *Annals of Statistics*, **39**, 838–864.

Barnett, G., Kohn, R., and Sheather, S. (1996) Robust Bayesian estimation of autoregressive moving average models *Journal of Time Series Analysis*, **18**, 11–28.

Bartholomew, D.J. (1975) Errors of prediction for Markov chain models. *Journal of the Royal Statistical Society B*, **37**, 444–456.

Billard, L. and Meshkani, M.R. (1995) Estimation of a stationary Markov chain. *Journal of the American Statistical Association*, **90**, 307–315.

Cappé, O., Moulines, E., and Rydén, T. (2005) *Inference in Hidden Markov Models*. Berlin: Springer.

de Finetti, B. (1937) La prévision: ses lois logiques, ses sources subjectives, *Annales de l'Institut Henri Poincaré*, **7**, 1–68. [English translation In *Studies in Subjective Probability* (1980). H.E. Kyburg and H.E. Smokler (Eds.). Malabar, FL: Krieger, pp. 53–118.]

Delage, E. and Mannor, S. (2010) Percentile optimization for MDP with parameter uncertainty. *Operations Research*, **58**, 203–213.

Deltour, I., Richardson, S., and Le Hesran, J.Y. (1999) Stochastic algorithms for Markov models estimation with intermittent missing data. *Biometrics*, **55**, 565–573.

Diaconis, P. and Freedman, D. (1980) De Finetti's theorem for Markov chains *Annals of Probability*, **8**, 115–130.

Diaconis, P. and Rolles, S. (2006) Bayesian analysis for reversible Markov chains. *Annals of Statistics*, **34**, 1270–1292.

Dion, J.P. (1972) Estimation des probabilités initiales et de la moyenne d'un processus de Galton-Watson. *Ph.D. Thesis*. Montreal: University of Montreal.

Dion, J.P. (1974) Estimation of the mean and the initial probabilities of the branching processes. *Journal of Applied Probability*, **11**, 687–694.

Dubes, R.C. and Donoghue, P.J. (1970) Bayesian learning in Markov chains with observable states. *Proceedings of the Southeastern Symposium on Systems Theory*, University of Florida.

Eichelsbacher, P. and Ganesh, A. (2002) Bayesian inference for Markov chains. *Journal of Applied Probability*, **39**, 91–99.

Fearnhead, P. and Clifford, P. (2003) On-Line inference for hidden Markov models via particle filters. *Journal of the Royal Statistical Society B*, **65**, 887–899.

Ganesh, A.J. and O'Connell, N. (2000) A large deviation principle for Dirichlet posteriors. *Bernoulli*, **6**, 1021–1034.

Gharamani, Z. (2001) An introduction to hidden Markov models and Bayesian networks. *Journal of Pattern Recognition and Artificial Intelligence*, **15**, 9–42.

Ghavamzada, M. and Engel, Y. (2007) Bayesian policy gradient algorithms. *Advances in Neural Information Processing Systems*, **19**, 457–464.

Green, P. (1995) Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Guttorp, P. (1991) *Statistical Inference for Branching Processes*. New York: John Wiley & Sons, Inc.

Heyde, C.C. (1979) On assessing the potential severity of an outbreak of a rare infectious disease: a Bayesian approach. *Australian Journal of Statistics*, **21**, 282–292.

Holzmann, H., Munk, A., Suster, M., and Zucchini, W. (2007) Hidden Markov models for circular and linear-circular time series. *Environmental and Ecological Statistics*, **13**, 325–347.

Howard, R.A. (1960) *Dynamic Programming and Markov Process*. Cambridge, MA: MIT Press.

Huelsenbeck, J.P. and Ronquist, F. (2001) MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

Hung, W.-L. (2000) Bayesian bootstrap clones for censored Markov chains, *Biometrical Journal*, **4**, 501–510.

Johansson, M. and Olofsson, T. (2007) Bayesian model selection for Markov, hidden Markov, and multinomial models, *IEEE Signal Processing Letters*, **14**, 129–132.

Kearns, M., Mansour, Y., and Ng, A.Y. (2002) A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine Learning*, **49**, 193–208.

Lee, T.C. and Judge, G.G. (1968) Maximum likelihood and Bayesian estimation of transition probabilities. *Journal of the American Statistical Association*, **63**, 1162–1179.

Lunn, D.J., Thomas, A., Best, N. and Spiegelhalter, D. (2000) WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility, *Statistics and Computing*, **10**, 325–337.

Martin, J.J. (1967). *Bayesian Decision Problems and Markov Chains*. New York: John Wiley & Sons, Inc.

McCullock, R.E. and Tsay, R.S. (1994) Bayesian analysis of autoregressive time series via the Gibbs sampler. *Journal of Time Series Analysis*, **15**, 235–250.

McGrory, C.J. and Titterington, D.M. (2009) Variational Bayesian analysis for hidden Markov models, *Australian & New Zealand Journal of Statistics*, **51**, 227–244.

Mendoza, M. and Gutiérrez Peña, E. (2000) Bayesian conjugate analysis of the Galton-Watson process., *Test*, **9**, 149–171.

Meshkani, M.R. and Billard, L. (1992) Empirical Bayes estimators for a finite Markov chain. *Biometrika*, **79**, 185–193.

Molina, M., González, M., and Mota, M. (1998) Bayesian inference for bisexual Galton-Watson processes *Communications in Statistics - Theory and Methods*, **27**, 1055–1070.

Papangelou, F. (1996) Large deviations and the Bayesian estimation of higher-order Markov transition functions. *Journal of Applied Probability*, **33**, 18–27.

Prado, R. and West, M. (2010) *Time Series: Modeling, Computation, and Inference*. Boca Raton: Chapman and Hall.

Puterman, M.L. (1994) *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York: John Wiley & Sons, Inc.

Raftery, A.E. (1985) A model for higher order Markov chains.*Journal of the Royal Statistical Society B*, **47**, 528–539.

Richardson, S. and Green, P. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society B*, **59**, 731–792.

Roberts, S.J. and Penny, W.D. (2002) Variational Bayes for generalized autoregressive models. *IEEE Transactions on Signal Processing*, **50**, 2245–2257.

Silver, A.E. (1963) Markovian decision processes with uncertain transition probabilities or rewards. *Technical Report 1*, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA.

Scott, D. (1987) On posterior asymptotic normality and asymptotic normality of estimators for the Galton-Watson process. *Journal of the Royal Statistical Society B*, **49**, 209–214.

Scott, S.L. (2002) Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, **97**, 337–351.

Strens, M. (2000) A Bayesian framework for reinforcement learning. *Proceedings International Conference on Machine Learning*, Stanford University, California.

