3

Information Criterion

In this chapter, we discuss using Kullback–Leibler information as a criterion for evaluating statistical models that approximate the true probability distribution of the data and its properties. We also explain how this criterion for evaluating statistical models leads to the concept of the information criterion, AIC. To this end, we explain the basic framework of model evaluation and the derivation of AIC by adopting a unified approach.

3.1 Kullback–Leibler Information

3.1.1 Definition and Properties

Let $\boldsymbol{x}_n = \{x_1, x_2, \ldots, x_n\}$ be a set of n observations drawn randomly (independently) from an unknown probability distribution function $G(x)$. In the following, we refer to the probability distribution function $G(x)$ that generates data as the true model or the true distribution. In contrast, let $F(x)$ be an arbitrarily specified model. If the probability distribution functions $G(x)$ and $F(x)$ have density functions $g(x)$ and $f(x)$, respectively, then they are called *continuous models* (or *continuous distribution models*). If, given either a finite set or a countably infinite set of discrete points $\{x_1, x_2, \ldots, x_k, \ldots\}$, they are expressed as probabilities of events

$$\begin{aligned} g_i &= g(x_i) \equiv \Pr(\{\omega;\ X(\omega) = x_i\}), \\ f_i &= f(x_i) \equiv \Pr(\{\omega;\ X(\omega) = x_i\}), \quad i = 1, 2, \ldots, \end{aligned} \tag{3.1}$$

then these models are called *discrete models* (*discrete distribution models*).

We assume that the goodness of the model $f(x)$ is assessed in terms of the closeness as a probability distribution to the true distribution $g(x)$. As a measure of this closeness, Akaike (1973) proposed the use of the following *Kullback–Leibler information* [or Kullback–Leibler divergence, Kullback–Leibler (1951), hereinafter abbreviated as “K-L information”]:

$$I(G;F) = E_G\left[\log\left\{\frac{G(X)}{F(X)}\right\}\right], \tag{3.2}$$

where E_G represents the expectation with respect to the probability distribution G.

If the probability distribution functions are continuous models that have the density functions $g(x)$ and $f(x)$, then the K-L information can be expressed as

$$I(g;f) = \int_{-\infty}^{\infty} \log\left\{\frac{g(x)}{f(x)}\right\} g(x)dx. \tag{3.3}$$

If the probability distribution functions are discrete models for which the probabilities are given by $\{g(x_i); i = 1, 2, \ldots\}$ and $\{f(x_i); i = 1, 2, \ldots\}$, then the K-L information can be expressed as

$$I(g;f) = \sum_{i=1}^{\infty} g(x_i)\log\left\{\frac{g(x_i)}{f(x_i)}\right\}. \tag{3.4}$$

By unifying the continuous and discrete models, we can express the K-L information as follows:

$$\begin{aligned} I(g\,;f) &= \int \log\left\{\frac{g(x)}{f(x)}\right\} dG(x) \\ &= \begin{cases} \displaystyle\int_{-\infty}^{\infty} \log\left\{\frac{g(x)}{f(x)}\right\} g(x)dx, & \text{for continuous model,} \\ \displaystyle\sum_{i=1}^{\infty} g(x_i)\log\left\{\frac{g(x_i)}{f(x_i)}\right\}, & \text{for discrete model.} \end{cases} \end{aligned} \tag{3.5}$$

Properties of K-L information. The K-L information has the following properties:

(i) $I(g;f) \geq 0$,
(ii) $I(g;f) = 0 \iff g(x) = f(x)$.

In view of these properties, we consider that the smaller the quantity of K-L information, the closer the model $f(x)$ is to $g(x)$.

Proof. First, let us consider the function $K(t) = \log t - t + 1$, which is defined for $t > 0$. In this case, the derivative of $K(t)$, $K'(t) = t^{-1} - 1$, satisfies the condition $K'(1) = 0$, and $K(t)$ takes its maximum, $K(1) = 0$, at $t = 1$. Therefore, the inequality $K(t) \leq 0$ holds for all t such that $t > 0$. The equality holds only for $t = 1$, which means that the relationship

$$\log t \leq t - 1 \quad (\text{the equality holds only when } \ t = 1)$$

holds.

For the continuous model, by substituting $t = f(x)/g(x)$ into this expression, we obtain

$$\log \frac{f(x)}{g(x)} \le \frac{f(x)}{g(x)} - 1.$$

By multiplying both sides of the equation by $g(x)$ and integrating them, we obtain

$$\begin{aligned} \int \log \left\{ \frac{f(x)}{g(x)} \right\} g(x)dx &\le \int \left\{ \frac{f(x)}{g(x)} - 1 \right\} g(x)dx \\ &= \int f(x)dx - \int g(x)dx = 0. \end{aligned}$$

This gives

$$\int \log \left\{ \frac{g(x)}{f(x)} \right\} g(x)dx = - \int \log \left\{ \frac{f(x)}{g(x)} \right\} g(x)dx \ge 0,$$

thus demonstrating (i). Clearly, the equality holds only when $g(x) = f(x)$.

For the discrete model, it suffices to replace the density functions $g(x)$ and $f(x)$ by the probability functions $g(x_i)$ and $f(x_i)$, respectively, and sum the terms over $i = 1, 2, \ldots$ instead of integrating.

Measures of the similarity between distributions. As a measure of the closeness between distributions, the following quantities have been proposed in addition to the K-L information [Kawada (1987)]:

$$\begin{aligned} \chi^2(g;f) &= \sum_{i=1}^{k} \frac{g_i^2}{f_i} - 1 = \sum_{i=1}^{k} \frac{(f_i - g_i)^2}{f_i} && \chi^2\text{-statistics}, \\ I_K(g;f) &= \int \left\{ \sqrt{f(x)} - \sqrt{g(x)} \right\}^2 dx && \text{Hellinger distance}, \\ I_\lambda(g;f) &= \frac{1}{\lambda} \int \left\{ \left(\frac{g(x)}{f(x)} \right)^\lambda - 1 \right\} g(x)dx && \text{Generalized information}, \\ D(g;f) &= \int u\left(\frac{g(x)}{f(x)} \right) g(x)dx && \text{Divergence}, \\ L_1(g;f) &= \int |g(x) - f(x)|dx && L^1\text{-norm}, \\ L_2(g;f) &= \int \{g(x) - f(x)\}^2 dx && L^2\text{-norm}. \end{aligned}$$

In the above divergence, $D(g;f)$, letting $u(x) = \log x$ produces K-L information $I(g;f)$; similarly, letting $u(x) = \lambda^{-1}(x^\lambda - 1)$ produces generalized information $I_\lambda(g;f)$. In $I_\lambda(g;f)$, when $\lambda \to 0$, we obtain K-L information $I(g;f)$. In this book, following Akaike (1973), the model evaluation criterion based on the K-L information will be referred to generically as an *information criterion.*

3.1.2 Examples of K-L Information

We illustrate K-L information by using several specific examples.

Example 1 (K-L information for normal models) Suppose that the true model $g(x)$ and the specified model $f(x)$ have normal distributions $N(\xi, \tau^2)$ and $N(\mu, \sigma^2)$, respectively. If E_G is an expectation with respect to the true model, the random variable X is distributed according to $N(\xi, \tau^2)$, and therefore, the following equation holds:

$$\begin{aligned} E_G\left[(X-\mu)^2\right] &= E_G\left[(X-\xi)^2 + 2(X-\xi)(\xi-\mu) + (\xi-\mu)^2\right] \\ &= \tau^2 + (\xi-\mu)^2. \end{aligned} \tag{3.6}$$

Thus, for the normal distribution $f(x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-(x-\mu)^2/(2\sigma^2)\right\}$, we obtain

$$\begin{aligned} E_G\left[\log f(X)\right] &= E_G\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(X-\mu)^2}{2\sigma^2}\right] \\ &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{\tau^2 + (\xi-\mu)^2}{2\sigma^2}. \end{aligned} \tag{3.7}$$

In particular, if we let $\mu = \xi$ and $\sigma^2 = \tau^2$ in this expression, it follows that

$$E_G\left[\log g(X)\right] = -\frac{1}{2}\log(2\pi\tau^2) - \frac{1}{2}. \tag{3.8}$$

Therefore, the K-L information of the model $f(x)$ with respect to $g(x)$ is given by

$$\begin{aligned} I(g\,; f) &= E_G\left[\log g(X)\right] - E_G\left[\log f(X)\right] \\ &= \frac{1}{2}\left\{\log\frac{\sigma^2}{\tau^2} + \frac{\tau^2 + (\xi-\mu)^2}{\sigma^2} - 1\right\}. \end{aligned} \tag{3.9}$$

