2

Statistical Models

In this chapter, we describe probability distributions, which provide fundamental tools for statistical models, and show that conditional distributions are used to acquire various types of information in the model-building process. By using regression and time series models as specific examples, we also discuss why evaluation of statistical models is necessary.

2.1 Modeling of Probabilistic Events and Statistical Models

Before considering statistical models, let us first discuss how to represent events that we know occur in a deterministic way. In the simple case in which an event is fixed and invariable, the state of the event can be expressed in the form $x = a$. In general, however, x varies depending on some factor. If x is dependent on an external factor u, then it can be expressed as a function of u, e.g., $x = h(u)$. In some cases, x is determined according to past events or based on the present state, in which case x can be expressed as some function of the factor.

Most real-life events, however, contain uncertainty, and in many cases our information about external factors is incomplete. In such cases, the value of x cannot be specified as a fixed value or a deterministic function of factors, and in such cases we use a probability distribution.

Given a random variable X defined on the sample space Ω, for any real value $x(\in \mathrm{R})$, the probability $\Pr(\{\omega \in \Omega\,; X(\omega) \leq x\})$ of an event such that $X(\omega) \leq x$ can be determined. If we regard such a probability as a function of x and express it as

$$\begin{aligned} G(x) &= \Pr(\{\omega \in \Omega\,; X(\omega) \leq x\}) \\ &= \Pr(X \leq x)\,, \end{aligned} \tag{2.1}$$

then the function $G(x)$ is referred to as the *distribution function* of X. By determining the distribution function $G(x)$, we can characterize the random variable X. In particular, if there exists a nonnegative function $g(t) \geq 0$ that satisfies

$$G(x) = \int_{-\infty}^{x} g(t)dt, \tag{2.2}$$

then X is said to be *continuous*, and the function $g(t)$ is called a *probability density function*. A continuous probability distribution can be defined by determining the density function $g(t)$.

On the other hand, if the random variable X takes either a finite or a countably infinite number of discrete values $x_1, x_2, \ldots$, then the variable X is said to be *discrete*. The probability of taking a discrete point $X = x_i$ is determined by

$$\begin{aligned} g_i = g(x_i) &= \Pr(\{\omega \in \Omega\,; X(\omega) = x_i\}) \\ &= \Pr(X = x_i), \qquad\qquad i = 1, 2, \ldots, \end{aligned} \tag{2.3}$$

where $g(x)$ is called a *probability function*, for which the distribution function is given by $G(x) = \sum_{\{i;x_i \leq x\}} g(x_i)$, where $\sum_{\{i;x_i \leq x\}}$ represents the sum of the discrete values such that $x_i \leq x$.

If we assume that the observations $\boldsymbol{x}_n = \{x_1, x_2, \ldots, x_n\}$ are generated from the distribution function $G(x)$, then $G(x)$ is referred to as the *true distribution*, or the *true model*. On the other hand, the distribution function $F(x)$ used to approximate the true distribution is referred to as a *model* and is assumed to have either a density function or a probability function $f(x)$. If a model is specified by p-dimensional parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$, then the model can be written as $f(x|\boldsymbol{\theta})$. If the parameters are represented as a point in the set $\Theta \subset \mathrm{R}^p$, then $\{f(x|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$ is called a *parametric family of probability distributions or models*.

An estimated model $f(x|\hat{\boldsymbol{\theta}})$ obtained by replacing an unknown parameter $\boldsymbol{\theta}$ with an estimator $\hat{\boldsymbol{\theta}}$ is referred to as a *statistical model*. The process of constructing a model that appropriately represents some phenomenon is referred to as *modeling*. In statistical modeling, it is necessary to estimate unknown parameters. However, settng up an appropriate family of probability models prior to estimating the parameters is of greater importance.

We first describe some probability distributions as fundamental models. After that, we will show that the mechanism of incorporating information from other variables can be represented in the form of a conditional distribution model.

2.2 Probability Distribution Models

The most fundamental form of a model is the probability distribution model or the probability model. More sophisticated models, such as conditional

distribution models described in the next section, are also constructed using the probability distribution model.

Example 1 (Normal distribution model) The most widely used continuous probability distribution model is the normal distribution model, or Gaussian distribution model. The probability density function for the normal distribution is given by

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < x < \infty\,. \tag{2.4}$$

This distribution is completely specified by the two parameters μ and σ^2, which are the mean and the variance, respectively. A probability distribution model, such as the normal distribution model, that can be expressed in a specific functional form containing a finite number of parameters $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ is called a *parametric probability distribution model*.

In addition to the normal distribution model, the following parametric probability distribution models are well known:

Example 2 (Cauchy distribution model) If the probability density function is given by

$$f(x|\mu,\tau^2) = \frac{1}{\pi}\frac{\tau}{(x-\mu)^2+\tau^2}\,, \quad -\infty < x < \infty\,, \tag{2.5}$$

then the distribution is called a *Cauchy distribution*. The parameters μ and τ^2 define the center of the distribution and the spread of the distribution, respectively. While the Cauchy distribution is symmetric with respect to the mode at μ, its mean and variance are not well-defined.

Example 3 (Laplace distribution model) A random variable X is said to have a Laplace distribution if its probability density function is

$$f(x|\mu,\tau) = \frac{1}{2\tau}\exp\left(-\frac{|x-\mu|}{\tau}\right), \qquad -\infty < x < \infty, \tag{2.6}$$

where $-\infty < \mu < \infty$ and $\tau > 0$. The mean and variance are respectively given by $E[X] = \mu$ and $V(X) = 2\tau^2$. The distribution function of the Laplace random variable is

$$F(x|\mu,\tau) = \begin{cases} \dfrac{1}{2}\exp\left(\dfrac{x-\mu}{\tau}\right), & x \le \mu, \\ 1 - \dfrac{1}{2}\exp\left(-\dfrac{x-\mu}{\tau}\right), & x > \mu. \end{cases} \tag{2.7}$$

Example 4 (Pearson's family of distributions model) If the probability density function is given by

$$f(x|\mu,\tau^2,b)=\frac{\Gamma(b)\tau^{2b-1}}{\Gamma(b-\frac{1}{2})\Gamma(\frac{1}{2})}\frac{1}{\{(x-\mu)^2+\tau^2\}^b},\quad -\infty<x<\infty\,, \tag{2.8}$$

then the distribution is known as a *Pearson's family of distributions*, in which the quantities μ and τ^2 are referred to as the *center* and *dispersion parameters*, as in the case of the Cauchy distribution. The quantity b is a parameter that specifies the shape of the distribution. By varying the value of b, it is possible to represent a variety of distributions. When $b=1$, the distribution is Cauchy, and when $b=(k+1)/2$ where k is an integer, the distribution is a t-distribution with k degrees of freedom. Also, the distribution becomes a normal distribution when $b\to\infty$.