Sung, M., Soyer, R., and Nhan , N. (2007) Bayesian analysis of non-homogeneous Markov chains: Application to mental health data. *Statistics in Medicine*, **26**, 3000–3017.

Tsay, R.S. (2005) *Analysis of Financial Time Series*. New York: John Wiley & Sons, Inc.

Thompson, W.R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**, 285–294.

Wang, T., Lizotte, D., Bowling, M., and Schuurmans, D. (2005) Bayesian sparse sampling for on-line reward optimization. *Proceedings International Conference on Machine Learning*, Bonn, Germany.

Welton, N.J. and Ades, A.E. (2005) Estimation of Markov chain transition probabilities and rates from fully and partially observed data: Uncertainty propagation, evidence synthesis, and model calibration. *Medical Decision Making*, **25**, 633–645.

Williams, R. (1992) Simple statistical gradient following algorithms for connectionist reinforcement learning. *Machine Learning*, **8**, 229–256.

4

Continuous time Markov chains and extensions

4.1 Introduction

In this chapter, we consider inference, prediction and decision-making tasks with continuous time Markov chains (CTMCs) and some of their extensions. Our interest in such processes is twofold. First, they constitute an extension of discrete time Markov chains, which were dealt with in Chapter 3. Throughout this chapter, we shall use some of the results shown there. Second, CTMCs have many applications, either directly or as basic building blocks in areas such as queueing, reliability analysis, risk analysis, or biomedical applications, some of which are presented in later chapters.

CTMCs are continuous time stochastic processes with discrete state space. We shall concentrate on homogeneous CTMCs with finite state space. In those processes, the system remains an exponential time at each state and, when leaving such state, it evolves according to probabilities that depend only on the leaving state. The basic probabilistic results for CTMCs of this type are outlined in Section 4.2.

The parameters of interest of the CTMC are the transition probabilities and the exponential permanence rates. Given a completely observed CTMC, inference for the transition probabilities can be carried out as in Chapter 3. In Section 4.3, we show how to extend this procedure to consider the CTMC rates and, as a relevant by-product, we deal with inference for the intensity matrix of the process. Short- and long-term forecasting are also considered. In Section 4.4, we illustrate the proposed procedures with an application to hardware availability.

In Section 4.5, we consider semi-Markovian processes, which generalize CTMCs by allowing the permanence times to be nonexponential and then, in Section 4.6, we outline some decision-making issues related to CTMCs and sketch a Markov chain Monte Carlo (MCMC) approach to solving semi-Markovian decision processes

Bayesian Analysis of Stochastic Process Models, First Edition. David Rios Insua, Fabrizio Ruggeri and Michael P. Wiper.

when there is uncertainty in the process parameters, which we illustrate through a maintenance example. The chapter finishes with a brief discussion.

4.2 Basic setup and results

In this section, we shall outline the most important probabilistic results for CTMCs. We shall assume that $\{X_t\}_{t\in T}$ is a continuous time stochastic process that evolves within a finite state space, say $E = \{1, 2, \ldots, K\}$. When the process enters into state i, it remains there for an exponentially distributed time period with mean $1/\nu_i$. At the end of this time period, the process will move to a different state $j \neq i$ with probability p_{ij}, such that $\sum_{j=1}^{K} p_{ij} = 1$, $\forall i$, and $p_{ii} = 0$. Clearly, for physical or logical reasons, some additional p_{ij} could also be zero. As in Chapter 3, the transition probability matrix is defined to be $\mathbf{P} = (p_{ij})$. This defines an embedded (discrete time) Markov chain. The process $\{X_t\}$ will be designated a CTMC with parameters $\mathbf{P}$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)^T$.

One important class of CTMCs, which will be analyzed in detail in later chapters are birth–death processes.

Example 4.1: A birth–death process is a particular example of a CTMC with state space $\{0, 1, 2, \ldots, K\}$, where the states represent the population size. Transitions in this process can occur either as single births, with rate λ_i or single deaths, with rate μ_i, for $i = 0, \ldots, K$, where $\mu_0 = \lambda_K = 0$. Therefore, the transition probabilities for this process are $p_{i,i+1} = \lambda_i/(\lambda_i + \mu_i)$, $p_{i,i-1} = \mu_i/(\lambda_i + \mu_i)$ and $p_{ij} = 0$ for $i = 0, \ldots, K$ and $j \notin \{i-1, i+1\}$. Also, the times between transitions are exponentially distributed with rate $\nu_i = \lambda_i + \mu_i$.

The birth–death process is equivalent to a Markovian queueing system where, given that there are i people in the system, arrivals occur with rate λ_i and a service is completed with rate μ_i. Processes of this type are examined in Chapter 7.

A pure birth process with infinite state space $\{0, 1, 2, \ldots\}$, $\mu_i = 0$ and $\lambda_i = \lambda$ for all i is called a Poisson process, which is the theme of Chapter 5. △

The parameters

$$r_{ij} = \nu_i p_{ij}$$

are called jumping intensities (from state i into state j). In addition, we set $r_{ii} = -\sum_{j\neq i} r_{ij} = -\nu_i$, $i \in \{1, \ldots, K\}$, and place all r_{ij} in the intensity matrix $\boldsymbol{\Lambda} = (r_{ij})$, also called the infinitesimal generator of the process, which have a key role in later computations.

The short-term behavior of the CTMC may be described through the forward Kolmogorov system of differential equations. Consider the transition probability functions

$$P_{ij}(t) = P(X_{t+s} = j | X_s = i) = P(X_t = j | X_0 = i), \tag{4.1}$$

which describe the probability that the system is in state j if it is currently in state i and a time t elapses. Then, under suitable regularity conditions (see, e.g., Ross, 2009), we have

$$P'_{ij}(t) = \sum_{k \neq j} r_{kj} P_{ik}(t) - \nu_j P_{ij}(t) = \sum_{k} r_{kj} P_{ik}(t).$$

Note that we may write this system jointly as

$$\mathbf{P}'(t) = \mathbf{\Lambda}\mathbf{P}(t) \tag{4.2}$$

$$\mathbf{P}(0) = \boldsymbol{I},$$

where $\mathbf{P}(t) = (P_{ij}(t))$ is the matrix of transition probability functions and $\boldsymbol{I}$ is the identity matrix. The analytic solution of this system is $\mathbf{P}(t) = \exp(\mathbf{\Lambda}t)$, which can be solved, for given t, using matrix exponentiation, a problem reviewed in, for example, Moler and Van Loan (2003).