Example 2 (K-L information for normal and Laplace models) Assume that the true model is a two-sided exponential (Laplace) distribution $g(x) = \frac{1}{2}\exp(-|x|)$ and that the specified model $f(x)$ is $N(\mu, \sigma^2)$. In this case, we obtain

$$\begin{aligned} E_G\left[\log g(X)\right] &= -\log 2 - \frac{1}{2}\int_{-\infty}^{\infty} |x| e^{-|x|} dx \\ &= -\log 2 - \int_0^{\infty} x e^{-x} dx \\ &= -\log 2 - 1, \end{aligned} \tag{3.10}$$

$$\begin{aligned} E_G\left[\log f(X)\right] &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{4\sigma^2}\int_{-\infty}^{\infty} (x-\mu)^2 e^{-|x|} dx \\ &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{4\sigma^2}(4 + 2\mu^2). \end{aligned} \tag{3.11}$$

Then the K-L information of the model $f(x)$ with respect to $g(x)$ is given by

$$I(g\,;f) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{2+\mu^2}{2\sigma^2} - \log 2 - 1. \tag{3.12}$$

Example 3 (K-L information for two discrete models) Assume that two dice have the following probabilities for rolling the numbers one to six:

$$\begin{aligned} f_a &= \{0.2,\ 0.12,\ 0.18,\ 0.12,\ 0.20,\ 0.18\}, \\ f_b &= \{0.18,\ 0.12,\ 0.14,\ 0.19,\ 0.22,\ 0.15\}. \end{aligned}$$

In this case, which is the fairer die? Since an ideal die has the probabilities $g = \{1/6,\ \ 1/6,\ \ 1/6,\ \ 1/6,\ \ 1/6,\ \ 1/6\}$, we take this to be the true model. When we calculate the K-L information, $I(g; f)$, the die that gives the smaller value must be closer to the ideal fair die. Calculating the value of

$$I(g; f) = \sum_{i=1}^{6} g_i \log \frac{g_i}{f_i}, \tag{3.13}$$

we obtain $I(g\,;f_a) = 0.023$ and $I(g\,;f_b) = 0.020$. Thus, in terms of K-L information, it must be concluded that die f_b is the fairer of the two.

3.1.3 Topics on K-L Information

Boltzmann's entropy. The negative of the K-L information, $B(g\,;f) = -I(g\,;f)$, is referred to as *Boltzmann's entropy.* In the case of the discrete distribution model $f = \{f_1, \ldots, f_k\}$, the entropy can be interpreted as a quantity that varies proportionally with the logarithm of the probability W in which the relative frequency of the sample obtained from the specified model agrees with the true distribution.

Proof. Suppose that we have n independent samples from a distribution that follows the model f, and assume that either a frequency distribution $\{n_1, \ldots, n_k\}$ $(n_1 + n_2 + \cdots + n_k = n)$ or a relative frequency $\{g_1, g_2, \ldots, g_k\}$ $(g_i = n_i/n)$ is obtained. Since the probability with which such a frequency distribution $\{n_1, \ldots, n_k\}$ is obtained is

$$W = \frac{n!}{n_1! \cdots n_k!} f_1^{n_1} \cdots f_k^{n_k}, \tag{3.14}$$

we take the logarithm of this quantity, and, using Stirling's approximation $(\log n! \sim n \log n - n)$, we obtain

$$\log W = \log n! - \sum_{i=1}^{k} \log n_i! + \sum_{i=1}^{k} n_i \log f_i$$

$$\begin{aligned}
&\sim n\log n - n - \sum_{i=1}^{k} n_i \log n_i + \sum_{i=1}^{k} n_i + \sum_{i=1}^{k} n_i \log f_i \\
&= -\sum_{i=1}^{k} n_i \log\left\{\frac{n_i}{n}\right\} + \sum_{i=1}^{k} n_i \log f_i \\
&= \sum_{i=1}^{k} n_i \log\left\{\frac{f_i}{g_i}\right\} = n\sum_{i=1}^{k} g_i \log\left\{\frac{f_i}{g_i}\right\} \\
&= n \cdot B(g\,;f).
\end{aligned}$$

Hence, it follows that $B(g\,;f) \sim n^{-1}\log W$; that is, $B(g;f)$ is approximately proportional to the logarithm of the probability of which the relative frequency of the sample obtained from the specified model agrees with the true distribution.

We notice that, in the above statement, the K-L information is not the probability of obtaining the distribution defined by a model from the true distribution. Rather, it is thought of as the probability of obtaining the observed data from the model.

On the functional form of K-L information. If the differentiable function F defined on $(0,\infty)$ satisfies the relationship

$$\sum_{i=1}^{k} g_i F(f_i) \leq \sum_{i=1}^{k} g_i F(g_i) \tag{3.15}$$

for any two probability functions $\{g_1,\ldots,g_k\}$ and $\{f_1,\ldots,f_k\}$, then $F(g) = \alpha + \beta\log g$ for some α, β with $\beta > 0$.

Proof. In order to demonstrate that $F(g) = \alpha + \beta\log g$, it suffices to show that $gF'(g) = \beta > 0$ and hence that $\partial F/\partial g = \beta/g$. Let $h = (h_1,\ldots,h_k)^T$ be an arbitrary vector that satisfies $\sum_{i=1}^{k} h_i = 0$ and $|h_i| \leq \max\{g_i, 1-g_i\}$. Since $g+\lambda h$ is a probability distribution, it follows from the assumption that

$$\varphi(\lambda) \equiv \sum_{i=1}^{k} g_i F(g_i + \lambda h_i) \leq \sum_{i=1}^{k} g_i F(g_i) = \varphi(0).$$

Therefore, since

$$\varphi'(\lambda) = \sum_{i=1}^{k} g_i F'(g_i + \lambda h_i) h_i, \quad \varphi'(0) = \sum_{i=1}^{k} g_i F'(g_i) h_i = 0$$

are always true, by writing $h_1 = C,\ h_2 = -C,\ h_i = 0\ (i = 3,\ldots,k)$, we have

$$g_1 F'(g_1) = g_2 F'(g_2) = \text{const} = \beta.$$

The equality for other values of i can be shown in a similar manner.

This result does not imply that the measure that satisfies $I(g:f) \geq 0$ is intrinsically limited to the K-L information. Rather, as indicated by (3.16) in the next section, the result shows that any measure that can be decomposed into two additive terms is limited to the K-L information.

3.2 Expected Log-Likelihood and Corresponding Estimator

The preceding section showed that we can evaluate the appropriateness of a given model by calculating the K-L information. However, K-L information can be used in actual modeling only in limited cases, since K-L information contains the unknown distribution g, so that its value cannot be calculated directly.

K-L information can be decomposed into

$$I(g\,;f) = E_G\left[\log\left\{\frac{g(X)}{f(X)}\right\}\right] = E_G\left[\log g(X)\right] - E_G\left[\log f(X)\right]. \tag{3.16}$$

Moreover, because the first term on the right-hand side is a constant that depends solely on the true model g, it is clear that in order to compare different models, it is sufficient to consider only the second term on the right-hand side. This term is called the *expected log-likelihood.* The larger this value is for a model, the smaller its K-L information is and the better the model is.

Since the expected log-likelihood can be expressed as

$$\begin{aligned} E_G\left[\log f(X)\right] &= \int \log f(x) dG(x) \\ &= \begin{cases} \displaystyle\int_{-\infty}^{\infty} g(x)\log f(x)dx, & \text{for continuous models,} \\ \displaystyle\sum_{i=1}^{\infty} g(x_i)\log f(x_i), & \text{for discrete models,} \end{cases} \end{aligned} \tag{3.17}$$

it still depends on the true distribution g and is an unknown quantity that eludes explicit computation. However, if a good estimate of the expected log-likelihood can be obtained from the data, this estimate can be used as a criterion for comparing models. Let us now consider the following problem.

Let $\boldsymbol{x}_n = \{x_1, x_2, \ldots, x_n\}$ be data observed from the true distribution $G(x)$ or $g(x)$. An estimate of the expected log-likelihood can be obtained by replacing the unknown probability distribution G contained in (3.17) with an empirical distribution function $\hat{G}$ based on data $\boldsymbol{x}_n$. The empirical distribution function is the distribution function for the probability function $\hat{g}(x_\alpha)= 1/n$ $(\alpha = 1, 2, \ldots, n)$ that has the equal probability $1/n$ for each of n observations

$\{x_1, x_2, \ldots, x_n\}$ (see Section 5.1). In fact, by replacing the unknown probability distribution G contained in (3.17) with the empirical distribution function $\hat{G}(x)$, we obtain

$$\begin{aligned} E_{\hat{G}}\left[\log f(X)\right] &= \int \log f(x) d\hat{G}(x) \\ &= \sum_{\alpha=1}^{n} \hat{g}(x_\alpha) \log f(x_\alpha) \\ &= \frac{1}{n} \sum_{\alpha=1}^{n} \log f(x_\alpha). \end{aligned} \tag{3.18}$$

According to the law of large numbers, when the number of observations, n, tends to infinity, the mean of the random variables $Y_\alpha = \log f(X_\alpha)$ ($\alpha = 1, 2, \ldots, n$) converges in probability to its expectation, that is, the convergence

$$\frac{1}{n} \sum_{\alpha=1}^{n} \log f(X_\alpha) \longrightarrow E_G\left[\log f(X)\right], \qquad n \to +\infty, \tag{3.19}$$

holds. Therefore, it is clear that the estimate based on the empirical distribution function in (3.18) is a natural estimate of the expected log-likelihood. The estimate of the expected log-likelihood multiplied by n, i.e.,

$$n \int \log f(x) d\hat{G}(x) = \sum_{\alpha=1}^{n} \log f(x_\alpha), \tag{3.20}$$

is the log-likelihood of the model $f(x)$. This means that the *log-likelihood*, frequently used in statistical analyses, is clearly understood as being an approximation to the K-L information.

