

that to the order of  $O(N_0^{-2})$ ,  $\{X(t), t \geq 0\}$  is a diffusion process with state space  $\Omega = [0, \infty)$  and with coefficients

$$\{m(x, t) = \alpha_N(t) + x\gamma(t), v(x, t) = \frac{1}{N_0}x\omega(t)\},$$

where  $\gamma(t) = b_I(t) - d_I(t)$ ,  $\omega(t) = b_I(t) + d_I(t)$  and  $\alpha_N(t)$  is the mutation rate from normal stem cells to initiated cells.

#### 1.4. State Space Models and Hidden Markov Models

To validate the stochastic models and to estimate unknown parameters in the model, one usually generates observed data from the system. Based on these data sets, statisticians have constructed statistical models to make inferences about the unknown parameters and to validate the model. To combine information from both the mechanism and the data, the state space model then combines the stochastic model and the statistical model into one model. Thus, the state space model has two sub-models:

- (1) The stochastic system model which is the stochastic model of the system, and
- (2) the observation model which is the statistical model based on some observed data from the system.

**Definition 1.8.** Let  $X(t)$  be a stochastic process with parameter space  $T$  and with state space  $S$ . Let  $\{Y(t_j) = Y_j, j = 1, \dots, n\}$  be the observed values on  $X(t)$  at the time points  $t_1 \leq t_2 \leq \dots \leq t_{n-1} \leq t_n$ . Suppose that  $Y_j = f[X(t), t \leq t_j] + e_j$  for some function  $f()$  of  $X(t), t \leq t_j$ , where  $e_j$  is the random measurement error for measuring  $Y_j$ . Then the combination  $\{X(t), t \in T; Y_j, j = 1, \dots, n\}$  is called a *state space model of the stochastic system* with stochastic system model given by the stochastic process  $\{X(t), t \in T\}$  and with the observation model given by the statistical model  $Y_j = f[X(t), t \leq t_j] + e_j$  for the system. In other word, a state space model of a stochastic system is the stochastic model of the system plus some statistical model based on some observed data from the system.

From this definition, it appears that if some data are available on the system, then one may always construct a state space model for the system. For this state space model, the stochastic process of the system is the stochastic

system model whereas the statistical model of the system is the observation model. As such, one may look at the state space model as a device to combine information from both sources: The mechanism of the system via stochastic models and the information from the data on the system. It is advantageous over both the stochastic model and the statistical model used alone as it combines advantages and information from both models.

The state space model was originally proposed by Kalman in the 60's for engineering control and communication [28]. Since then it has been successfully used in satellite research and military missile research. It has also been used by economists in econometric research [29] and by mathematician and statisticians in time series research [30] for solving many difficult problems which appear to be extremely difficult from other approaches. In 1995, the state space model was first proposed by Wu and Tan in AIDS research [31, 32]. Since then it has been used by Cazelles and Chau [33] and by Tan and his associates for modeling AIDS epidemic [34, 35]; it has also been used by Tan and his associates for studying the HIV pathogenesis in HIV-infected individuals [36–39]. Recently, Tan and his associates [40–42] have developed state space models for carcinogenesis. In Chaps. 8 and 9, we will illustrate and discuss these models and demonstrate some of its applications to cancer and AIDS.

**Definition 1.9.** A state space model is called a *hidden Markov model* if the stochastic system model is a Markov process.

Hidden Markov models usually apply to observed data on a Markov process because the observed data are usually masked by random measurement errors in measuring the observations. As such, it is appropriate to define hidden Markov models as above because the Markov process is hidden in the observed equations. In this section we will illustrate this by an example from the AIDS epidemiology. This example has been used by Satten and Longini [17] to estimate the transition rates in the San Francisco homosexual population.