The simplest case is when $\mathbf{\Lambda}$ is diagonalizable with different eigenvalues, which holds with no significant loss of generality (Geweke *et al.*, 1986). We then decompose $\mathbf{\Lambda} = \boldsymbol{SDS}^{-1}$, where $\boldsymbol{D}$ is the diagonal matrix with the distinct eigenvalues $\lambda_1, \ldots, \lambda_K$ of $\mathbf{\Lambda}$ as its entries, and $\boldsymbol{S}$ is an invertible matrix consisting of the eigenvectors corresponding to the eigenvalues in $\mathbf{\Lambda}$. Then, we have

$$\exp(\mathbf{\Lambda}t) = \sum_{i=0}^{\infty} \frac{(\mathbf{\Lambda}t)^i}{i!} = \sum_{i=0}^{\infty} \frac{(\boldsymbol{SDS}^{-1})^i t^i}{i!} = \boldsymbol{S}\left[\sum_{i=0}^{\infty} \frac{(\boldsymbol{D}t)^i}{i!}\right]\boldsymbol{S}^{-1}$$

$$= \boldsymbol{S}\begin{pmatrix} \exp(\lambda_1 t) & 0 & \ldots & 0 \\ 0 & \exp(\lambda_2 t) & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \exp(\lambda_K t) \end{pmatrix}\boldsymbol{S}^{-1}$$

As with the discrete time case, forecasting the long-term behavior of a CTMC means that we need to consider the equilibrium distribution. Under suitable conditions (see, e.g., Ross, 2009) for given $\mathbf{P}$ and $\boldsymbol{\nu}$, we know that, if it exists, the equilibrium distribution $\{\pi_j\}_{j=1}^{K}$ is obtained through the solution of the system

$$\nu_j \pi_j = \sum_{i \neq j} r_{ij} \pi_i, \qquad \forall j \in \{1, \ldots, K\}, \tag{4.3}$$

$$\sum_j \pi_j = 1; \qquad \pi_j \geq 0.$$

4.3 Inference and prediction for CTMCs

Here, we study inference and prediction for CTMCs. We first consider inference for chain parameters and then examine the forecasting of both the short- and long-term behavior of a CTMC. We will suppose throughout the most general case where the transition matrix, **P**, and the transition rates, $\boldsymbol{\nu}$, are unknown and unrelated, that is, that the elements of **P** are not known functions of $\boldsymbol{\nu}$.

Assume that we observe the initial state of the chain, say x_0 and the times, t_i, and states, x_i, for $i = 1, \ldots, n$, of the first n transitions of the chain. Then, the likelihood function can be written as

$$l(\mathrm{P}, \boldsymbol{\nu}|\text{data}) = \prod_{i=1}^{n} \nu_{x_{i-1}} \exp(-\nu_{x_{i-1}}(t_i - t_{i-1}))p_{x_{i-1}x_i} \propto \prod_{i=1}^{K} \nu_i^{n_i} \exp(-\nu_i T_i) \prod_{j=1}^{K} p_{ij}^{n_{ij}}, \tag{4.4}$$

where n_{ij} is the number of observed transitions from i to j, T_i is the total time spent in state i and $n_i = \sum_{j=1}^{K} n_{ij}$ is the total number of transitions out of state i, for $i, j \in \{1, \ldots, K\}$. Given the lack of memory property of the exponential distribution, many alternative experiments have likelihood functions of the same form.

4.3.1 Inference for the chain parameters

The likelihood function in (4.4) can be written as

$$l(\mathrm{P}, \boldsymbol{\nu}|\text{data}) = l_1(\mathrm{P}|\text{data})l_2(\boldsymbol{\nu}|\text{data}),$$

where $l_1(\mathrm{P}|\text{data}) = \prod_{i=1}^{K} \prod_{j=1}^{K} p_{ij}^{n_{ij}}$ and $l_2(\boldsymbol{\nu}|\text{data}) = \prod_{i=1}^{K} \nu_i^{n_i} \exp(-\nu_i T_i)$, which implies that, given independent prior distributions for **P** and $\boldsymbol{\nu}$, the posterior distributions will also be independent and inference for **P** and $\boldsymbol{\nu}$ can be carried out separately. This setup is described through the influence diagram in Figure 4.1.

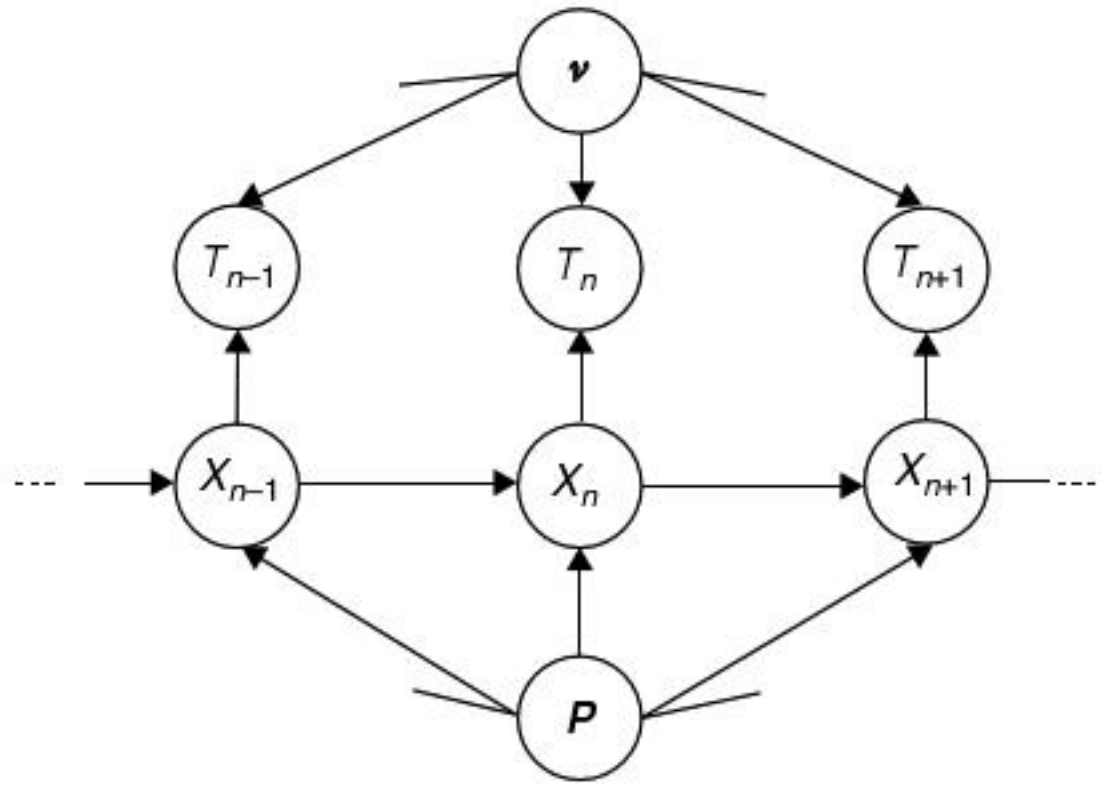

Figure 4.1 Influence diagram for a CTMC.