Example 4 (Expected log-likelihood for normal models) Let both of the continuous models $g(x)$ and $f(x)$ be the standard normal distribution $N(0, 1)$ with mean 0 and variance 1. Let us generate n observations, $\{x_1, x_2, \ldots, x_n\}$, from the true model $g(x)$ to construct the empirical distribution function $\hat{G}$. In the next step, we calculate the value of (3.18),

$$E_{\hat{G}}\left[\log f(X)\right] = -\frac{1}{2}\log(2\pi) - \frac{1}{2n}\sum_{\alpha=1}^{n} x_\alpha^2.$$

Table 3.1 shows the results of obtaining the mean and the variance of $E_{\hat{G}}[\log f(X)]$ by repeating this process 1,000 times.

Since the average of the 1,000 trials is very close to the true value, that is, the expected log-likelihood

$$E_G[\log f(X)] = \int g(x) \log f(x) dx = -\frac{1}{2}\log(2\pi) - \frac{1}{2} = -1.4189,$$

Table 3.1. Distribution of the log-likelihood of a normal distribution model. The mean, variance, and standard deviation are obtained by running 1,000 Monte Carlo trials. The expression $E_G[\log f(X)]$ represents the expected log-likelihood.

n	10	100	1,000	10,000	$E_G[\log f(X)]$
Mean	−1.4188	−1.4185	−1.4191	−1.4189	−1.4189
Variance	0.05079	0.00497	0.00050	0.00005	—
Standard deviation	0.22537	0.07056	0.02232	0.00696	—

the results suggest that even for a small number of observations, the log-likelihood has little bias. By contrast, the variance decreases in inverse proportion to n.

3.3 Maximum Likelihood Method and Maximum Likelihood Estimators

3.3.1 Log-Likelihood Function and Maximum Likelihood Estimators

Let us consider the case in which a model is given in the form of a probability distribution $f(x|\boldsymbol{\theta})(\boldsymbol{\theta} \in \Theta \subset R^p)$, having unknown p-dimensional parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$. In this case, given data $\boldsymbol{x}_n = \{x_1, x_2, \ldots, x_n\}$, the log-likelihood can be determined for each $\boldsymbol{\theta} \in \Theta$. Therefore, by regarding the log-likelihood as a function of $\boldsymbol{\theta} \in \Theta$, and representing it as

$$\ell(\boldsymbol{\theta}) = \sum_{\alpha=1}^{n} \log f(x_\alpha|\boldsymbol{\theta}), \tag{3.21}$$

the log-likelihood is referred to as the *log-likelihood function.* A natural estimator of $\boldsymbol{\theta}$ is defined by finding the maximizer $\boldsymbol{\theta} \in \Theta$ of the $\ell(\boldsymbol{\theta})$, that is, by determining $\boldsymbol{\theta}$ that satisfies the equation

$$\ell(\hat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}). \tag{3.22}$$

This method is called the *maximum likelihood method*, and $\hat{\boldsymbol{\theta}}$ is called the *maximum likelihood estimator*. If the data used in the estimation must be specified explicitly, then the maximum likelihood estimator is denoted by $\hat{\boldsymbol{\theta}}(\boldsymbol{x}_n)$. The model $f(x|\hat{\boldsymbol{\theta}})$ determined by $\hat{\boldsymbol{\theta}}$ is called the *maximum likelihood model*, and the term $\ell(\hat{\boldsymbol{\theta}}) = \sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\boldsymbol{\theta}})$ is called the *maximum log-likelihood.*

3.3.2 Implementation of the Maximum Likelihood Method by Means of Likelihood Equations

If the log-likelihood function $\ell(\boldsymbol{\theta})$ is continuously differentiable, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is given as a solution of the likelihood equation

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_i} = 0, \quad i = 1, 2, \ldots, p \quad \text{or} \quad \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}, \tag{3.23}$$

where $\partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ is a p-dimensional vector, the i^{th} component of which is given by $\partial \ell(\boldsymbol{\theta})/\partial \theta_i$, and $\mathbf{0}$ is the p-dimensional zero vector, all the components of which are 0. In particular, if the likelihood equation is a linear equation having p-dimensional parameters, the maximum likelihood estimator can be expressed explicitly.

Example 5 (Normal model) Let us consider the normal distribution model $N(\mu, \sigma^2)$ with respect to the data $\{x_1, x_2, \ldots, x_n\}$. Since the log-likelihood function is given by

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{\alpha=1}^{n} (x_\alpha - \mu)^2, \tag{3.24}$$

the likelihood equation takes the form

$$\begin{aligned} \frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{\alpha=1}^{n} (x_\alpha - \mu) = \frac{1}{\sigma^2} \left(\sum_{\alpha=1}^{n} x_\alpha - n\mu \right) = 0, \\ \frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{\alpha=1}^{n} (x_\alpha - \mu)^2 = 0. \end{aligned}$$

It follows, then, that the maximum likelihood estimators for μ and σ^2 are

$$\hat{\mu} = \frac{1}{n} \sum_{\alpha=1}^{n} x_\alpha, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{\alpha=1}^{n} (x_\alpha - \hat{\mu})^2. \tag{3.25}$$

For the following 20 observations

−7.99	−4.01	−1.56	−0.99	−0.93	−0.80	−0.77	−0.71	−0.42	−0.02
0.65	0.78	0.80	1.14	1.15	1.24	1.29	2.81	4.84	6.82

the maximum likelihood estimates of μ and σ^2 are calculated as

$$\hat{\mu} = \frac{1}{n} \sum_{\alpha=1}^{n} x_\alpha = 0.166, \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{\alpha=1}^{n} (x_\alpha - \hat{\mu})^2 = 8.545, \tag{3.26}$$

and the maximum log-likelihood is

$$\ell(\hat{\mu}, \hat{\sigma}^2) = -\frac{n}{2} \log(2\pi\hat{\sigma}^2) - \frac{n}{2} = -49.832. \tag{3.27}$$

Example 6 (Bernoulli model) The log-likelihood function based on n observations $\{x_1, x_2, \ldots, x_n\}$ drawn from the Bernoulli distribution $f(x|p) = p^x(1-p)^{1-x}$ $(x = 0, 1)$ is

$$\begin{aligned}\ell(p) &= \log\left\{\prod_{\alpha=1}^{n} p^{x_\alpha}(1-p)^{1-x_\alpha}\right\} \\ &= \sum_{\alpha=1}^{n} x_\alpha \log p + \left(n - \sum_{\alpha=1}^{n} x_\alpha\right)\log(1-p). \qquad (3.28)\end{aligned}$$

Consequently, the likelihood equation is

$$\frac{\partial \ell(p)}{\partial p} = \frac{1}{p}\sum_{\alpha=1}^{n} x_\alpha - \frac{1}{1-p}\left(n - \sum_{\alpha=1}^{n} x_\alpha\right) = 0. \qquad (3.29)$$

Thus, the maximum likelihood estimator for p is given by

$$\hat{p} = \frac{1}{n}\sum_{\alpha=1}^{n} x_\alpha. \qquad (3.30)$$

Example 7 (Linear regression model) Let $\{y_\alpha, x_{\alpha 1}, x_{\alpha 2}, \ldots, x_{\alpha p}\}$ $(\alpha = 1, 2, \ldots, n)$ be n sets of data that are observed with respect to a response variable y and p explanatory variables $\{x_1, x_2, \ldots, x_p\}$. In order to describe the relationship between the variables, we assume the following linear regression model with Gaussian noise:

$$y_\alpha = \boldsymbol{x}_\alpha^T \boldsymbol{\beta} + \varepsilon_\alpha, \quad \varepsilon_\alpha \sim N(0, \sigma^2), \quad \alpha = 1, 2, \ldots, n, \qquad (3.31)$$

where $\boldsymbol{x}_\alpha = (1, x_{\alpha 1}, x_{\alpha 2}, \ldots, x_{\alpha p})^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$. Since the probability density function of y_α is

$$f(y_\alpha|\boldsymbol{x}_\alpha; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\left(y_\alpha - \boldsymbol{x}_\alpha^T\boldsymbol{\beta}\right)^2\right\}, \qquad (3.32)$$

the log-likelihood function is expressed as

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_{\alpha=1}^{n} \log f(y_\alpha|\boldsymbol{x}_\alpha; \boldsymbol{\theta}) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{\alpha=1}^{n}\left(y_\alpha - \boldsymbol{x}_\alpha^T\boldsymbol{\beta}\right)^2 \qquad (3.33) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}),\end{aligned}$$

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ and $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^T$. By taking partial derivatives of the above equation with respect to the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2)^T$, the likelihood equation is given by

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= -\frac{1}{2\sigma^2}\left(-2X^T\boldsymbol{y} + 2X^TX\boldsymbol{\beta}\right) = \mathbf{0},\\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta}) = 0. \end{aligned} \qquad (3.34)$$

Consequently, the maximum likelihood estimators for $\boldsymbol{\beta}$ and σ^2 are given by

$$\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\boldsymbol{y}, \qquad \hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}). \qquad (3.35)$$

3.3.3 Implementation of the Maximum Likelihood Method by Numerical Optimization

Although in the preceding section we showed cases in which it was possible to obtain an explicit solution to the likelihood equations, in general likelihood equations are complex nonlinear functions of the parameter vector $\boldsymbol{\theta}$. In this subsection, we describe how to obtain the maximum likelihood estimator in such situations.