Example 5 (Mixture of normal distributions model) If the density function can be represented by

$$f(x|m,\boldsymbol{\theta})=\sum_{j=1}^{m}\alpha_j\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left\{-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right\},\quad -\infty<x<\infty, \tag{2.9}$$

then the distribution is called a *mixture of normal distributions*, where $\boldsymbol{\theta}=(\mu_1,\ldots,\mu_m,\sigma_1^2,\ldots,\sigma_m^2,\alpha_1,\ldots,\alpha_{m-1})^T$ and $\sum_{j=1}^{m}\alpha_j=1$. A mixture of normal distributions is constructed by combining m normal distributions with weights α_j, in which case m is referred to as the number of components. A wide range of probability distribution models can be expressed by appropriate selection of the parameters m, α_j, μ_j, and σ_j^2.

Figure 2.1 shows various examples of probability distribution models. The model in the upper left panel is the standard normal distribution model with mean 0 and variance 1. The model in the upper right panel is a Cauchy distribution model with $\mu=0$ and $\tau^2=1$. One feature of this model is that it has fatter left and right tails. By using a Cauchy distribution rather than a normal distribution, it is possible to model a phenomenon in which large absolute values have small but nonnegligible probabilities. This property can be used to detect outliers, perform a robust estimation, or detect jumps in a trend. The lower left panel shows Pearson distributions with $b=0.6,0.75,1,1.5$, and 3. By varying the value of b, it is possible to continuously represent various distributions, ranging from distributions that have even fatter tails than the Cauchy distribution to the normal distribution. The lower right panel shows an example of a mixture of normal distributions, which is capable of representing complex distributions even in the simplest case when $m=2$.

Example 6 (Binomial distribution model) Let X be a binary random variable taking the values of either 0 or 1, and let the probability of an event's occurring be given by

$$\Pr(X=1)=p,\quad \Pr(X=0)=1-p,\qquad (0<p<1)\,. \tag{2.10}$$

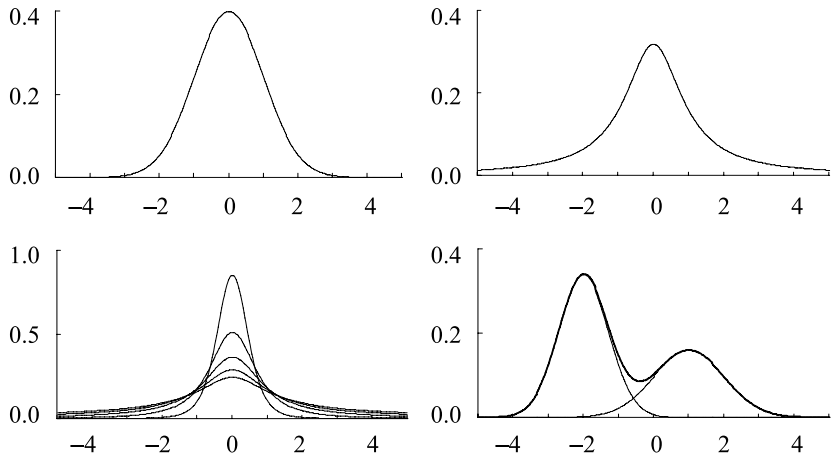

Fig. 2.1. Various examples of probability distributions: standard normal distribution (upper left); Cauchy distribution with $m = 0$ and $\tau^2 = 1$ (upper right); Pearson distributions with $b = 0.6, 0.75, 1, 1.5$, and 3 (lower left); and a mixture of normal distributions (lower right).

This probability distribution is referred to as a *Bernoulli distribution*, and its probability function is given by

$$f(x|p) = p^x(1-p)^{1-x}, \qquad x = 0, 1\,. \tag{2.11}$$

We further assume that the sequence of random variables $X_1, X_2, \ldots, X_n$ is independently distributed having the same Bernoulli distribution. Then the random variable $X = X_1 + X_2 + \cdots + X_n$ denotes the number of occurrences of an event in n trials, and its probability function is given by

$$f(x|p) = {}_nC_x p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n\,. \tag{2.12}$$

Such a probability distribution is called a *binomial distribution* with parameters n and p. The mean and variance are $E[X] = np$ and $V(X) = np(1-p)$, respectively.

Example 7 (Poisson distribution model) When very rare events are observed in short intervals, the distribution of the number of events is given by

$$f(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad x = 0, 1, 2, \ldots \qquad (0 < \lambda < \infty)\,. \tag{2.13}$$

This distribution is called a *Poisson distribution*. The mean and variance are $E[X] = \lambda$ and $V(X) = \lambda$. The Poisson distribution is derived as an approximation to the binomial distribution by writing $np = \lambda$ for the probability

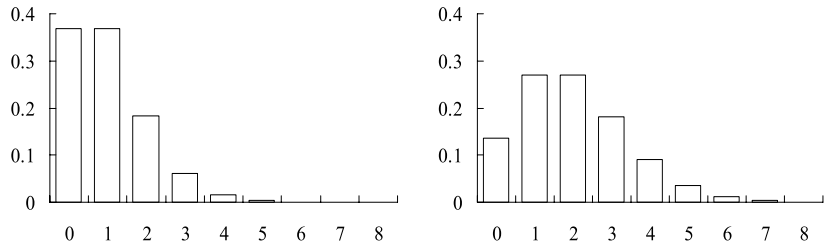

Fig. 2.2. Poisson distributions: left: $\lambda = 1$; right: $\lambda = 2$.

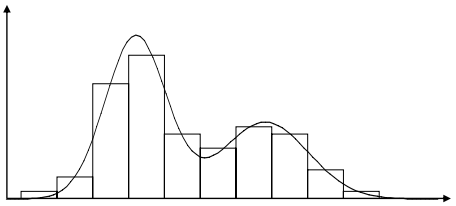

Fig. 2.3. A continuous distribution model and its approximation by a histogram.

function of the binomial distribution, while keeping λ constant. In fact, if n tends to infinity and p approaches 0, then for a fixed integer x,

$$ {}_nC_x p^x (1-p)^{n-x} = \frac{n!}{(n-x)!}\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-x} \quad \rightarrow \quad \frac{\lambda^x}{x!}e^{-\lambda}. \tag{2.14} $$

Figure 2.2 shows Poisson distributions for the cases when the parameter λ is 1 and 2. Discrete distributions of various shapes can be represented depending on the value of λ.