Inference for the transition probabilities can then proceed as in Chapter 3. Assuming a known initial state and a matrix beta prior distribution as outlined in Section 3.3.2. then the posterior distribution is also matrix beta from (3.4). The case of an unknown initial state can also be dealt with the methods of Section 3.3.6. A natural conjugate prior for the permanence rates is also available. If we assume that the rates have independent gamma prior distributions, $\nu_i \sim \text{Ga}(a_i, b_i)$, for $i = 1, \ldots, K$, then combining prior and likelihood, we see that $\nu_i|\text{data} \sim \text{Ga}(a_i + n_i, b_i + T_i)$ for $i = 1, \ldots, K$.

Given the above posteriors, we may provide inference about the intensity matrix, which will be of relevance later on, as follows:

- When the posterior distributions are sufficiently concentrated, we could summarize them through the posterior modes, $\hat{\nu}_i$ and $\hat{p}_{ij}$, to estimate $\hat{r}_{ij} = \hat{\nu}_i \hat{p}_{ij}$, $i \neq j$. For $i = j$, we set $\hat{r}_{ii} = -\hat{\nu}_i$, $i = 1, \ldots, m$. For example, for the Dirichlet-multinomial model it would be,

$$\hat{\nu}_i = \frac{\alpha_i + n_i - 1}{\beta_i + \sum_{j=1}^{n_i} t_{ij}}; \; \hat{p}_{ij} = \frac{\delta_{ij} + n_{ij} - 1}{\sum_{l \neq i}(n_{il} + \delta_{il}) - k + 1}; \; \hat{r}_{ij} = \hat{\nu}_i \hat{p}_{ij}, \hat{r}_{ii} = -\hat{\nu}_i.$$

- Otherwise, we would use posterior samples $\{\nu^\eta\}_{\eta=1}^N$ and $\{\mathbf{P}^\eta\}_{\eta=1}^N$, to obtain samples from the posterior $\{r_{ij}^\eta = \nu_i^\eta p_{ij}^\eta\}_{\eta=1}^N$, $i \neq j$. For $i = j$, we would use the posterior sample $\{r_{ii}^\eta = -\nu_i^\eta\}_{\eta=1}^N$, $i = 1, \ldots, k$. We may then summarize all samples appropriately, through, for example, their sample means $\frac{1}{N}\sum_{\eta=1}^N r_{ij}^\eta, \forall i, j$.

4.3.2 Forecasting short-term behavior

Here, we shall consider forecasting the short-term behavior of a CTMC. This can be based on the solution of the system of differential equations described in (4.2), which characterize short-term behavior, when parameters $\mathbf{P}$ and $\boldsymbol{\nu}$ are fixed. However, we need to take into account the uncertainty about parameters to estimate the predictive matrix of transition probabilities $\mathbf{P}(t)|\text{data}$. Various options can be considered.

First, when the posterior distributions of $\mathbf{P}$ and $\boldsymbol{\nu}$ are sufficiently concentrated, we could summarize them through the posterior modes, $\hat{\boldsymbol{\nu}}$ and $\hat{\mathrm{P}}$, so that, assuming $\boldsymbol{\Lambda}(\hat{\mathbf{P}}, \hat{\boldsymbol{\nu}})$ is diagonalizable with K different eigenvalues, we can estimate $\mathrm{P}(t)|\text{data}$ through

$$\boldsymbol{S}(\hat{\mathbf{P}}, \hat{\boldsymbol{\nu}}) \begin{pmatrix} \exp(\lambda_1(\hat{\mathrm{P}}, \hat{\boldsymbol{\nu}})t) & 0 & \ldots & 0 \\ 0 & \exp(\lambda_2(\hat{\mathbf{P}}, \hat{\boldsymbol{\nu}})t) & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \exp(\lambda_k(\hat{\mathbf{P}}, \hat{\boldsymbol{\nu}})t) \end{pmatrix} \boldsymbol{S}(\hat{\mathbf{P}}, \hat{\boldsymbol{\nu}})^{-1}.$$

More generally, we could obtain Monte Carlo samples, $\boldsymbol{\nu}^{(s)}, \mathbf{P}^{(s)}$, for $s = 1, \ldots, S$. Then, for each s, solve the corresponding decomposition. This would provide us with a sample $\mathbf{P}(t)^{(s)}$, which might be summarized according to, for example, the sample mean, $\frac{1}{S}\sum_s \mathbf{P}(t)^{(s)}$.

This procedure is easily implemented when K is relatively small. However, for large K, it may be that the matrix exponentiation operation may be too computationally intensive to be used within a Monte Carlo type scenario. One possibility is to use a reduced order model (ROM), as described in Section 2.4.1. In this case, if m is the maximum number of matrix exponentiations that our computational budget allows, we would proceed as follows:

1. For $s = 1, \ldots, S$, sample $\boldsymbol{\nu}^{(s)}, \mathbf{P}^{(s)}$ from the relevant posteriors.
2. Cluster the sampled values into m clusters and spread the centroids to obtain the ROM range $(\boldsymbol{\nu}^{(i)}, \mathbf{P}^{(i)})$ for $i = 1, \ldots, m$.
3. Compute the optimal ROM probabilities by solving

$$\min_{q_1,\ldots,q_m} e(q_1, \ldots, q_m)$$
$$\text{s.t. } \sum_{r=1}^{m} q_r = 1, \quad q_r \geq 0, r = 1, \ldots, m.$$