When a given likelihood equation cannot be solved explicitly, a numerical optimization method is frequently employed, which involves starting from an appropriately chosen initial value $\boldsymbol{\theta}_0$ and successively generating quantities $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots,$ in order to cause convergence to the solution $\hat{\boldsymbol{\theta}}$. Assuming that the estimated value $\boldsymbol{\theta}_k$ can be determined at some stage, we determine the next point, $\boldsymbol{\theta}_{k+1}$, which yields a larger likelihood, using the method described below.

In the maximum likelihood method, in order to determine the $\hat{\boldsymbol{\theta}}$ that maximizes $\ell(\boldsymbol{\theta})$, we find $\boldsymbol{\theta}$ that satisfies the necessary condition, namely the likelihood equation $\partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \mathbf{0}$. However, since $\boldsymbol{\theta}_k$ does not exactly satisfy $\partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \mathbf{0}$, we generate the next point, $\boldsymbol{\theta}_{k+1}$, in order to approximate 0 closer. For this purpose, we first perform a Taylor series expansion of $\partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ in the neighborhood of $\boldsymbol{\theta}_k$,

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \approx \frac{\partial \ell(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \ell(\boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}(\boldsymbol{\theta} - \boldsymbol{\theta}_k). \qquad (3.36)$$

Then by writing

$$\boldsymbol{g}(\boldsymbol{\theta}) = \left(\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1}, \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_2}, \cdots, \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_p}\right)^T,$$

$$H(\boldsymbol{\theta}) = \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} = \left(\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right), \quad i, j = 1, 2, \ldots, p, \qquad (3.37)$$

in terms of $\boldsymbol{\theta}$ that satisfies $\partial \ell(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \mathbf{0}$, we obtain

$$\mathbf{0} = \boldsymbol{g}(\boldsymbol{\theta}) \approx \boldsymbol{g}(\boldsymbol{\theta}_k) + H(\boldsymbol{\theta}_k)(\boldsymbol{\theta} - \boldsymbol{\theta}_k), \qquad (3.38)$$

where the quantity $\boldsymbol{g}(\boldsymbol{\theta}_k)$ is a gradient vector and $H(\boldsymbol{\theta}_k)$ is a Hessian matrix. By virtue of (3.38), it follows that $\boldsymbol{\theta} \approx \boldsymbol{\theta}_k - H(\boldsymbol{\theta}_k)^{-1}\boldsymbol{g}(\boldsymbol{\theta}_k)$. Therefore, using

$$\boldsymbol{\theta}_{k+1} \equiv \boldsymbol{\theta}_k - H(\boldsymbol{\theta}_k)^{-1}\boldsymbol{g}(\boldsymbol{\theta}_k),$$

we determine the next point, $\boldsymbol{\theta}_{k+1}$. This technique, called *the Newton–Raphson method*, is known to converge rapidly near the root, or in other words, provided an appropriate initial value is chosen.

Thus, while the Newton–Raphson method is considered to be an efficient technique, several difficulties may be encountered when it is applied to maximum likelihood estimation: (1) in many cases, it may prove difficult to calculate the Hessian matrix, which is the 2nd-order partial derivative of the log-likelihood; (2) for each matrix, the method requires calculating the inverse matrix of $H(\boldsymbol{\theta}_k)$; and (3) depending on how the initial value is selected, the method may converge very slowly or even diverge.

In order to mitigate these problems, a *quasi-Newton method* is employed. This method does not involve calculating the Hessian matrix and automatically generates the inverse matrix, $H^{-1}(\boldsymbol{\theta}_k)$. In addition, step widths can be introduced either to accelerate convergence or to prevent divergence. Specifically, the following algorithm is employed in order to successively generate $\boldsymbol{\theta}_{k+1}$:

(i) Determine a search (descending) direction vector $\boldsymbol{d}_k = -H_k^{-1}\boldsymbol{g}_k$.
(ii) Determine the optimum step width λ_k that maximizes $\ell(\boldsymbol{\theta}_k + \lambda\boldsymbol{d}_k)$.
(iii) By taking $\boldsymbol{\theta}_{k+1} \equiv \boldsymbol{\theta}_k + \lambda_k\boldsymbol{d}_k$, determine the next point, $\boldsymbol{\theta}_{k+1}$, and set $\boldsymbol{y}_k \equiv \boldsymbol{g}(\boldsymbol{\theta}_{k+1}) - \boldsymbol{g}(\boldsymbol{\theta}_k)$.
(iv) Update an estimate of $H(\boldsymbol{\theta}_k)^{-1}$ by using either the Davidon–Fletcher–Powell (DFP) algorithm or the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm:

$$H_{k+1}^{-1} = H_k^{-1} + \frac{\boldsymbol{s}_k\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k} - \frac{H_k^{-1}\boldsymbol{y}_k\boldsymbol{y}_k^T H_k^{-1}}{\boldsymbol{y}_k^T H_k^{-1}\boldsymbol{y}_k}, \tag{3.39}$$

$$H_{k+1}^{-1} = H_k^{-1} + \frac{\boldsymbol{s}_k\boldsymbol{y}_k^T H_k^{-1}}{\boldsymbol{s}_k^T\boldsymbol{y}_k} - \frac{H_k^{-1}\boldsymbol{y}_k\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k} + \left\{1 + \frac{\boldsymbol{y}_k H_k^{-1}\boldsymbol{y}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k}\right\}\frac{\boldsymbol{s}_k\boldsymbol{s}_k^T}{\boldsymbol{s}_k^T\boldsymbol{y}_k},$$

where $\boldsymbol{s}_k = \boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_k$.

When applying the quasi-Newton method, one starts with appropriate initial values, $\boldsymbol{\theta}_0$ and H_0^{-1}, and successively determines $\boldsymbol{\theta}_k$ and H_k^{-1}. As an initial value for H_0^{-1}, the identity matrix I, an appropriately scaled matrix of the unit matrix, or an approximate value of $H(\boldsymbol{\theta}_0)^{-1}$ is used. In situations in which it is also difficult to calculate the gradient vector $\boldsymbol{g}(\boldsymbol{\theta})$ of a log-likelihood function, $\boldsymbol{g}(\boldsymbol{\theta})$ can be determined solely from the log-likelihood by numerical differentiation.

Other methods besides the Newton–Raphson method and the quasi-Newton method described above (for example, the simplex method) can be used to obtain the maximum likelihood estimate, since it suffices to determine $\boldsymbol{\theta}$ that maximizes the log-likelihood function.

Example 8 (Cauchy distribution model) Consider the Cauchy distribution model expressed by

$$f(x|\mu,\tau^2) = \sum_{\alpha=1}^{n} \log f(x_\alpha|\mu,\tau^2) = \frac{1}{\pi}\frac{\tau}{(y-\mu)^2+\tau^2} \tag{3.40}$$

for the data shown in Example 5. The log-likelihood of the Cauchy distribution model is given by

$$\ell(\mu,\tau^2) = \frac{n}{2}\log\tau^2 - n\log\pi - \sum_{\alpha=1}^{n}\log\left\{(x_\alpha-\mu)^2+\tau^2\right\}. \tag{3.41}$$

Then the first derivatives of $\ell(\mu,\tau^2)$ with respect to μ and τ^2 are

$$\begin{aligned} \frac{\partial\ell}{\partial\mu} &= 2\sum_{\alpha=1}^{n}\frac{x_\alpha-\mu}{(x_\alpha-\mu)^2+\tau^2}, \\ \frac{\partial\ell}{\partial\tau^2} &= \frac{n}{2\tau^2} - \sum_{\alpha=1}^{n}\frac{1}{(x_\alpha-\mu)^2+\tau^2}. \end{aligned} \tag{3.42}$$

The maximum likelihood estimates of the parameters μ and τ^2 are then obtained by maximizing the log-likelihood using the quasi-Newton method. Table 3.2 shows the results of the quasi-Newton method when the initial estimates are set to $\theta_0 = (\mu_0, \tau_0^2)^T = (0,1)^T$. The quasi-Newton method only required five iterations to find the maximum likelihood estimates.