Example 8 (Histogram model) A histogram can be obtained by dividing the domain $x_{\min} \leq X \leq x_{\max}$ of the random variable into appropriate intervals $B_1, \ldots, B_k$, determining the frequencies $n_1, \ldots, n_k$ of the observations that fall in the intervals $B_j = \{x; x_{j-1} \leq x < x_j\}$, and graphing the results. If we set $n = n_1 + \cdots + n_k$, and define the relative frequency as $f_j = n_j/n$, a histogram can be thought of as defining the discrete distribution model $f = \{f_1, \ldots, f_k\}$ that is obtained by converting a continuous variable into a discrete variable. On the other hand, if the histogram is thought of as approximating a density function with a stepwise function, the histogram itself can be regarded as a type of continuous distribution model (Figure 2.3).

Example 9 (Probability model) A wide variety of phenomena can be expressed in terms of probability distributions according to the underlying

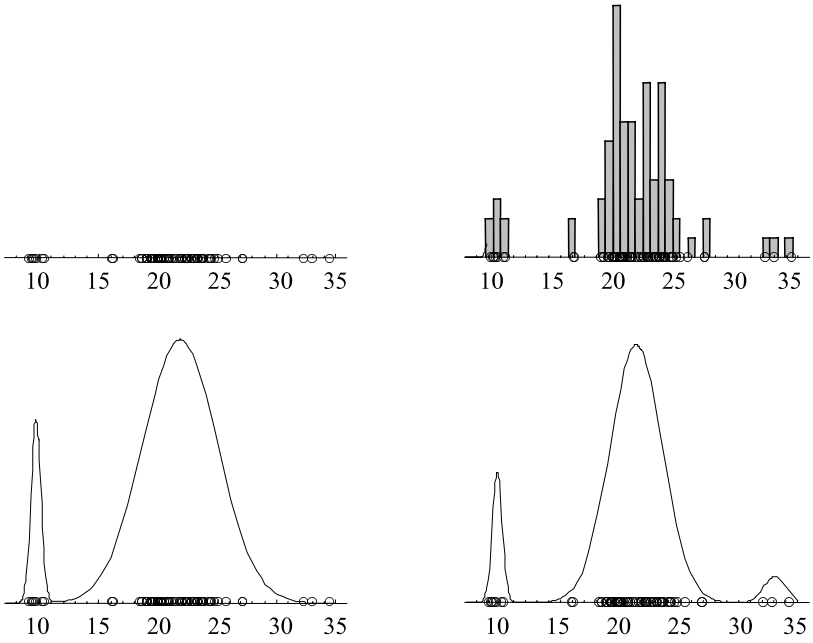

Fig. 2.4. The distribution of the velocities of 82 galaxies [Roeder (1990)]. Data (top left), the histogram (top right), and a mixture of normal distributions model (bottom left: $m = 2$; bottom right: $m = 3$).

problem. The problem is how to construct a probability model based on observed data.

Figure 2.4 shows the observed velocities, x, of 82 galaxies [Roeder (1990)]. Let us approximate the distribution of galaxy velocities using the mixture of normal distributions model in (2.9). If we estimate the parameters for the mixture of normal distributions based on observed data and replace the unknown parameters with estimated values, then the resulting density function $f(x|m, \hat{\boldsymbol{\theta}})$ is a statistical model. A critical issue in fitting the mixture of normal distributions model is the selection of the number of components, m. A two-component model has five parameters, while a three-component model has eight parameters. We must determine which model among the various candidate models best describes the probabilistic structure of the random variable X. Essential to answering this question is the criteria for evaluating the goodness of a statistical model.

Thus far, we have considered univariate random variables. There are many real-world situations, however, in which several variables must be considered simultaneously, for example, temperature and pressure in meteorological

data, or interest rate and GDP in economic data. In such cases, $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ becomes a multivariate random vector, for which the distribution function is defined as a function of p variables that are given in terms of $\boldsymbol{x} = (x_1, \ldots, x_p)^T \in \mathrm{R}^p$,

$$\begin{aligned} G(x_1, \ldots, x_p) &= \Pr(\{\omega \in \Omega : X_1(\omega) \leq x_1, \ldots, X_p(\omega) \leq x_p\}) \\ &= \Pr(X_1 \leq x_1, \ldots, X_p \leq x_p)\,. \end{aligned} \tag{2.15}$$

In parallel with the univariate case, a density function for the multivariate distribution can be defined. For a continuous distribution, a nonnegative function $f(x_1, \ldots, x_p) \geq 0$ that satisfies

$$\begin{aligned} &\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_p) dx_1 \cdots dx_p = 1, \\ &G(x_1, \cdots, x_p) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_p} f(t_1, \ldots, t_p) dt_1 \cdots dt_p \end{aligned} \tag{2.16}$$

is called the *probability density function* of the multivariate random vector $\boldsymbol{X}$.

Consider a discrete case, in which a p-dimensional random vector $\boldsymbol{X} = (X_1, \cdots, X_p)^T$ assumes either a finite or a countably infinite number of discrete values $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$, where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$, $i = 1, 2, \ldots$. Then the probability function of the random vector $\boldsymbol{X}$ is defined by

$$g(\boldsymbol{x}_i) = \mathrm{P_r}(X_1 = x_{i1}, \ldots, X_p = x_{ip}), \quad i = 1, 2, \ldots. \tag{2.17}$$

The probability function satisfies

$$g(\boldsymbol{x}_i) \geq 0, \quad i = 1, 2, \ldots, \quad \text{and} \quad \sum_{i=1}^{\infty} g(\boldsymbol{x}_i) = 1, \tag{2.18}$$

and the distribution function can be expressed as

$$G(x_1, \cdots, x_p) = \sum_{\{i; x_{i1} \leq x_1\}} \cdots \sum_{\{i; x_{ip} \leq x_p\}} g(x_{i1}, \ldots, x_{ip}). \tag{2.19}$$

Example 10 (Multivariate normal distribution) A p-dimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ is said to have a p-variate normal distribution with mean vector $\boldsymbol{\mu}$ and variance covariance matrix Σ if its probability density function is given by

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right\}, \tag{2.20}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$ and Σ is a $p \times p$ symmetric positive definite matrix whose $(i, j)^{th}$ component is given by σ_{ij}. We write $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$.