4. For $i = 1$ to m
   (a) Compute $\boldsymbol{\Lambda}(\boldsymbol{\nu}^{(i)}, \mathbf{P}^{(i)})$
   (b) Decompose $\boldsymbol{\Lambda}(\boldsymbol{\nu}^{(i)}, \mathbf{P}^{(i)}) = \boldsymbol{S}(\boldsymbol{\nu}^{(i)}, \mathbf{P}^{(i)})\boldsymbol{D}(\boldsymbol{\nu}^{(i)}, \mathbf{P}^{(i)})\boldsymbol{S}^{-1}(\boldsymbol{\nu}^{(i)}, \mathbf{P}^{(i)})$
   (c) Compute $\mathbf{P}(t)|\nu^{(i)}, \mathbf{P}^{(i)}$ through

$$\boldsymbol{S}\left(\mathbf{P}^{(i)}, \boldsymbol{\nu}^{(i)}\right) \begin{pmatrix} \exp(\lambda_1(\mathbf{P}^{(i)}, \boldsymbol{\nu}^{(i)})t) & 0 & \ldots & 0 \\ 0 & \exp(\lambda_2(\mathbf{P}^{(i)}, \boldsymbol{\nu}^{(i)})t) & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \exp(\lambda_m(\mathbf{P}^{(i)}, \boldsymbol{\nu}^{(i)})t) \end{pmatrix} \times \boldsymbol{S}\left(\mathbf{P}^{(i)}, \boldsymbol{\nu}^{(i)}\right)^{-1}.$$

5. Approximate $\mathbf{P}(t)|$data through

$$\sum_{i=1}^{m} q_i \ \mathbf{P}(t)|\boldsymbol{\nu}^{(i)}, \mathbf{P}^{(i)}.$$

Alternatively, for very high-dimensional problems, a purely simulation-based approach, which entirely eliminates the need for matrix exponential computations can also be implemented. In this case, for each element, $\boldsymbol{\nu}^{(s)}, \mathbf{P}^{(s)}$, of a Monte Carlo sample of size S, a set of state transitions and their corresponding transition times can be generated. Then, for given t, we can define $X_t^{(s)}$ to be the state value at time t. We can now approximate $P(t)(x)|$data through $\frac{1}{S}\sum_{s=1}^{S} I(X_t^{(s)} = x)$, where $I(\cdot)$ is an indicator function. An advantage of this approach, when compared with the previous techniques, is that essentially no extra computation is required to compute the distributions at different times, whereas in the previous cases, separate matrix exponential computations are needed for each t. Approaches of this type are analyzed in more detail in Chapter 9.

4.3.3 Forecasting long-term behavior

Depending on the concentration of the posterior and the computational budget available, long-term forecasting of the CTMC behavior can also be undertaken in a number of different ways.

First, if the posterior distributions are sufficiently concentrated, we could substitute the parameters by, for example, their posterior modes, and solve system (4.3), to obtain an approximate point summary of the predictive equilibrium distribution $\{\hat{\pi}_i\}_{i=1}^m$. However, as earlier, this approach does not give a measure of uncertainty.

Otherwise, we may obtain samples from the posteriors, $\boldsymbol{\nu}^{(s)}$, $\mathbf{P}^{(s)}$, for $s = 1, \ldots, S$ and, consequently, obtain the sampled probabilities, $\pi_i^{(s)}$, for $s = 1, \ldots, S$, from the predictive equilibrium distribution through the repeated solution of system (4.3). If needed, we could summarize it through, for example, their means,

$$\hat{\pi}_i = \frac{1}{S}\sum_{s=1}^{S} \pi_i^{(s)} \quad \text{for } i = 1, \ldots, K.$$

For large K, solving the system of equations required may be costly computationally and we may opt for using a ROM as explained earlier.

4.3.4 Predicting times between transitions

Given the current state, prediction of the time to the next transition is much more straightforward. If the current state is i, given the gamma posterior distribution, $\nu_i|\text{data} \sim \text{Ga}(a_i + n_i, b_i + T_i)$, then if T is the time to the next transition, we have

$$P(T \leq t|\text{data}) = \left(\frac{b_i + T_i}{b_i + T_i + t}\right)^{a_i+n_i}.$$

Predictions of times up to more than one transition can also be handled by using Monte Carlo approaches as outlined earlier.

4.4 Case study: Hardware availability through CTMCs

In recent years, there has been increasing interest in reliability, availability, and maintainability (RAM) analyses of hardware (HW) systems and, in particular, safety critical systems. Sometimes such systems can be modeled using CTMCs, which, in this context, describe stochastic processes which evolve through a discrete set of states, some of which correspond to ON configurations and the rest to OFF configurations. Transition from an ON to an OFF state entails a system failure, whereas a transition from an OFF to an ON system implies a repair. Here, we shall emphasize availability, which is a key performance parameter for information technology systems. Indeed, there are many hardware configurations aimed at attaining very high system availability, for example, 99.999% of time, through transfer of workload when

one, or more, system components fail, or intermediate failure states with automated recovery; see Kim *et al.* (2005) for details. Thus, we are concerned with hardware systems, which we assume can be modeled through a CTMC. We shall consider that states $\{1, 2, \ldots, l\}$ correspond to operational (ON) configurations, whereas states $\{l+1, \ldots, K\}$ correspond to OFF configurations.

A classical approach to availability estimation of CTMC HW systems would calculate maximum likelihood estimates for the CTMC parameters and then compute the equilibrium distribution given these, and finally, estimate the long-term fraction of time that the system remains in ON configurations. A shortfall of this approach is that it does not account for parameter uncertainty, whereas the fully Bayesian framework we adopt here automatically incorporates this uncertainty. Also, both short-term and long-term forecasting can be carried out.

Initially, we shall consider steady-state prediction of the system. In this case, the availability is the sum of the equilibrium probabilities for the ON states, conditional on the rates and transition probabilities, $\boldsymbol{\nu}$, $\mathbf{P}$, that is

$$A|\nu, P = \sum_{i=1}^{l} \pi_i|\boldsymbol{\nu}, \mathbf{P}.$$

If the posterior parameter distributions are precise, we may use the approximate predictive steady-state availability, based on the approximate equilibrium distribution

$$\hat{A}|\text{data} \simeq \sum_{i=1}^{l} \hat{\pi}_i$$

to estimate the predictive availability. Otherwise, if the posteriors are not concentrated, we would obtain a predictive steady-state availability sample, based on the sample obtained in Section 4.3.3

$$\{A^{(s)} = \sum_{i=1}^{l} \pi_i^{(s)}\}_{s=1}^{S},$$

and summarize it accordingly, for example, through $\frac{1}{S}\sum_{s=1}^{S} A^{(s)}$. Finally, if the computational budget only allows for m equilibrium distribution computations, then we could approach the posterior availability through

$$\sum_{i=1}^{m} q_i \left(\sum_{j=1}^{l} \pi_j^{(i)}|\nu^{(i)}, \mathbf{P}^{(i)} \right),$$

based on the ROM equilibrium distribution mentioned in Section 4.3.3.