Table 3.2. Estimation of the parameters of the Cauchy distribution model by a quasi-Newton algorithm.

k	μ_k	τ_k^2	$\ell(\theta_k)$	$\partial\ell/\partial\mu$	$\partial\ell/\partial\tau^2$
0	0.00000	1.00000	48.12676	−0.83954	−1.09776
1	0.23089	1.30191	47.87427	0.18795	−0.14373
2	0.17969	1.35705	47.86554	−0.04627	−0.04276
3	0.18940	1.37942	47.86484	0.00244	−0.00106
4	0.18886	1.38004	47.86484	−0.00003	−0.00002
5	0.18887	1.38005	47.86484	0.00000	0.00000

Example 9 (Time series model) In general, the time series are mutually correlated and the log-likelihood of the time series model cannot be expressed as the sum of the logarithms of the density function of each observation.

However, the likelihood can generally be expressed by using the conditional distributions as follows:

$$L(\theta) = f(y_1, \dots, y_N|\theta) = \prod_{n=1}^{N} f(y_n|y_1, \dots, y_{n-1}). \tag{3.43}$$

Here, for some simple models, each conditional distribution on the right-hand side of the above expression can be obtained from the specified model. For example, for the autoregressive model,

$$y_n = \sum_{j=1}^{m} a_j y_{n-j} + \varepsilon_n, \quad \varepsilon_n \sim N(0, \sigma^2), \tag{3.44}$$

for $n > m$, the conditional distribution is obtained by

$$f(y_n|y_1, \dots, y_{n-1}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} \left(y_n - \sum_{j=1}^{m} a_j y_{n-j} \right)^2 \right\}. \tag{3.45}$$

By ignoring the first m conditional distributions, the log-likelihood of an AR model can be approximated by

$$\ell(\theta) = -\frac{N-m}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=m+1}^{N} \left(y_n - \sum_{j=1}^{m} a_j y_{n-j} \right)^2, \tag{3.46}$$

where $\theta = (a_1, \dots, a_m, \sigma^2)^T$. The least squares estimates of the parameters of the AR model are easily obtained by maximizing the approximate log-likelihood. However, for exact maximum likelihood estimation, we need to use the state-space representation of the model shown below.

In general, we assume that the time series y_n is expressed by a state-space model

$$\begin{aligned} \boldsymbol{x}_n &= F_n \boldsymbol{x}_{n-1} + G_n \boldsymbol{v}_n, \\ y_n &= H_n \boldsymbol{x}_n + w_n, \end{aligned} \tag{3.47}$$

where $\boldsymbol{x}_n$ is a properly defined k-dimensional state vector; F_n, G_n, and H_n are $k \times k$, $k \times \ell$, and $1 \times k$ matrices; and $\boldsymbol{v}_n \sim N_\ell(0, Q_n)$ and $w_n \sim N(0, \sigma^2)$. Then the one-step-ahead predictor $\boldsymbol{x}_{n|n-1}$ and its variance covariance matrix $V_{n|n-1}$ of the state vector $\boldsymbol{x}_n$ given the observations $y_1, ..., y_{n-1}$ can be obtained very efficiently by using the Kalman filter recursive algorithm as follows [Anderson and Moore (1979) and Kitagawa and Gersch (1996)]:

One-step-ahead prediction

$$\begin{aligned} \boldsymbol{x}_{n|n-1} &= F_n \boldsymbol{x}_{n-1|n-1}, \\ V_{n|n-1} &= F_n V_{n-1|n-1} F_n^T + G_n Q_n G_n^T. \end{aligned} \tag{3.48}$$

Filter

$$\begin{aligned}K_n &= V_{n|n-1}H_n^T(H_nV_{n|n-1}H_n^T+\sigma^2)^{-1},\\ \boldsymbol{x}_{n|n} &= \boldsymbol{x}_{n|n-1}+K_n(y_n-H_n\boldsymbol{x}_{n|n-1}),\\ V_{n|n} &= (I-K_nH_n)V_{n|n-1}.\end{aligned} \tag{3.49}$$

Then the one-step-ahead predictive distribution of the observation y_n given $\{y_1,\ldots,y_{n-1}\}$ can be expressed as

$$p(y_n|y_1,\ldots,y_{n-1}) = \frac{1}{\sqrt{2\pi r_n}}\exp\left\{-\frac{(y_n-H_n\boldsymbol{x}_{n|n-1})^2}{2r_n}\right\} \tag{3.50}$$

with $r_n = H_nV_{n|n-1}H_n^T+R_n$. Therefore, if the model contains some unknown parameter vector $\boldsymbol{\theta}$, the log-likelihood of the time series model expressed in the state-space model is given by

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2}\left\{N\log 2\pi+\sum_{n=1}^{N}\log r_n+\sum_{n=1}^{N}\frac{(y_n-H_n\boldsymbol{x}_{n|n-1})^2}{r_n}\right\}. \tag{3.51}$$

The maximum likelihood estimate of the parameter $\hat{\boldsymbol{\theta}}$ is obtained by maximizing (3.51) with respect to those parameters used in a numerical optimization method.

3.3.4 Fluctuations of the Maximum Likelihood Estimators

Assume that the true distribution $g(x)$ that generates data is the standard normal distribution $N(0,1)$ with mean 0 and variance 1 and that the specified model $f(x|\theta)$ is a normal distribution in which either the mean μ or the variance σ^2 is unknown. Figures 3.1 and 3.2 are plots of the log-likelihood function

$$\ell(\mu) = -\frac{n}{2}\log(2\pi)-\frac{1}{2}\sum_{\alpha=1}^{n}(x_\alpha-\mu)^2, \tag{3.52}$$

based on n observations with an unknown mean μ and the variance $\sigma^2=1$. The horizontal axis represents the value of μ, and the vertical axis represents the corresponding value of $\ell(\mu)$. Figures 3.1 and 3.2 show log-likelihood functions based on $n=10$ and $n=100$ observations, respectively. In these figures, random numbers are used to generate 10 sets of observations $\{x_1,x_2,\ldots,x_n\}$ following the distribution $N(0\ ,1)$, and the log-likelihood functions $\ell(\mu)$ $(-2\le\mu\le 2)$ calculated from the observation sets are overlaid. The value of μ that maximizes these functions is the maximum likelihood estimate of the mean, which is plotted on the horizontal axis with lines pointing downward from the axis. The estimator has a scattered profile depending on the data involved. In the figures, the bold curves represent the expected log-likelihood function

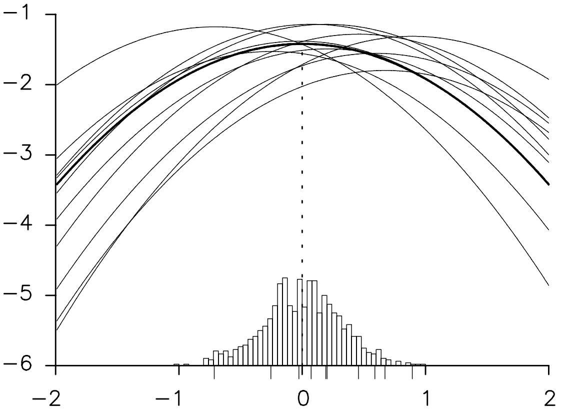

Fig. 3.1. Distributions of expected log-likelihood (bold lines), log-likelihood (thin lines), and maximum likelihood estimators with respect to the mean μ of normal distributions; $n = 10$.

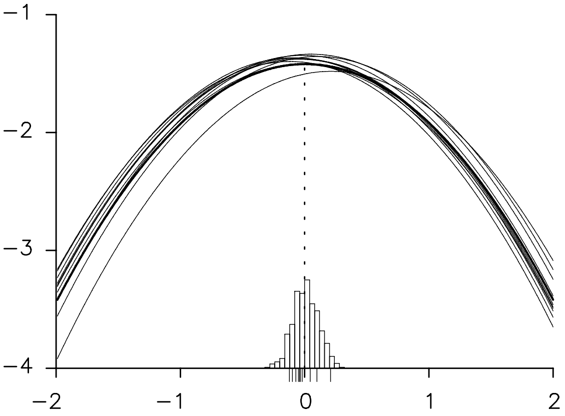

Fig. 3.2. Distributions of the expected log-likelihood (bold), log-likelihood (thin), and maximum likelihood estimator with respect to the mean μ of the normal distribution; $n = 100$.

$$nE_G\left[\log f(X|\mu)\right] = n\int g(x)\log f(x|\mu)dx = -\frac{n}{2}\log(2\pi) - \frac{n(1+\mu^2)}{2},$$

and the values of the true parameter μ_0 corresponding to the function are plotted as dotted lines. The difference between these values and the maximum likelihood estimate is the estimation error of μ. The histogram in the figure, which shows the distribution of the maximum likelihood estimates resulting from similar calculations repeated 1,000 times, indicates that the maximum