Example 11 (Multinomial distribution) Suppose that there exist $k+1$ possible outcomes $E_1, \ldots, E_{k+1}$ in a trial. Let $\mathrm{P}(E_i) = p_i$, where $\sum_{i=1}^{k+1} p_i = 1$, and let X_i $(i = 1, \ldots, k+1)$ denote the number of times outcome E_i occurs in n trials, where $\sum_{i=1}^{k+1} X_i = n$. If the trials are repeated independently, then a multinomial distribution with parameters n, $p_1, \ldots, p_k$ is defined as a discrete distribution having the probability function

$$\Pr(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{\prod_{i=1}^{k+1} x_i!} \prod_{i=1}^{k+1} p_i^{x_i}, \tag{2.21}$$

where $x_i = 0, 1, \ldots, n$ (note that $x_{k+1} = n - \sum_{i=1}^{k} x_i$). The mean, variance, and covariance are respectively given by $E[X_i] = np_i$, $i = 1, \ldots, k$, $V(X_i) = np_i(1 - p_i)$, and $\mathrm{Cov}(X_i, X_j) = -np_ip_j$ $(i \neq j)$.

2.3 Conditional Distribution Models

From the viewpoint of statistical modeling, the probability distribution is the most fundamental model in the situation in which the distribution of the random variable X is independent of various other factors. In practice, however, information associated with these variables can be used in various ways. The essence of statistical modeling lies in finding such information and incorporating it into a model in an appropriate form. In the following, we consider cases in which a random variable depends on other variables, on past history, on a spatial pattern, or on prior information. The important thing is that such modeling approaches can be considered as essentially estimating conditional distributions. Thus, the essence of statistical modeling can be thought of as obtaining an appropriate conditional distribution.

In general, if the distribution of the random variable Y is determined in a manner that depends on a p-dimensional variable $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^T$, then the distribution of Y is expressed as $F(y|\boldsymbol{x})$ or $f(y|\boldsymbol{x})$, and this is called a *conditional distribution model*. There are several ways in which the random variable depends on the other variables $\boldsymbol{x}$. In the following, we consider typical conditional distribution models.

2.3.1 Regression Models

The regression model is used to model the relationship between a response variable y and several explanatory variables $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^T$. This is equivalent to assuming that the probability distribution of the response variable y varies depending on the explanatory variables $\boldsymbol{x}$ and that a conditional distribution is given in the form of $f(y|\boldsymbol{x})$.

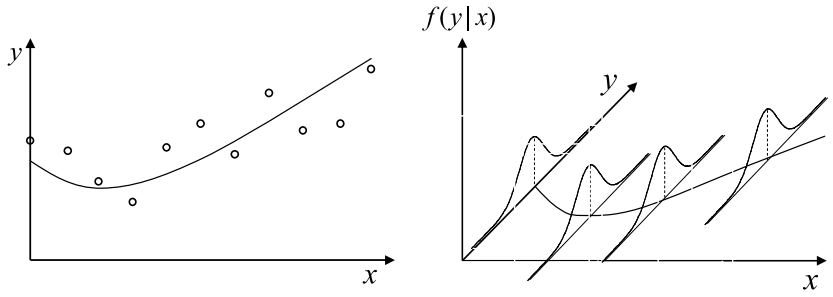

Fig. 2.5. Regression model (left) and conditional distribution model (right) in which the mean of the response variable varies as a function of the explanatory variable x.

Let $\{(y_\alpha, \boldsymbol{x}_\alpha);\ \alpha = 1, 2, \ldots, n\}$ be n sets of data obtained in terms of the response variable y and p explanatory variables $\boldsymbol{x}$. Then the model

$$y_\alpha = u(\boldsymbol{x}_\alpha) + \varepsilon_\alpha, \qquad \alpha = 1, 2, \ldots, n\,, \tag{2.22}$$

of the observed data is called a *regression model*, where $u(\boldsymbol{x})$ is a function of the explanatory variables $\boldsymbol{x}$, and the error terms or noise ε_α are assumed to be independently distributed with mean $E[\varepsilon_\alpha] = 0$ and variance $V(\varepsilon_\alpha) = \sigma^2$. We often assume that the noise ε_α follows the normal distribution $N(0, \sigma^2)$. In such a case, y_α has the normal distribution $N(u(\boldsymbol{x}_\alpha), \sigma^2)$ with mean $u(\boldsymbol{x}_\alpha)$ and variance σ^2, and its density function is given by

$$f(y_\alpha|\boldsymbol{x}_\alpha) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_\alpha - u(\boldsymbol{x}_\alpha))^2}{2\sigma^2}\right\}, \quad \alpha = 1, 2, \ldots, n\,. \tag{2.23}$$

This distribution is a type of conditional distribution model in which the mean varies according to $E[Y|\boldsymbol{x}] = u(\boldsymbol{x})$ in a manner that depends on the values of the explanatory variables $\boldsymbol{x}$.

The left panel in Figure 2.5 shows 11 observations and the mean function $u(x)$ of the one-dimensional explanatory variable x and the response variable y. The data y_α at a given point x_α are observed as

$$y_\alpha = \mu_\alpha + \varepsilon_\alpha, \qquad \alpha = 1, 2, \ldots, 11\,, \tag{2.24}$$

with true mean value $E[Y_\alpha|x_\alpha] = \mu_\alpha$ and noise ε_α. The quantity $u(x)$ represents the mean structure of the event, and ε_α is the noise that induces probabilistic fluctuations in the data y_α. The right panel in Figure 2.5 shows a conditional distribution determined using a regression model. Fixing the value of the explanatory variable x gives the probability distribution $f(y|x)$, for which the mean is $u(x)$. Therefore, the regression model in (2.23) determines a class of distributions that move in parallel with the value of x.