As discussed in Lee (2000), we may be also interested in a type of short-term availability, called interval availability. Define the random variable Y_t

$$Y_t|\boldsymbol{\nu}, \mathbf{P} = \begin{cases} 1, \text{ if } X_t|\boldsymbol{\nu}, \mathbf{P} \in \{1, 2, \ldots, l\}, \\ 0, \qquad\qquad \text{otherwise} \end{cases}$$

and

$$A_t|\boldsymbol{\nu}, \mathbf{P} = \frac{1}{t}\int_0^t (Y_u|\boldsymbol{\nu}, \mathbf{P})\ \mathrm{d}u.$$

Then, the interval availability is defined through

$$I_t|\boldsymbol{\nu}, \mathbf{P} = E[A_t|\boldsymbol{\nu}, \mathbf{P}] = \frac{1}{t}\sum_{j=1}^{l}\int_0^t \pi_j(u|\boldsymbol{\nu}, \mathbf{P})\,\mathrm{d}u,$$

where $\pi_j(u|\boldsymbol{\nu}, \mathbf{P}) = P(X_u = j|\boldsymbol{\nu}, \mathbf{P})$. We may approximate it with a one dimensional integration method, like Simpson's rule.

The key computation is that of the $\pi_j(t|\boldsymbol{\nu}, \mathbf{P})$ terms, $j = 1, \ldots, K$. To do this, we solve the Chapman–Kolmogorov system of differential equations (see, e.g., Ross, 2009),

$$\pi'(t|\nu, P) = (\Lambda|\boldsymbol{\nu}, \mathbf{P}) \cdot \pi(t|\boldsymbol{\nu}, \mathbf{P}); \quad t \in [0, T),$$

$$\pi(0|\boldsymbol{\nu}, \mathbf{P}) = \pi^{(0)},$$

where $\pi(t|\boldsymbol{\nu}, \mathbf{P}) = (\pi_1(t|\boldsymbol{\nu}, \mathbf{P}), \ldots, \pi_K(t|\boldsymbol{\nu}, \mathbf{P}))$, $\boldsymbol{\pi}^{(0)} = (\pi_1^{(0)}, \pi_2^{(0)}, \ldots, \pi_K^{(0)})$ is the initial state probability vector, and $\boldsymbol{\Lambda}|\boldsymbol{\nu}, \mathbf{P}$ is the intensity matrix, conditional on $\boldsymbol{\nu}, \mathbf{P}$. Its analytic solution is

$$\pi(t|\boldsymbol{\nu}, \mathbf{P}) = \pi^{(0)} \exp(\boldsymbol{\Lambda} t|\boldsymbol{\nu}, \mathbf{P}).$$

Note that again the key operation is that of matrix exponentiation.

We may then define the posterior interval availability through

$$I_t|\text{data} = \int\int E[A_t|\boldsymbol{\nu}, \mathbf{P}]\pi(\boldsymbol{\nu}, \mathbf{P}|\text{data})\, d\mathbf{P}\, d\boldsymbol{\nu}.$$

As discussed previously in this chapter, at least three approaches can be considered.

When the posteriors are precise enough, we may summarize them, for example, through the posterior modes $\hat{\mathbf{P}}$, $\hat{\boldsymbol{\nu}}$, and use $E[A_t|\hat{\mathbf{P}}, \hat{\boldsymbol{\nu}}.]$ as a summary of the predictive

availability. Otherwise, if the posteriors are not precise, for appropriate posterior samples $\{\mathbf{P}^{(s)}, \boldsymbol{\nu}^{(s)}\}_{s=1}^{S}$, we could use

$$\frac{1}{S}\sum_{s=1}^{S} E\left[A_t|\mathbf{P}^{(s)}, \boldsymbol{\nu}^{(s)}\right].$$

Finally, if the computational budget allows only for m availability computations, then we could approach the predictive availability through

$$\sum_{i=1}^{m} q_i\, E\left[A_t|\mathbf{P}^{(i)}, \boldsymbol{\nu}^{(i)}\right],$$

based on the ROM sample obtained in Section 4.3.3.

Example 4.2: We shall consider a system which is described by a dual-duplex model, with transition diagram as in Figure 4.2, along with the jumping intensities $r_{ij} = \nu_i p_{ij}$. The dual-duplex system is designed to detect a fault using a hardware comparator that switches to a hot standby redundancy. To improve reliability and safety, the dual-duplex system is designed in double modular redundancy. Because the dual-duplex system has high reliability, availability, and safety, it can be applied in embedded control systems like airplanes. It has two ON states $\{1, 2\}$ and two OFF states $\{3, 4\}$. The transition probability matrix is

$$\mathbf{P} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & p_{12} & p_{13} & 0 \\ p_{21} & 0 & p_{23} & p_{24} \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{array}.$$

The permanence rates are ν_1, ν_2, ν_3 and ν_4.

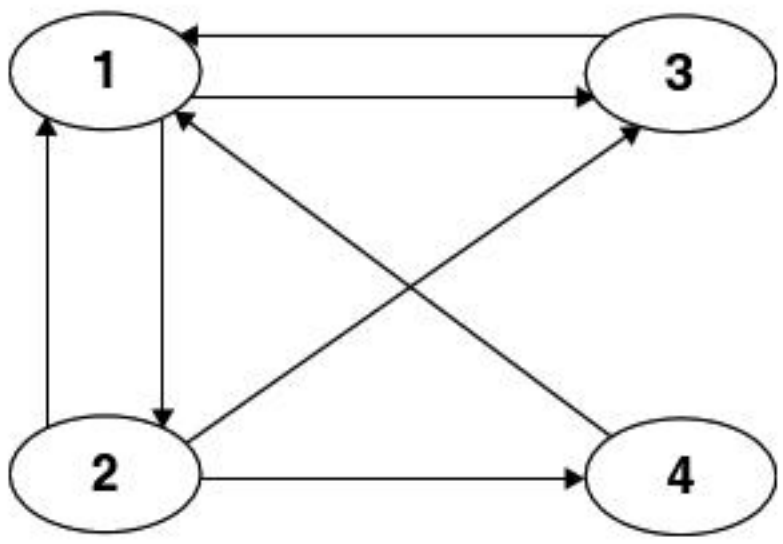

Figure 4.2 Transition diagram for the dual-duplex system.