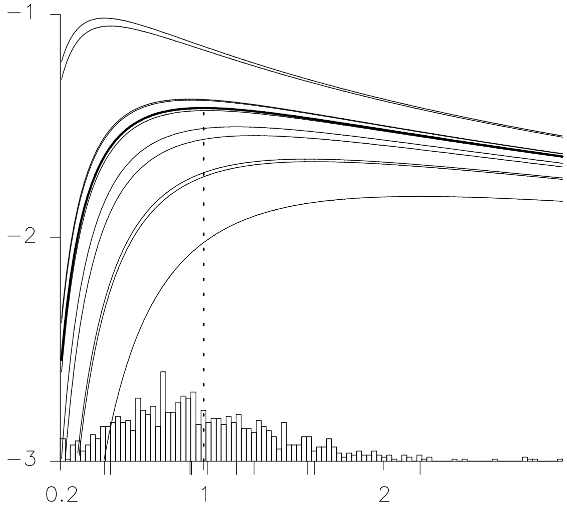

Fig. 3.3. Distributions of the expected log-likelihood (bold), log-likelihood (thin), and maximum likelihood estimator with respect to the variance σ^2 of the normal distribution; $n = 10$.

likelihood estimator has a distribution over a range of ±1 in the case of $n = 10$, and ±0.3 in the case of $n = 100$.

Figures 3.3 and 3.4 show 10 overlaid plots of the following log-likelihood function, obtained from $n = 10$ and $n = 100$ observations, respectively, with unknown variance σ^2 and the mean $\mu = 0$:

$$\ell(\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{\alpha=1}^{n} x_\alpha^2.$$

In this case, $\ell(\sigma^2)$ is an asymmetric function of σ^2, and the corresponding distribution of the maximum likelihood estimator is also asymmetric. In this case, too, the figures suggest that the distribution of the estimators converges to the true value as n increases. In the figures, the bold curve represents the expected log-likelihood function

$$nE_G\left[\log f(X|\sigma^2)\right] = n\int g(x)\log f(x|\sigma^2)dx = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{n}{2\sigma^2},$$

and the value of the corresponding true parameter is shown by the dotted line. The difference between this value and the maximum likelihood estimator is the estimation error of σ^2. The histograms in the figures show the distribution of the maximum likelihood estimator when the same calculations are repeated 1,000 times.

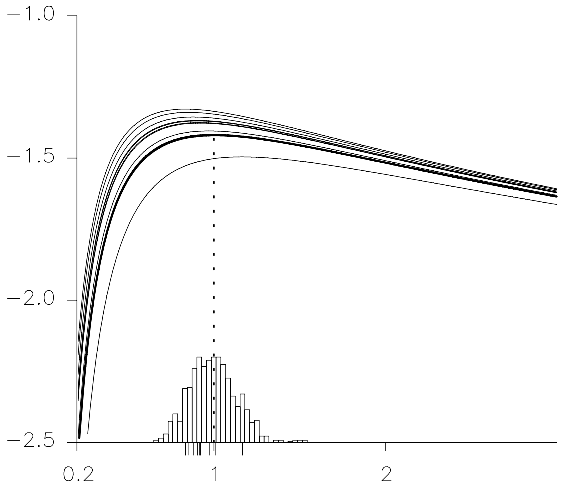

Fig. 3.4. Distributions of the expected log-likelihood (bold), log-likelihood (thin), and maximum likelihood estimator with respect to the variance σ^2 of the normal distribution; $n = 100$.

3.3.5 Asymptotic Properties of the Maximum Likelihood Estimators

This section discusses the asymptotic properties of the maximum likelihood estimator of a continuous parametric model $\{f(x|\boldsymbol{\theta});\ \boldsymbol{\theta} \in \Theta \subset R^p\}$ with p-dimensional parameter vector $\boldsymbol{\theta}$.

Asymptotic normality. Assume that the following regularity condition holds for the density function $f(x|\boldsymbol{\theta})$:

(1) The function $\log f(x|\boldsymbol{\theta})$ is three times continuously differentiable with respect to $\boldsymbol{\theta} = (\theta_1,\, \theta_2,\, \ldots,\, \theta_p)^T$.

(2) There exist integrable functions on R, $F_1(x)$, and $F_2(x)$ and a function $H(x)$ such that

$$\int_{-\infty}^{\infty} H(x) f(x|\boldsymbol{\theta}) dx < M,$$

for an appropriate real value M, and the following inequalities hold for any $\boldsymbol{\theta} \in \Theta$:

$$\left| \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_i} \right| < F_1(x), \qquad \left| \frac{\partial^2 \log f(x|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right| < F_2(x),$$

$$\left| \frac{\partial^3 \log f(x|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j \partial \theta_k} \right| < H(x), \quad i, j, k = 1, 2, \ldots, p.$$

(3) The following inequality holds for an arbitrary $\boldsymbol{\theta} \in \Theta$:

$$0 < \int_{-\infty}^{\infty} f(x|\boldsymbol{\theta}) \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_j} dx < \infty, \quad i, j = 1, \ldots, p. \tag{3.53}$$

Then, under the above conditions the following properties can be derived:
(a) Assume that $\boldsymbol{\theta}_0$ is a solution of

$$\int f(x|\boldsymbol{\theta}) \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} dx = \mathbf{0} \tag{3.54}$$

and that data $\boldsymbol{x}_n = \{x_1, x_2, \ldots, x_n\}$ are obtained according to the density function $f(x|\boldsymbol{\theta}_0)$. In addition, let $\hat{\boldsymbol{\theta}}_n$ be the maximum likelihood estimator based on n observations. Then the following properties hold:

(i) The likelihood equation

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{\alpha=1}^{n} \frac{\partial \log f(x_\alpha|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0} \tag{3.55}$$

has a solution that converges to $\boldsymbol{\theta}_0$.

(ii) The maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ converges in probability to $\boldsymbol{\theta}_0$ when $n \to +\infty$.

(iii) The maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ has asymptotic normality, that is, the distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ converges in law to the p-dimensional normal distribution $N_p(\mathbf{0}, I(\boldsymbol{\theta}_0)^{-1})$ with the mean vector $\mathbf{0}$ and the variance covariance matrix $I(\boldsymbol{\theta}_0)^{-1}$, where the matrix $I(\boldsymbol{\theta}_0)$ is the value of the matrix $I(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, which is given by

$$I(\boldsymbol{\theta}) = \int f(x|\boldsymbol{\theta}) \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} dx. \tag{3.56}$$

This matrix $I(\boldsymbol{\theta})$, with $(i, j)^{th}$ component given as (3.53) under condition (3), is called the *Fisher information matrix.*

Although the asymptotic normality stated above assumes the existence of $\boldsymbol{\theta}_0 \in \Theta$ that satisfies the assumption $g(x) = f(x|\boldsymbol{\theta}_0)$, similar results, given below, can also be obtained even when the assumption does not hold:

(b) Assume that $\boldsymbol{\theta}_0$ is a solution of

$$\int g(x) \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} dx = \mathbf{0} \tag{3.57}$$

and that data $\boldsymbol{x}_n = \{x_1, x_2, \cdots, x_n\}$ are observed according to the distribution $g(x)$. In this case, the following statements hold with respect to the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$:

(i) The maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ converges in probability to $\boldsymbol{\theta}_0$ as $n \to +\infty$.

(ii) The distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ with respect to the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ converges in law to the p-dimensional normal distribution with the mean vector $\mathbf{0}$ and the variance covariance matrix $J^{-1}(\boldsymbol{\theta}_0)I(\boldsymbol{\theta}_0)J^{-1}(\boldsymbol{\theta}_0)$ as $n \to +\infty$. In other words, when $n \to +\infty$, the following holds:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to N_p\left(\mathbf{0}, J^{-1}(\boldsymbol{\theta}_0)I(\boldsymbol{\theta}_0)J^{-1}(\boldsymbol{\theta}_0)\right), \tag{3.58}$$

where the matrices $I(\boldsymbol{\theta}_0)$ and $J(\boldsymbol{\theta}_0)$ are the $p \times p$ matrices evaluated at $\boldsymbol{\theta}= \boldsymbol{\theta}_0$ and are given by the following equations:

$$\begin{aligned} I(\boldsymbol{\theta}) &= \int g(x) \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} dx \\ &= \left(\int g(x) \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log f(x|\boldsymbol{\theta})}{\partial \theta_j} dx \right), \end{aligned} \tag{3.59}$$

$$\begin{aligned} J(\boldsymbol{\theta}) &= -\int g(x) \frac{\partial^2 \log f(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} dx \\ &= -\left(\int g(x) \frac{\partial^2 \log f(x|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} dx \right), \quad i, j = 1, \ldots, p. \end{aligned} \tag{3.60}$$