Example 12 (Linear regression model) If the regression function or the mean function $u(\boldsymbol{x})$ can be approximated by a linear function of $\boldsymbol{x}$, then the model in (2.22) can be expressed as

$$\begin{aligned} y_\alpha &= \beta_0 + \beta_1 x_{\alpha 1} + \cdots + \beta_p x_{\alpha p} + \varepsilon_\alpha \\ &= \boldsymbol{x}_\alpha^T \boldsymbol{\beta} + \varepsilon_\alpha, \qquad\qquad \alpha = 1, 2, \ldots, n\,, \end{aligned} \tag{2.25}$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$, $\boldsymbol{x}_\alpha = (1, x_{\alpha 1}, x_{\alpha 2}, \ldots, x_{\alpha p})^T$ and is referred to as a *linear regression model*. A linear regression model with Gaussian noise can be expressed by the density function

$$f(y_\alpha | \boldsymbol{x}_\alpha; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_\alpha - \boldsymbol{x}'_\alpha \boldsymbol{\beta})^2}{2\sigma^2} \right\}, \quad \alpha = 1, 2, \ldots, n\,, \tag{2.26}$$

where the unknown parameters in the model are $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2)^T$. In the linear regression model, the critical issue is to determine a set of explanatory variables that appropriately describes changes in the distribution of the response variable y; this problem is referred to as the *variable selection* problem.

Example 13 (Polynomial regression model) A polynomial regression model with Gaussian noise,

$$y_\alpha = \beta_0 + \beta_1 x_\alpha + \cdots + \beta_m x_\alpha^m + \varepsilon_\alpha, \quad \varepsilon_\alpha \sim N(0, \sigma^2), \tag{2.27}$$

assumes that the regression function $u(x)$ can be approximated by $\beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m$ with respect to the one-dimensional explanatory variable x. For each order m, the parameters of the polynomial regression model are $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)^T$ and the error variance is σ^2. In a polynomial regression model, the crucial task is determining the order m, which is referred to as the *order selection* problem. As shown in Example 16, a model having an order that is too low cannot adequately represent the data structure. On the other hand, a model with an order that is too high causes the model to react excessively to random variations in the data, masking the essential relationship.

Various functions in addition to polynomials are used to represent a regression function. Trigonometric function models are expressed as

$$y_\alpha = a_0 + \sum_{j=1}^{m} \{ a_j \cos(j\omega x_\alpha) + b_j \sin(j\omega x_\alpha) \} + \varepsilon_\alpha\,. \tag{2.28}$$

In addition, various forms of other orthogonal functions can be used to approximate the regression function.

Example 14 (Nonlinear regression models) Thus far, given a regression function $E[Y|\boldsymbol{x}] = u(\boldsymbol{x})$, we have constructed models by assuming functional

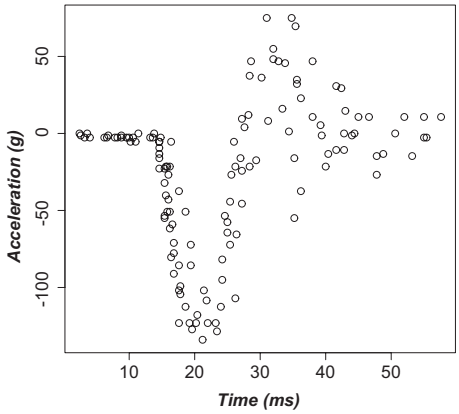

Fig. 2.6. Motorcycle impact data.

forms such as polynomials. The analysis of complex and diverse phenomena, however, requires developing more flexible models. Figure 2.6, for example, plots the measured acceleration $Y(g$; gravity) of the crash dummy's head at a time X (ms, millisecond) from the moment of collision in repeated motorcycle collision experiments [Härdle (1990)]. Neither polynomial models nor models using specific nonlinear functions are adequate for describing the structure of phenomena characterized by data that exhibit this type of complex nonlinear structure.

It is assumed that at each point x_α, y_α is observed as $y_\alpha = \mu_\alpha + \varepsilon_\alpha$, $\alpha = 1, 2, \ldots, n$, with noise ε_α. In order to approximate μ_α, $\alpha = 1, 2, \ldots, n$, in a way that reflects the structure of the phenomenon, we use a regression model

$$y_\alpha = u(x_\alpha; \boldsymbol{\theta}) + \varepsilon_\alpha, \qquad \alpha = 1, 2, \ldots, n\,. \tag{2.29}$$

For $u(x; \boldsymbol{\theta})$, various models are used depending on the analysis objective, including (1) splines [Green and Silverman (1994)], (2) B-splines [de Boor (1978), Imoto (2001)], (3) kernel functions [Simonoff (1996)], and (4) multilayer neural network models [Bishop (1995), Ripley (1996)]. Our purpose here is to identify the mean structure of a phenomenon from data based on these flexible models.

Example 15 (Changing variance model) Whereas in the regression models described above, only the mean structure changes as a function of the explanatory variables $\boldsymbol{x}$, in changing variance models the variance of the response variable y also changes as a function of $\boldsymbol{x}$, and such a change is expressed in the form $\sigma^2(\boldsymbol{x})$. In this case, the conditional distribution of y is given by $N(u(\boldsymbol{x}), \sigma^2(\boldsymbol{x}))$. Figure 2.7 shows an example of a conditional distribution determined by a changing variance model in which it has a constant mean. It

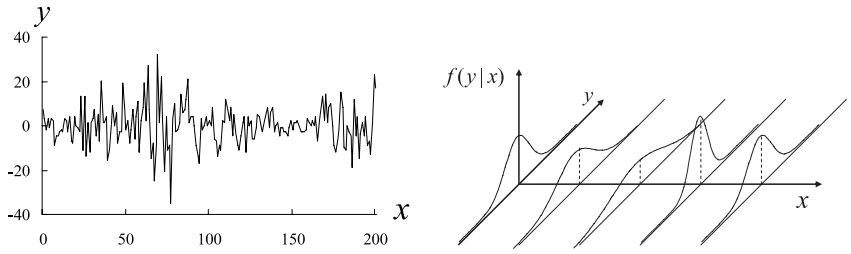

Fig. 2.7. Conditional distributions of changing variance models.

shows that the variance of the distribution changes depending on the value of x. These types of changing variance models are important for analyzing earthquake data and financial data.