For rows 1 and 2, we assume Dir(0, 1, 1, 0) and Dir(1, 0, 1, 1) priors, respectively. On the basis of the data counts $n_{12} = 10$, $n_{13} = 4$, $n_{21} = 7$, $n_{23} = 1$, $n_{24} = 2$, we get

$$(p_{11}, p_{12}, p_{13}, p_{14})|\text{data} \sim \text{Dir}(0, 1 + 10, 1 + 4, 0)$$
$$(p_{21}, p_{22}, p_{23}, p_{24})|\text{data} \sim \text{Dir}(1 + 7, 0, 1 + 1, 1 + 2).$$

We are relatively sure that there will be around one failure every 10 hours for ν_1 and ν_2, and, therefore, assume priors $\nu_1 \sim \text{Ga}(0.1, 1)$ and $\nu_2 \sim \text{Ga}(0.1, 1)$. We are less sure about ν_3, ν_4, expecting around 5 repairs per hour, therefore, assuming priors Ga(10, 2) and Ga(10, 2) for ν_3, ν_4. On the basis of the data available (for state 1, 14 times which add up 127.42; 10 with sum 86.81, for state 2; 5 with sum 1.09, for state 3; and, 2 with sum 0.27, for state 4), we get the posterior parameters in Table 4.1.

Table 4.1 Posterior parameters of permanence rates.

	α	β
ν_1	$0.1 + 14$	$1 + 127.42$
ν_2	$0.1 + 10$	$1 + 86.81$
ν_3	$10 + 5$	$2 + 1.09$
ν_4	$10 + 2$	$2 + 0.27$

For fixed P, ν, the system that provides the equilibrium solution in this case is

$$\begin{cases} \nu_1\pi_1 = r_{21}\pi_2 + r_{31}\pi_3 + r_{41}\pi_4, \\ \nu_2\pi_2 = r_{12}\pi_1, \\ \nu_3\pi_3 = r_{13}\pi_1 + r_{23}\pi_2, \\ \nu_4\pi_4 = r_{24}\pi_2, \\ \pi_1 + \pi_2 + \pi_3 + \pi_4 = 1, \\ \pi_i \geq 0, \end{cases}$$

which solves to give $\pi_1 = \frac{\nu_2\nu_3\nu_4}{\Delta}$, $\pi_2 = \frac{r_{12}\nu_3\nu_4}{\Delta}$, $\pi_3 = 1 - (\pi_1 + \pi_2 + \pi_3)$, $\pi_4 = \frac{r_{12}r_{24}\nu_3}{\Delta}$, with $\Delta = \nu_2\nu_3\nu_4 + r_{12}\nu_3\nu_4 + r_{13}\nu_2\nu_4 + r_{12}r_{23}\nu_4 + r_{12}r_{24}\nu_3$. Figure 4.3 shows density plots for the posterior equilibrium distribution. We may summarize this through the mean probabilities which are

$$\hat{\pi}_1 = 0.5931,\ \hat{\pi}_2 = 0.3990,\ \hat{\pi}_3 = 0.0059,\ \hat{\pi}_4 = 0.0020.$$

Finally, to estimate the system availability, we use the outlined procedure to compute the value of the state probability vector $\pi(t)|\boldsymbol{\nu}, \mathbf{P}$ at each point of the interval $[0, t)$, divided into 200 subintervals for Simpson's rule. We plot the system availability in Figure 4.4, when the system was initially in state 1. We have also plotted 95% predictive bands around the central values for each situation. We can observe that, for the case of availability, the uncertainty is, in practice, negligible, with

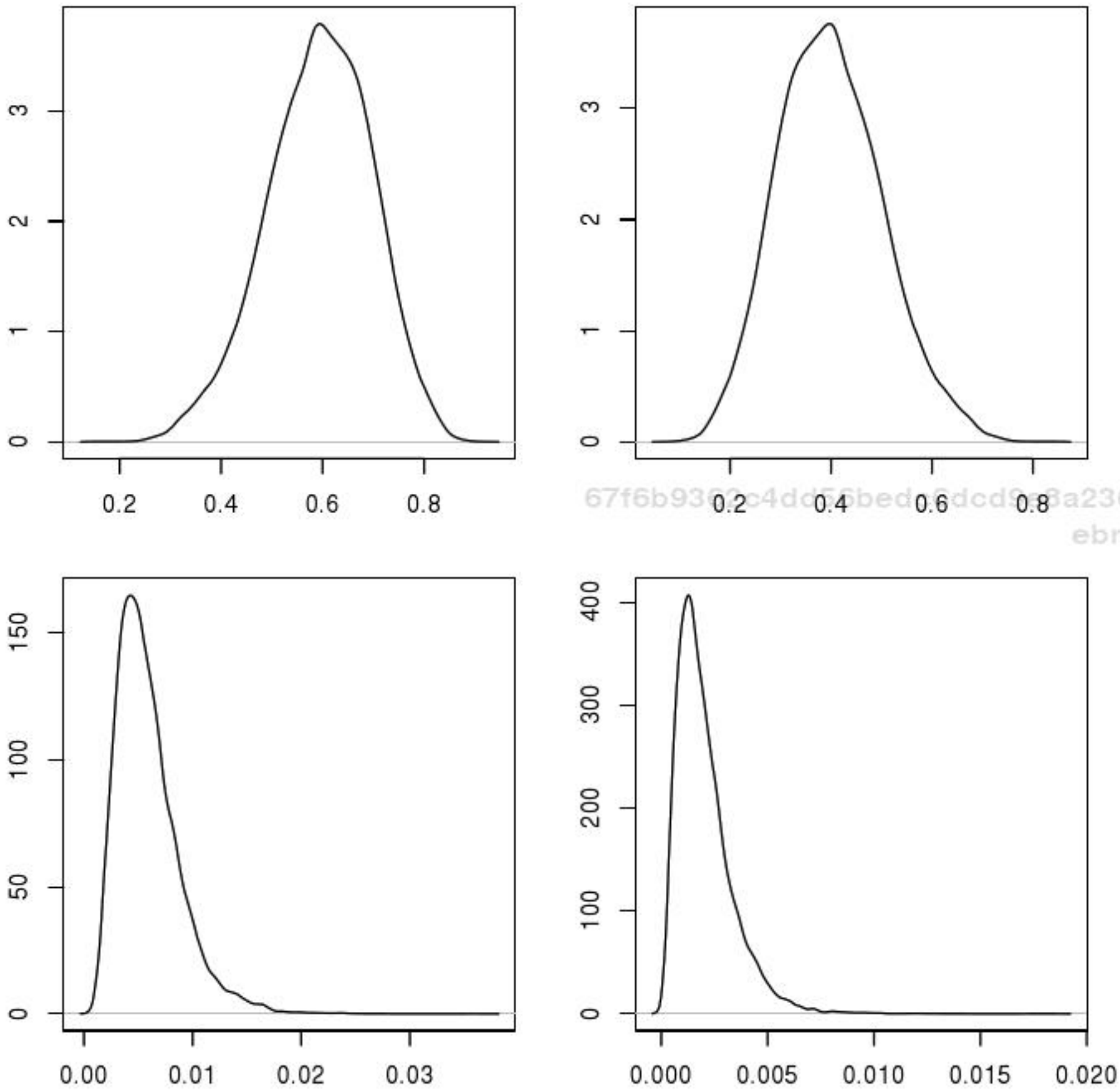

Figure 4.3 Posterior equilibrium distribution.