Outline of the Proof. By using a Taylor expansion of the first derivative of the maximum log-likelihood $\ell(\hat{\boldsymbol{\theta}}_n) = \sum_{\alpha=1}^n \log f(x_\alpha|\hat{\boldsymbol{\theta}}_n)$ around $\boldsymbol{\theta}_0$, we obtain

$$\mathbf{0} = \frac{\partial \ell(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}} = \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \cdots. \tag{3.61}$$

From the Taylor series expansion formula, the following approximation for the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_n$ can be obtained:

$$-\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}. \tag{3.62}$$

By the law of large numbers, when $n \to +\infty$, it can be shown that

$$-\frac{1}{n} \frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = -\frac{1}{n} \sum_{\alpha=1}^n \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \log f(x_\alpha|\boldsymbol{\theta}) \bigg|_{\boldsymbol{\theta}_0} \to J(\boldsymbol{\theta}_0), \tag{3.63}$$

where $|_{\boldsymbol{\theta}_0}$ is the value of the derivative at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

By virtue of the fact that when the p-dimensional random vector is written as $\boldsymbol{X}_\alpha = \partial \log f(X_\alpha|\boldsymbol{\theta})/\partial \boldsymbol{\theta}|_{\boldsymbol{\theta}_0}$ in the multivariate central limit theorem of

Remark 1 below and the right-hand side of (3.62) is $E_G[\boldsymbol{X}_\alpha] = 0$, $E_G[\boldsymbol{X}_\alpha \boldsymbol{X}_\alpha^T] = I(\boldsymbol{\theta}_0)$, it follows that

$$\sqrt{n}\frac{1}{n}\frac{\partial \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \sqrt{n}\frac{1}{n}\sum_{\alpha=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(x_\alpha|\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}_0} \to N_p(\boldsymbol{0}, I(\boldsymbol{\theta}_0)). \tag{3.64}$$

Then it follows from (3.62), (3.63), and (3.64) that, when $n \to +\infty$, we obtain

$$\sqrt{n}J(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \longrightarrow N_p(\boldsymbol{0}, I(\boldsymbol{\theta}_0)). \tag{3.65}$$

Therefore, the convergence in law

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \longrightarrow N_p\left(\boldsymbol{0}, J^{-1}(\boldsymbol{\theta}_0)I(\boldsymbol{\theta}_0)J^{-1}(\boldsymbol{\theta}_0)\right) \tag{3.66}$$

holds as n tends to infinity. In fact, it has been shown that this asymptotic normality holds even when the existence of higher-order derivatives is not assumed [Huber (1967)].

If the distribution $g(x)$ that generated the data is included in the class of parametric models $\{f(x|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subset R^p\}$, from Remark 2 shown below, the equality $I(\boldsymbol{\theta}_0) = J(\boldsymbol{\theta}_0)$ holds, and the asymptotic variance covariance matrix for $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ becomes

$$J^{-1}(\boldsymbol{\theta}_0)I(\boldsymbol{\theta}_0)J^{-1}(\boldsymbol{\theta}_0) = I(\boldsymbol{\theta}_0)^{-1}, \tag{3.67}$$

and the result (a) (iii) falls out.

Remark 1 (Multivariate central limit theorem) Let $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n, \ldots\}$ be a sequence of mutually independent random vectors drawn from a p-dimensional probability distribution and that have mean vector $E[\boldsymbol{X}_\alpha] = \boldsymbol{\mu}$ and variance covariance matrix $E[(\boldsymbol{X}_\alpha - \boldsymbol{\mu})(\boldsymbol{X}_\alpha - \boldsymbol{\mu})^T] = \Sigma$. Then the distribution of $\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{\mu})$ with respect to the sample mean vector $\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{\alpha=1}^{n} \boldsymbol{X}_\alpha$ converges in law to a p-dimensional normal distribution with mean vector $\boldsymbol{0}$ and variance covariance matrix Σ when $n \to +\infty$. In other words, when $n \to +\infty$, it holds that

$$\frac{1}{\sqrt{n}}\sum_{\alpha=1}^{n}(\boldsymbol{X}_\alpha - \boldsymbol{\mu}) = \sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}) \ \to \ N_p(\boldsymbol{0}, \Sigma). \tag{3.68}$$

Remark 2 (Relationship between the matrices $I(\boldsymbol{\theta})$ and $J(\boldsymbol{\theta})$) The following equality holds with respect to the second derivative of the log-likelihood function:

$$\begin{aligned} &\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(x|\boldsymbol{\theta}) \\ &\quad = \frac{\partial}{\partial \theta_i}\left\{\frac{\partial}{\partial \theta_j} \log f(x|\boldsymbol{\theta})\right\} \end{aligned}$$

$$
\begin{aligned}
&= \frac{\partial}{\partial \theta_i} \left\{ \frac{1}{f(x|\boldsymbol{\theta})} \frac{\partial}{\partial \theta_j} f(x|\boldsymbol{\theta}) \right\} \\
&= \frac{1}{f(x|\boldsymbol{\theta})} \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(x|\boldsymbol{\theta}) - \frac{1}{f(x|\boldsymbol{\theta})^2} \frac{\partial}{\partial \theta_i} f(x|\boldsymbol{\theta}) \frac{\partial}{\partial \theta_j} f(x|\boldsymbol{\theta}) \\
&= \frac{1}{f(x|\boldsymbol{\theta})} \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(x|\boldsymbol{\theta}) - \frac{\partial}{\partial \theta_i} \log f(x|\boldsymbol{\theta}) \frac{\partial}{\partial \theta_j} \log f(x|\boldsymbol{\theta}).
\end{aligned}
$$

By taking the expectation of the both sides with respect to the distribution $G(x)$, we obtain

$$
\begin{aligned}
&E_G \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(x|\boldsymbol{\theta}) \right] \\
&= E_G \left[\frac{1}{f(x|\boldsymbol{\theta})} \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(x|\boldsymbol{\theta}) \right] - E_G \left[\frac{\partial}{\partial \theta_i} \log f(x|\boldsymbol{\theta}) \frac{\partial}{\partial \theta_j} \log f(x|\boldsymbol{\theta}) \right].
\end{aligned}
$$

Hence, in general, we know that $I(\boldsymbol{\theta}) \neq J(\boldsymbol{\theta})$. However, if there exists a parameter vector $\boldsymbol{\theta}_0 \in \Theta$ such that $g(x) = f(x|\boldsymbol{\theta}_0)$, the first term on the right-hand side becomes

$$
\begin{aligned}
E_G \left[\frac{1}{f(x|\boldsymbol{\theta}_0)} \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(x|\boldsymbol{\theta}_0) \right] &= \int \frac{\partial^2}{\partial \theta_i \partial \theta_j} f(x|\boldsymbol{\theta}_0) dx \\
&= \frac{\partial^2}{\partial \theta_i \partial \theta_j} \int f(x|\boldsymbol{\theta}_0) dx = 0,
\end{aligned}
$$

and therefore the equality $I_{ij}(\theta_0) = J_{ij}(\theta_0)$ $(i, j = 1, 2, \ldots, p)$ holds; hence, we have $I(\boldsymbol{\theta}_0) = J(\boldsymbol{\theta}_0)$.

3.4 Information Criterion AIC

3.4.1 Log-Likelihood and Expected Log-Likelihood

The argument that has been presented thus far can be summarized as follows. When we build a model using data, we assume that the data $\boldsymbol{x}_n = \{x_1, x_2, \ldots, x_n\}$ are generated according to the true distribution $G(x)$ or $g(x)$. In order to capture the structure of the given phenomena, we assume a parametric model $\{f(x|\boldsymbol{\theta});\ \boldsymbol{\theta} \in \Theta \subset R^p\}$ having p-dimensional parameters, and we estimate it by using the maximum likelihood method. In other words, we construct a statistical model $f(x|\hat{\boldsymbol{\theta}})$ by replacing the unknown parameter $\boldsymbol{\theta}$ contained in the probability distribution by the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$. Our purpose here is to evaluate the goodness or badness of the statistical model $f(x|\hat{\boldsymbol{\theta}})$ thus constructed. We now consider the evaluation of a model from the standpoint of making a prediction.

Our task is to evaluate the expected goodness or badness of the estimated model $f(z|\hat{\boldsymbol{\theta}})$ when it is used to predict the independent future data $Z = z$

generated from the unknown true distribution $g(z)$. The K-L information described below is used to measure the closeness of the two distributions:

$$\begin{aligned} I\{g(z); f(z|\hat{\boldsymbol{\theta}})\} &= E_G\left[\log\left\{\frac{g(Z)}{f(Z|\hat{\boldsymbol{\theta}})}\right\}\right] \\ &= E_G\Big[\log g(Z)\Big] - E_G\left[\log f(Z|\hat{\boldsymbol{\theta}})\right], \end{aligned} \tag{3.69}$$

where the expectation is taken with respect to the unknown probability distribution $G(z)$ by fixing $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{x}_n)$.