**Example 1.19. The hidden Markov models of HIV epidemic as state space models.** Consider a population involving only HIV-infected individuals and AIDS cases. Following Satten and Longini [17], we partition the HIV-infected individuals into 6 sub-stages by the number of  $CD4^{(+)}$  T cells per  $mm^3$  of blood as given in Example 1.12. Let  $i$  stand for the  $I_i$  stage with  $I_6$  denoting the AIDS stage. Let  $X(t)$  denote the stochastic process

representing the infective stages with state space  $\Omega = \{1, \dots, 6\}$  and with parameter space  $T = \{0, 1, \dots, \infty\}$  with 0 denoting the starting time. Let  $\gamma_{ji}$  be the one-step transition probability from  $I_j$  to  $I_i$  ( $j, i = 1, \dots, 6$ ) and with  $\gamma_{6i} = \delta_{6i}, i = 1, \dots, 6$ . Then  $X(t)$  is a homogeneous Markov chain with 6 states and with discrete time. In this Markov chain, the state  $I_6$  is the absorbing state (persistent state) and all other states are transient states (For definition of persistent states and transient states, see, Definition 2.3.) Let  $Y_i(t)$  be the observed number of the  $I_i$  people at time  $t$ . Then, because the CD4<sup>(+)</sup> T cell counts are subjected to measurement errors, in terms of the observed numbers, the process is a hidden Markov chain. In this section, we proceed to show that this hidden Markov chain can be expressed as a state space model which consists of the stochastic system model and the observation model (For more detail about state space models, see Chaps. 8 and 9.) To this end, let  $W_{ij}(t)$  denote the number of  $I_j$  people at time  $t+1$  given  $I_i(t)$   $I_i$  people at time  $t$  for  $i = 1, \dots, 5; j = 1, \dots, 6$  and  $Z_{ij}(r, t)$  the observed number of  $I_r$  people at time  $t+1$  counted among the  $W_{ij}(t)$  people. Assume now that the death rate is very small for people other than AIDS and that there are no immigration and no migration in the population. Then, given  $I_i(t)$  for ( $i = 1, \dots, 5$ ), the probability distribution of  $W_{ij}(t), j = 1, \dots, 6$  follows a five-dimensional multinomial distribution with parameters  $\{I_i(t); \gamma_{ij}, j = 1, \dots, 5\}$ . That is, with  $W_{i6}(t) = I_i(t) - \sum_{j=1}^5 W_{ij}(t)$ , we have that, for  $i = 1, \dots, 5$ ,

$$\{W_{ij}(t), j = 1, \dots, 5\} | I_i(t) \sim ML\{I_i(t); \gamma_{ij}, j = 1, \dots, 5\}.$$

Note that  $\sum_{j=1}^6 \gamma_{ij} = 1$  for  $i = 1, \dots, 6$  and  $I_6 \rightarrow I_6$  only.

Let  $I_6(t)$  include people who died from AIDS during  $[t, t+1)$ . Then, we have the following stochastic equations for  $I_j(t), j = 1, \dots, 6$ :

$$\begin{aligned} I_j(t+1) &= \sum_{i=1}^5 W_{ij}(t) + \delta_{j6} I_6(t) \\ &= \sum_{i=1}^5 I_i(t) \gamma_{ij} + \delta_{j6} I_6(t) + \epsilon_j(t+1), \end{aligned} \quad (1.2)$$

where

$$\epsilon_j(t+1) = \sum_{i=1}^5 [W_{ij}(t) - I_i(t) \gamma_{ij}].$$

Denote by  $F' = (\gamma_{ij})$  the one-step transition matrix,  $\underline{I}(t) = \{I_1(t), \dots, I_6(t)\}'$  and  $\underline{\epsilon}(t+1) = \{\epsilon_1(t+1), \dots, \epsilon_6(t+1)\}'$ . Then in matrix notation, the above system of equations become:

$$\underline{I}(t+1) = F\underline{I}(t) + \underline{\epsilon}(t+1). \quad (1.3)$$

This is the stochastic system model for the state space model associated with the above hidden Markov chain.

To account for the random measurement error, we assume that the measurement errors follow Gaussian distributions and that measurement errors for AIDS cases is very small to be ignored. Let  $\nu_i$  ( $i = 1, \dots, 5$ ) denote the mean number of CD4<sup>(+)</sup> T cells per mm<sup>3</sup> of blood for the  $I_i$  stage. (One may take  $\nu_1 = 1000/\text{mm}^3$ ,  $\nu_2 = 800/\text{mm}^3$ ,  $\nu_3 = 600/\text{mm}^3$ ,  $\nu_4 = 425/\text{mm}^3$ ,  $\nu_5 = 275/\text{mm}^3$ .) Let  $X_i = \frac{Z - \nu_i}{100}$   $i = 1, \dots, 5$ , where  $Z$  is the observed number of CD4<sup>(+)</sup>. Then, given the  $I_i$  stage ( $i = 1, \dots, 5$ ), the conditional distribution of  $X_i$  is a truncated Gaussian with mean 0 and variance  $\sigma^2$  and with state space  $[-\frac{\nu_i}{100}, \frac{2000 - \nu_i}{100}]$  independently for  $i = 1, \dots, 5$ . For ( $i = 1, \dots, 5$ ), let