Generally, a regression model is composed of a model that approximates the mean function $E[Y|\boldsymbol{x}]$ representing the structure of phenomenon and a probability distribution model that describes the probabilistic fluctuation of the data. Since models that approximate the mean function depend on several parameters, we write $u(\boldsymbol{x};\boldsymbol{\beta})$. Observed data with Gaussian noise are then given as

$$y_\alpha = u(\boldsymbol{x}_\alpha;\boldsymbol{\beta}) + \varepsilon_\alpha, \qquad \alpha = 1, 2, \ldots, n\,, \tag{2.30}$$

and are represented by the density function

$$f(y_\alpha|\boldsymbol{x}_\alpha;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_\alpha - u(\boldsymbol{x}_\alpha;\boldsymbol{\beta}))^2}{2\sigma^2}\right\}, \quad \alpha = 1, 2, \ldots, n\,, \tag{2.31}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2)^T$.

In the case of a regression model expressed by a density function, we estimate the parameter vector $\boldsymbol{\theta}$ of the model by using the maximum likelihood method, and we denote it as $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^T, \hat{\sigma}^2)^T$. Then the density function in which the unknown parameters in (2.31) are replaced with their corresponding estimators,

$$f(y_\alpha|\boldsymbol{x}_\alpha;\hat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left\{-\frac{(y_\alpha - u(\boldsymbol{x}_\alpha;\hat{\boldsymbol{\beta}}))^2}{2\hat{\sigma}^2}\right\}, \quad \alpha = 1, 2, \ldots, n\,, \tag{2.32}$$

is called a *statistical model*.

Although the main focus in regression models tends to be modeling for expected values, the distributions of error terms are also important. For a given regression function, different models can be obtained by changing the value of the variance. In addition, models that assume distributions other than

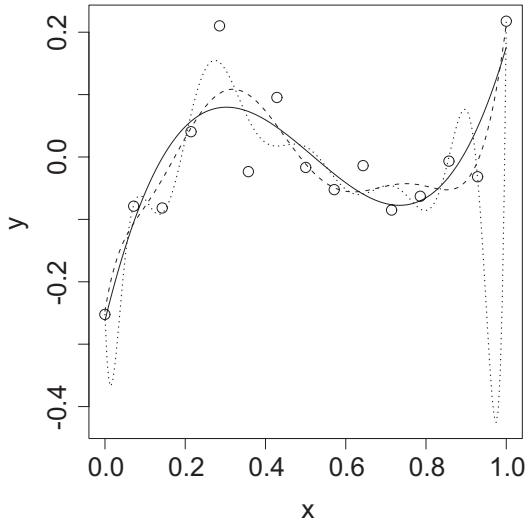

Fig. 2.8. Fitting polynomial regression models of order 3 (solid), 8 (broken), and 12 (dotted).

the normal distribution for the error terms (e.g., Cauchy distribution) are also conceivable.

Example 16 (Fitting a polynomial regression model) Figure 2.8 shows a plot of 15 observations obtained with respect to the explanatory variable x and the response variable y. By ordering the data as $\{(x_\alpha, y_\alpha);\ \alpha = 1, 2, \ldots, 15\}$, we fit the polynomial regression model in (2.27).

For each order m, we estimate the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)^T$ of the polynomial regression model by using either the least square method or the maximum likelihood method that maximizes the log-likelihood function

$$\sum_{\alpha=1}^{n} \log f(y_\alpha | x_\alpha; \boldsymbol{\beta}, \sigma^2) \tag{2.33}$$
$$= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{\alpha=1}^{n}\{y_\alpha - (\beta_0 + \beta_1 x_\alpha + \cdots + \beta_m x_\alpha^m)\}^2$$

and denote the results as $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m)^T$. The figure shows the estimated polynomial regression curves for orders 3, 8, and 12; it shows that estimated polynomials can vary greatly depending on the assumed order. Thus, the problem is deciding the order of the polynomial that should be adopted in the model.

If we consider the problem of order selection from the viewpoint of the goodness of fit of data in an estimated model, that is, from the standpoint of

minimizing the squared sum of residuals

$$\sum_{\alpha=1}^{n} (y_\alpha - \hat{y}_\alpha)^2 = \sum_{\alpha=1}^{n} \left\{ y_\alpha - \left(\hat{\beta}_0 + \hat{\beta}_1 x_\alpha + \cdots + \hat{\beta}_m x_\alpha^m \right) \right\}^2 , \quad (2.34)$$

then the higher the order of the model, the smaller the value will be. As a result, we select the highest order [i.e., the $(n-1)^{th}$ order] polynomial that passes through all data points. If the data are free of errors, the error term ε_α in (2.27) will be superfluous, in which case it is sufficient to select the most complex model out of the class of models expressed by a large number of parameters. However, for data that contain intrinsic or observational errors, models that overfit the observed data tend to model the errors excessively and do not adequately approximate the true structure of the phenomenon. Consequently, such models do not predict future events well.

In general, a model that is too complex overadjusts for the random fluctuation in the data, while, on the other hand, overly simplistic models fail to adequately describe the structure of the phenomenon being modeled. Therefore, the key to evaluating a model is to strike a balance between, badness of fit of the data and the model complexity.

Example 17 (Spline functions) Assume that in the data $\{(y_\alpha, x_\alpha); \alpha = 1, 2, \ldots, n\}$ observed with respect to a response variable y and an explanatory variable x, n observations, $x_1, x_2, \ldots, x_n$, are ordered in ascending order in the interval $[a, b]$ as follows:

$$a < x_1 < x_2 < \cdots < x_n < b. \quad (2.35)$$

The essential idea in spline function fitting is to divide the interval containing the data $\{x_1, \ldots, x_n\}$ into several subintervals and to fit a polynomial model in a segment-by-segment manner, rather than fitting a single polynomial model to n sets of observed data.

Let $\xi_1 < \xi_2 < \cdots < \xi_m$ denote the m points that divide (x_1, x_n). These points are referred to as *knots*. A commonly used spline function in practical applications is the cubic spline, in which a third-order polynomial is fitted segment by segment over the subintervals $[a, \xi_1]$, $[\xi_1, \xi_2], \ldots, [\xi_m, b]$, and the polynomials are smoothly connected at the knots. In other words, the model is fitted under the restriction that at each knot, the first and second derivatives of the third-order polynomial are continuous. As a result, the cubic spline function having the knots $\xi_1 < \xi_2 < \cdots < \xi_m$ is given by

$$u(x; \boldsymbol{\theta}) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{i=1}^{m} \theta_i (x - \xi_i)_+^3 , \quad (2.36)$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m, \beta_0, \beta_1, \beta_2, \beta_3)^T$ and $(x - \xi_i)_+ = \max\{0, x - \xi_i\}$.