relative errors less than 1%. This is so because the dual-duplex system is designed as a high-availability device. △

4.5 Semi-Markovian processes

In this section, we analyze semi-Markovian process (SMP) models. These generalize CTMCs by assuming that the times between transitions are not necessarily exponential. Formally, a semi-Markovian process is defined as follows. Let $\{X_t\}_{t\in T}$ be a continuous time stochastic process which evolves within a finite state space with states $E = \{1, 2, \ldots, K\}$. When the process enters a state i, it remains there for a random time T_i, with parameters $\boldsymbol{v}_i$, which is positive with probability 1. Let $f_i(.|\boldsymbol{v}_i)$ and $F_i(.|\boldsymbol{v}_i)$ be the density and distribution functions of T_i, respectively, and define $\mu_i = E[T_i|\boldsymbol{v}_i]$. When leaving state i, we assume, as earlier, that the process moves

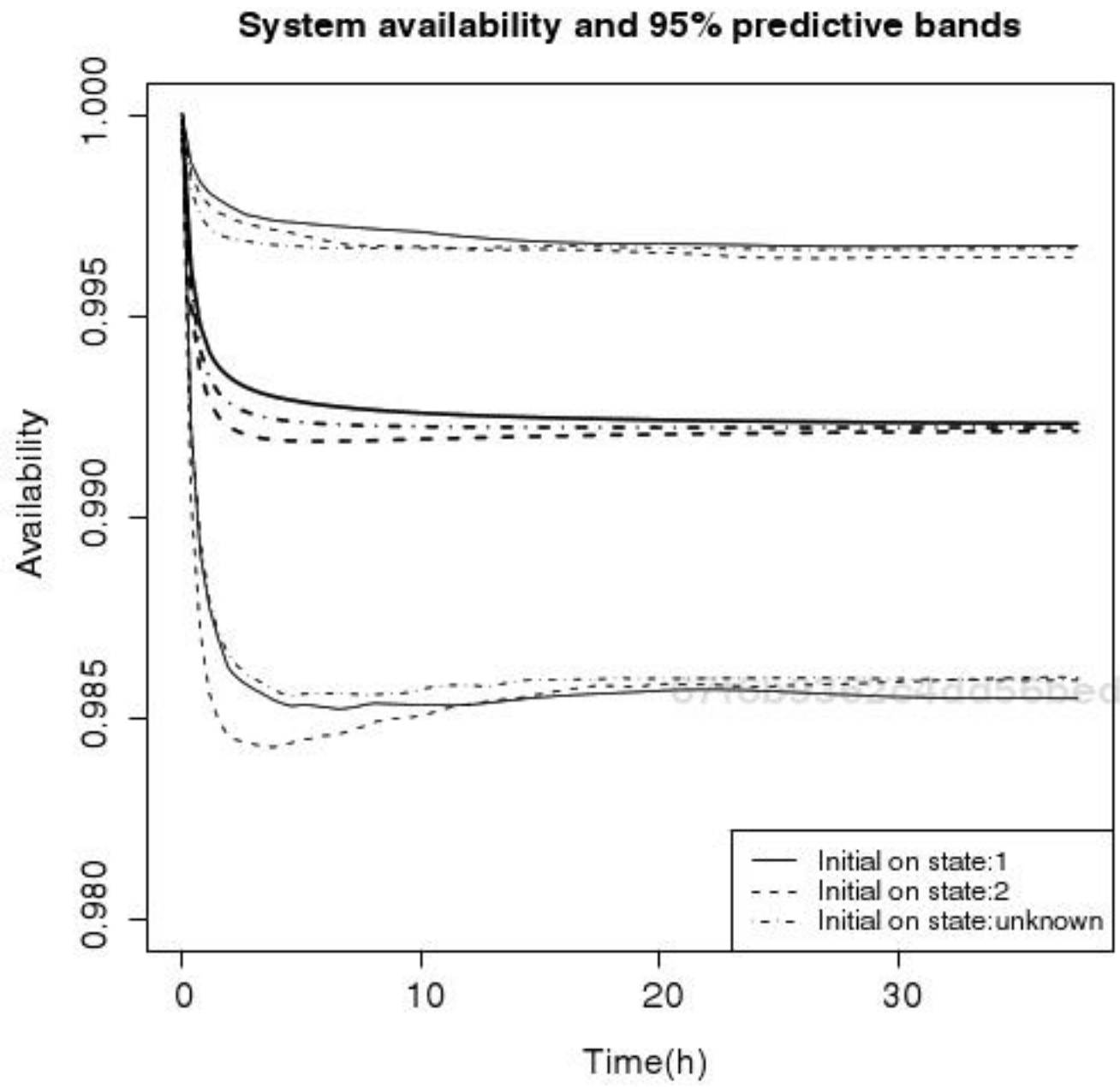

Figure 4.4 System availability.

to state j with probability p_{ij}, with $\sum_j p_{ij} = 1, \ \forall i$, and $p_{ii} = 0$. As for CTMCs, the transition probability matrix, $\mathbf{P} = (p_{ij})$, defines an embedded DTMC.

Thus, the parameters for the SMP will be $\mathbf{P} = (p_{ij})$ and $\boldsymbol{\nu} = (\nu_i)$. Given that both the states at transition and the times between transitions are observed, inference about $\mathbf{P}$ follows the same procedure as before. We shall assume that we have available the posterior distribution of $\boldsymbol{\nu}$, possibly through a sample.

We now describe long-term forecasting of the proportion of time the system spends in each of the states. To compute long-term proportions of time in state j, which we represent as π_j, for fixed $\mathbf{P}$ and $\boldsymbol{\nu}$, we need to do the following:

1. Compute, if it exists, the equilibrium distribution $\bar{\boldsymbol{\pi}}$ of the embedded Markov chain, whose transition matrix $\mathbf{P}$ is described by

$$\bar{\boldsymbol{\pi}} = \bar{\boldsymbol{\pi}}\mathbf{P}$$

$$\sum_{i=1}^{K} \bar{\pi}_i = 1, \qquad \bar{\pi}_i \geq 0.$$

2. Compute, if they exist, the expected holding times at each state, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$.
3. Compute

$$\pi = \sum_{i=1}^{K} \bar{\pi}_i \mu_i . \tag{4.5}$$