$$a_{i,0} = \frac{2000 - \nu_i}{100}, \quad a_{i,1} = \frac{900 - \nu_i}{100}, \quad a_{i,2} = \frac{700 - \nu_i}{100}, \quad a_{i,3} = \frac{500 - \nu_i}{100},$$

$$a_{i,4} = \frac{350 - \nu_i}{100}, \quad a_{i,5} = \frac{200 - \nu_i}{100}, \quad a_{i,6} = -\frac{\nu_i}{100}.$$

Denote, for ( $i = 1, \dots, 5; j = 1, \dots, 6$ ),

$$p_{ij} = C_i \int_{a_{i,j}}^{a_{i,j-1}} f(x) dx.$$

where  $f(x)$  is the pdf of the Gaussian distribution with mean 0 and variance  $\sigma^2$  and  $C_i^{-1} = \int_{a_{i,6}}^{a_{i,0}} f(x) dx$ .

Then for ( $i = 1, \dots, 5; j = 1, \dots, 5$ ), the conditional probability distribution of  $\{Z_{ij}(r, t), r = 1, \dots, 5\}$  given  $W_{ij}(t)$  is:

$$\{Z_{ij}(r, t), r = 1, \dots, 5\} | W_{ij}(t) \sim ML\{W_{ij}(t); \quad p_{jr}, r = 1, \dots, 5\}.$$

Note that  $\sum_{r=1}^6 p_{jr} = 1$  for  $j = 1, \dots, 5$ .

It follows that we have, with  $p_{6i} = \gamma_{6i} = \delta_{6i}$ :

$$\begin{aligned}
 Y_i(t+1) &= \sum_{u=1}^5 \left\{ \sum_{v=1}^5 Z_{uv}(i, t) + W_{u6} p_{6i} \right\} + \delta_{i6} I_6(t) \\
 &= \sum_{u=1}^5 \sum_{v=1}^6 W_{uv}(t) p_{vi} + \delta_{i6} I_6(t) + e_{i1}(t+1) \\
 &= \sum_{u=1}^5 I_u(t) \sum_{v=1}^6 \gamma_{uv} p_{vi} + \delta_{i6} I_6(t) + e_{i1}(t+1) + e_{i2}(t+1) \\
 &= \sum_{u=1}^6 I_u(t) \sum_{v=1}^6 \gamma_{uv} p_{vi} + e_i(t+1),
 \end{aligned}$$

where

$$e_{i1}(t+1) = \sum_{u=1}^5 \left\{ \sum_{v=1}^5 [Z_{uv}(i, t) - W_{uv}(t) p_{vi}] \right\},$$

and

$$e_{i2}(t+1) = \sum_{u=1}^5 \left\{ \sum_{v=1}^6 p_{vi} [W_{uv}(t) - I_u(t) \gamma_{uv}] \right\},$$

and  $e_i(t+1) = e_{i1}(t+1) + e_{i2}(t+1)$ .

Put  $P = (p_{ij})$  and  $H = P'F$ . Denote by  $\tilde{Y}(t) = \{Y_1(t), \dots, Y_6(t)\}'$  and  $\tilde{\varepsilon}(t) = \{e_1(t), \dots, e_6(t)\}'$ . Then, in matrix notation, we have:

$$\tilde{Y}(t+1) = H \tilde{I}(t) + \tilde{\varepsilon}(t+1). \quad (1.4)$$

Equation (1.4) is the observation model for the state space model associated with the above hidden Markov chain.

## 1.5. The Scope of the Book

The stochastic models described in Secs. 1.1–1.4 are the major models which arise from genetics, cancer and AIDS. In this book we will thus present a systematic treatment of these models and illustrate its applications to genetics, cancer and AIDS.