It is commonly known, however, that it is not appropriate to fit a cubic polynomial near a boundary since the estimated curve will vary excessively. In

order to address this difficulty, the natural cubic spline specifies that the cubic spline be a linear function at the two ends of the interval $(-\infty, \xi_1]$, $[\xi_m, +\infty)$, so that the natural cubic spline is given by

$$u(x;\boldsymbol{\theta}) = \beta_0 + \beta_1 x + \sum_{i=1}^{m-2} \theta_i \left\{ d_i(x) - d_{m-1}(x) \right\} , \tag{2.37}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_{m-2}, \beta_0, \beta_1)^T$ and

$$d_i(x) = \frac{(x-\xi_i)_+^3 - (x-\xi_m)_+^3}{\xi_m - \xi_i} .$$

When applying a spline in practical situations, we still need to determine the number of knots and their positions. From a computational standpoint, it is difficult to estimate the positions of knots as parameters. For this reason, we estimate the parameters $\boldsymbol{\theta}$ of the model by using the maximum penalized likelihood method described in Subsection 5.2.4 or the penalized least squares method discussed in Section 6.5. These topics are covered in Chapters 5 and 6. In the B-spline, a basis function is constructed by connecting the segment-wise polynomials, and it can substantially reduce the number of parameters in a model. This topic will be discussed in Section 6.2.

2.3.2 Time Series Model

Observed data, $x_1, \ldots, x_N$, for events that vary with time are referred to as a *time series*. The vast majority of real-world data, including meteorological data, environmental data, financial or economic data, and time-dependent experimental data, constitutes time series. The main aim of time series analysis is to identify the structure of the phenomenon represented by a sequence of measurements and to predict future observations. To analyze such time series data, we consider the conditional distribution

$$f(x_n | x_{n-1}, x_{n-2}, \ldots), \tag{2.38}$$

given observations up to the time $n-1$.

Example 18 (AR model and ARMA model) In particular, by assuming a linear structure in finite dimensions, we obtain an *autoregressive* (AR) *model* [Akaike (1969, 1970), Brockwell and Davis (1991)];

$$x_n = \sum_{j=1}^{p} a_j x_{n-j} + \varepsilon_n, \qquad \varepsilon_n \sim N(0, \sigma^2) , \tag{2.39}$$

where p denotes the order and indicates which information, obtained up to what time in the past, must be used in order to determine a future predictive

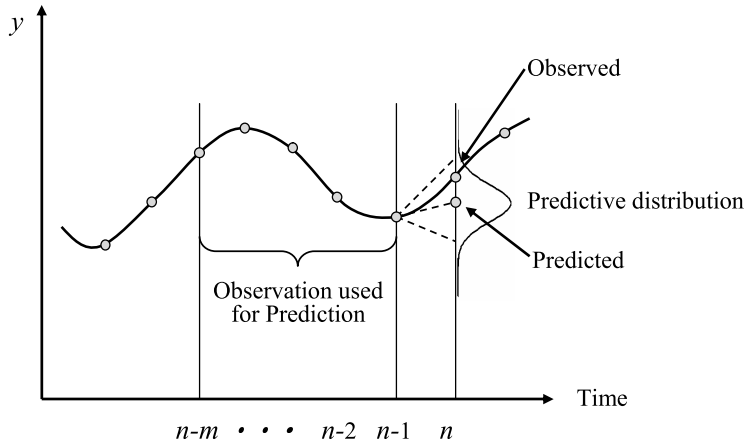

Fig. 2.9. Predictive distribution of time series.

Table 2.1. Residual variances and prediction error variances of AR models with a variety of orders.

p	$\hat{\sigma}_p^2$	PEV_p	p	$\hat{\sigma}_p^2$	PEV_p	p	$\hat{\sigma}_p^2$	PEV_p
0	6.3626	8.0359	7	0.3477	0.3956	14	0.3206	0.3802
1	1.1386	1.3867	8	0.3397	0.3835	15	0.3204	0.3808
2	0.3673	0.4311	9	0.3313	0.3817	16	0.3202	0.3808
3	0.3633	0.4171	10	0.3312	0.3812	17	0.3188	0.3823
4	0.3629	0.4167	11	0.3250	0.3808	18	0.3187	0.3822
5	0.3547	0.4030	12	0.3218	0.3797	19	0.3187	0.3822
6	0.3546	0.4027	13	0.3218	0.3801	20	0.3186	0.3831

distribution. A particular case is that of $p = 0$, which is called *white noise* if x_n is uncorrelated with its own past history. An AR model means that a conditional distribution (also referred to as a predictive distribution) of x_n can be given by the normal distribution having mean $\sum_{j=1}^{p} a_j x_{n-j}$ and variance σ^2.

Similar to the polynomial models, the selection of an appropriate order is an important problem in AR models. When time series data $x_1, \ldots, x_n$ are given, the coefficients a_j and the prediction error variance σ^2 are estimated using the least squares method or the maximum likelihood method. However, the estimated prediction error variance, $\hat{\sigma}_p^2$, of the AR model of order p is a monotonically decreasing function of p. Therefore, if the AR order is determined by this criterion, the maximum order will always be selected, which corresponds to the order selection for the polynomial model in Example 16.

The second column in Table 2.1 indicates the change in $\hat{\sigma}_p^2$ when AR models up to order 20 are fitted to the observations of the rolling angle of a ship [$n = 500$, Kitagawa and Gersch (1996)]. Here, $\hat{\sigma}_p^2$ decreases rapidly up to $p = 2$ and diminishes gradually thereafter. The third column in the table gives the prediction error variance

$$\mathrm{PEV}_p = \frac{1}{500} \sum_{i=501}^{1000} (x_i - \hat{x}_i^p)^2 \,, \tag{2.40}$$

when the subsequent data $x_{501}, \ldots, x_{1000}$ are predicted by

$$\hat{x}_i^p = \sum_{j=1}^{p} \hat{a}_j^p x_{i-j} \quad (i = 501, \ldots, 1000), \tag{2.41}$$

based on the estimated model of order p, where $\hat{a}_j^p$ is an estimate of the j-th coefficient a_j for the AR model of order p. The value of PEV_p is smallest at $p = 12$, and for higher orders, rather than decreasing, the prediction error variance increases.

Even when the time series has a complex structure and the AR model requires a high order p, in some cases an appropriate model can be obtained with fewer parameters by using past values of ε_n together with past values of the time series. The following model is referred to as an *autoregressive moving average* (*ARMA*) *model*:

$$x_n = \sum_{j=1}^{p} a_j x_{n-j} + \varepsilon_n - \sum_{j=1}^{q} b_j \varepsilon_{n-j}. \tag{2.42}$$

In general, if the conditional distribution of a time series x_n is represented by nonlinear functions of the series $x_{n-1}, x_{n-2}, \ldots$ and noise (also called "innovation"), $\varepsilon_n, \varepsilon_{n-1}, \ldots$, then the corresponding model is called a *nonlinear time series model*. If the time series $\boldsymbol{x}_n$ is a vector and the components are interrelated, a multivariate time series model is used for forecasting.

Example 19 (State-space models) A wide variety of time series models such as the ARMA model, trend model, seasonal adjustment model, and time-varying model can be represented using a state-space model. In a state-space model, the time series is expressed by using an unknown m-dimensional state vector $\boldsymbol{\alpha}_n$ as follows:

$$\begin{aligned} \boldsymbol{\alpha}_n &= F_n \boldsymbol{\alpha}_{n-1} + G_n \boldsymbol{v}_n, \\ x_n &= H_n \boldsymbol{\alpha}_n + w_n \,, \end{aligned} \tag{2.43}$$

where $\boldsymbol{v}_n$ and w_n are white noises that have the normal distributions $N_n(0, Q_n)$ and $N(0, \sigma_n^2)$, respectively. Concerning the state-space model, the Kalman filter algorithm is known to efficiently calculate the conditional distributions

$f(\boldsymbol{\alpha}_n|x_{n-1}, x_{n-2}, \ldots)$ and $f(\boldsymbol{\alpha}_n|x_n, x_{n-1}, \ldots)$ of the unknown state $\boldsymbol{\alpha}_n$ from observed time series; these conditional distributions are referred to as a *state prediction distribution* and a *filter distribution*, respectively. Many important problems in time series analysis, such as prediction and control, computation of likelihood, and decomposition into several components, can be solved by using the estimated state vector.

The generalized state-space model is a generalization of the state-space model [Kitagawa (1987)]. It represents the time series as follows:

$$\begin{aligned} \boldsymbol{\alpha}_n &\sim F(\boldsymbol{\alpha}_n|\boldsymbol{\alpha}_{n-1}), \\ x_n &\sim H(x_n|\boldsymbol{\alpha}_n)\,, \end{aligned} \tag{2.44}$$

where F and H denote appropriately specified conditional probability distributions. In other words, generalized state-space models directly model the two conditional distributions that are essential in time series modeling. This conditional distribution model can also be applied when observed data or states are discrete variables. It can be shown that the hidden Markov model is actually a special case of the generalized state-space model. Recently, a sequential Monte Carlo method for recursive estimation of unknown parameters of the generalized state-space models has been developed [see for example, Durbin and Koopman (2001), Harvey (1989), and Kitagawa and Gersch (1996)].

This method can thus be used to estimate the unknown state vector if the (general) state-space model is specified. Since the log-likelihood of the state-space model can be computed by using the predictive distribution of the state, unknown parameters of the model can be estimated using the maximum likelihood method. However, the state-space model is a very flexible model that is capable of expressing a very wide range of time series models. Therefore, in actual time series modeling, we have to compare a large variety of time series models and select an appropriate one.

2.3.3 Spatial Models

The spatial model represents the distribution of data by associating a spatial arrangement with it. For the case when data are arranged in a regular lattice, as depicted in the left plot of Figure 2.10, a model such as

$$p(x_{ij}|x_{i,j-1}, x_{i,j+1}, x_{i-1,j}, x_{i+1,j})\,, \tag{2.45}$$

that represents the data x_{ij} at point (i, j), for example, can be constructed as a conditional distribution of the surrounding four points. As a simple example, a model

$$x_{ij} = \frac{1}{4}(x_{i,j-1} + x_{i,j+1} + x_{i-1,j} + x_{i+1,j}) + \varepsilon_{ij} \tag{2.46}$$

is conceivable in which ε_{ij} is a normal distribution with mean 0 and variance σ^2.

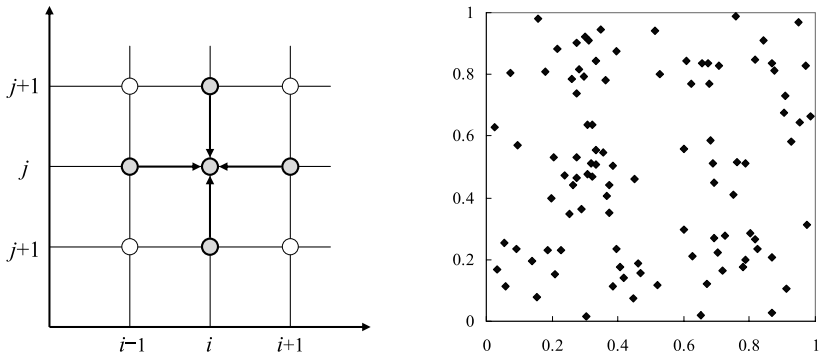

Fig. 2.10. An example of a prediction model for lattice data and spatial data.

On the other hand, in the general case in which the pointwise arrangement of data is not necessarily a lattice pattern, as illustrated in the right plot of Figure 2.10, a model that describes an equilibrium state can be obtained by modeling the local interaction of the points called particles.

Let us assume that the pointwise arrangement $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ of n particles is given. If we define a potential function $\phi(x, y)$ that models the force acting between two points, the sum of the potential energy at the point arrangement $\boldsymbol{x}$ can be given by

$$H(\boldsymbol{x}) = \sum_{1 \le i \le j \le n} \phi(x_i, x_j)\,. \tag{2.47}$$

Then the Gibbs distribution is defined by

$$f(\boldsymbol{x}) = C \exp\{-H(\boldsymbol{x})\}\,, \tag{2.48}$$

where C is a normalization constant defined such that the integration over the entire space is 1. In this method, models on spatial data can be obtained by establishing concrete forms of the potential function $\phi(x, y)$. For the analysis of spatial data, see Cressie (1991).

http://www.springer.com/978-0-387-71886-6

Information Criteria and Statistical Modeling
Konishi, S.; Kitagawa, G.
2008, XII, 276 p., Hardcover
ISBN: 978-0-387-71886-6