In view of the properties of the K-L information, the larger the expected log-likelihood

$$E_G\left[\log f(Z|\hat{\boldsymbol{\theta}})\right] = \int \log f(z|\hat{\boldsymbol{\theta}})dG(z) \tag{3.70}$$

of the model is, the closer the model is to the true one. Therefore, in the definition of the information criterion, the crucial issue is to obtain a good estimator of the expected log-likelihood. One such estimator is

$$\begin{aligned} E_{\hat{G}}\left[\log f(Z|\hat{\boldsymbol{\theta}})\right] &= \int \log f(z|\hat{\boldsymbol{\theta}})d\hat{G}(z) \\ &= \frac{1}{n}\sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\boldsymbol{\theta}}), \end{aligned} \tag{3.71}$$

in which the unknown probability distribution G contained in the expected log-likelihood is replaced with an empirical distribution function $\hat{G}$. This is the log-likelihood of the statistical model $f(z|\hat{\boldsymbol{\theta}})$ or the maximum log-likelihood

$$\ell(\hat{\boldsymbol{\theta}}) = \sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\boldsymbol{\theta}}). \tag{3.72}$$

It is worth noting here that the estimator of the expected log-likelihood $E_G[\log f(Z|\hat{\boldsymbol{\theta}})]$ is $n^{-1}\ell(\hat{\boldsymbol{\theta}})$ and that the log-likelihood $\ell(\hat{\boldsymbol{\theta}})$ is an estimator of $nE_G[\log f(Z|\hat{\boldsymbol{\theta}})]$.

3.4.2 Necessity of Bias Correction for the Log-Likelihood

In practical situations, it is difficult to precisely capture the true structure of given phenomena from a limited number of observed data. For this reason, we construct several candidate statistical models based on the observed data at hand and select the model that most closely approximates the mechanism of the occurrence of the phenomena. In this subsection, we consider the situation in which multiple models $\{f_j(z|\boldsymbol{\theta}_j);\ j = 1, 2, \ldots, m\}$ exist, and the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_j$ has been obtained for the parameters of the model, $\boldsymbol{\theta}_j$.

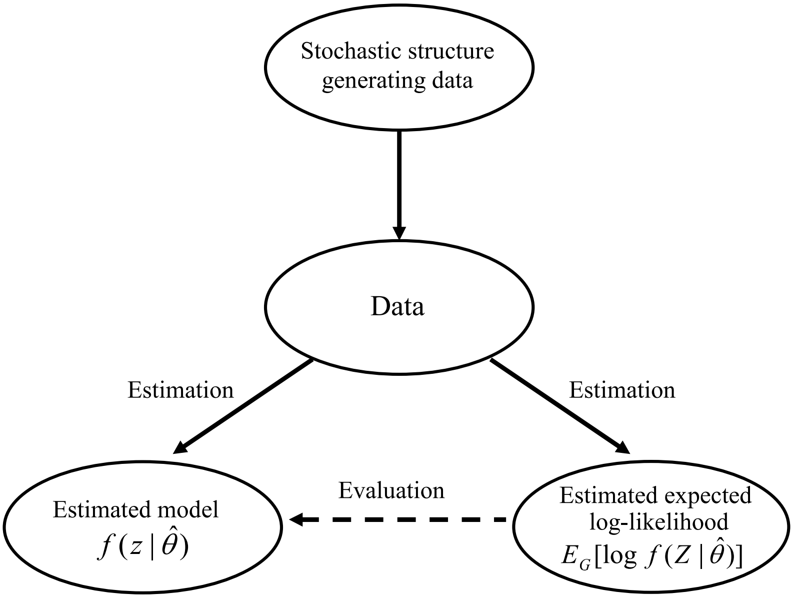
Stochastic structure generating data
Data
Estimation
Estimation
Evaluation
Estimated model
$f(z\,|\,\hat{\theta})$
Estimated expected log-likelihood
$E_G[\log f(Z\,|\,\hat{\theta})]$

Fig. 3.5. Use of data in the estimations of the parameter of a model and of the expected log-likelihood.

From the foregoing argument, it appears that the goodness of the model specified by $\hat{\boldsymbol{\theta}}_j$, that is, the goodness of the maximum likelihood model $f_j(z|\hat{\boldsymbol{\theta}}_j)$, can be determined by comparing the magnitudes of the maximum log-likelihood $\ell_j(\hat{\boldsymbol{\theta}}_j)$. However, it is known that this approach does not provide a fair comparison of models, since the quantity $\ell_j(\hat{\boldsymbol{\theta}}_j)$ contains a bias as an estimator of the expected log-likelihood $nE_G[\log f_j(z|\hat{\boldsymbol{\theta}}_j)]$, and the magnitude of the bias varies with the dimension of the parameter vector.

This result may seem to contradict the fact that generally $\ell(\boldsymbol{\theta})$ is a good estimator of $nE_G[\log f(Z|\boldsymbol{\theta})]$. However, as is evident from the process by which the log-likelihood in (3.71) was derived, the log-likelihood was obtained by estimating the expected log-likelihood by reusing the data $\boldsymbol{x}_n$ that were initially used to estimate the model in place of the future data (Figure 3.5). The use of the same data twice for estimating the parameters and for estimating the evaluation measure (the expected log-likelihood) of the goodness of the estimated model gives rise to the bias.

Relationship between log-likelihood and expected log-likelihood. Figure 3.6 shows the relationship between the expected log-likelihood function and the log-likelihood function

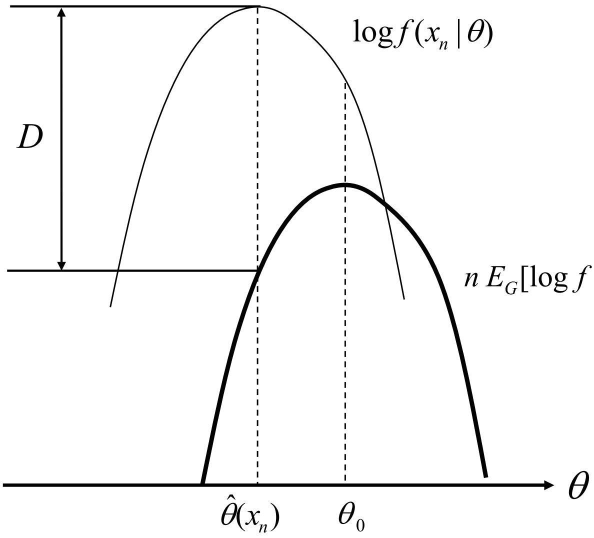

Fig. 3.6. Log-likelihood and expected log-likelihood.

$$n\eta(\theta) = nE_G\left[\log f(Z|\theta)\right], \qquad \ell(\theta) = \sum_{\alpha=1}^{n} \log f(x_\alpha|\theta), \tag{3.73}$$

for a model $f(x|\theta)$ with a one-dimensional parameter θ. The value of θ that maximizes the expected log-likelihood is the true parameter θ_0. On the other hand, the maximum likelihood estimator $\hat{\theta}(\boldsymbol{x}_n)$ is given as the maximizer of the log-likelihood function $\ell(\theta)$. The goodness of the model $f(z|\hat{\theta})$ defined by $\hat{\theta}(\boldsymbol{x}_n)$ should be evaluated in terms of the expected log-likelihood $E_G[\log f(Z|\hat{\theta})]$. However, in actuality, it is evaluated using the log-likelihood $\ell(\hat{\theta})$ that can be calculated from data. In this case, as indicated in Figure 3.6, the true criterion should give $E_G[\log f(Z|\hat{\theta})] \leq E_G[\log f(Z|\theta_0)]$ (see Subsection 3.1.1). However, in the log-likelihood, the relationship $\ell(\hat{\theta}) \geq \ell(\theta_0)$ always holds.

The log-likelihood function fluctuates depending on data, and the geometry between the two functions also varies; however, the above two inequalities always hold. If the two functions have the same form, then the log-likelihood is actually inferior to the extent that it appears to be better than the true model. The objective of the bias evaluation is to compensate for this phenomenon of reversal. Therefore, the prerequisite for a fair comparison of models is evaluation of and correction for the bias. In this subsection, we define an information criterion as a bias-corrected log-likelihood of the model.

Let us assume that n observations $\boldsymbol{x}_n$ generated from the true distribution $G(x)$ or $g(x)$ are realizations of the random variable $\boldsymbol{X}_n = (X_1, X_2, \cdots, X_n)^T$, and let

$$\ell(\hat{\boldsymbol{\theta}}) = \sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\boldsymbol{\theta}}(\boldsymbol{x}_n)) = \log f(\boldsymbol{x}_n|\hat{\boldsymbol{\theta}}(\boldsymbol{x}_n)) \tag{3.74}$$

represent the log-likelihood of the statistical model $f(z|\hat{\boldsymbol{\theta}}(\boldsymbol{x}_n))$ estimated by the maximum likelihood method. The bias of the log-likelihood as an estimator of the expected log-likelihood given in (3.70) is defined by

$$b(G) = E_{G(\boldsymbol{x}_n)}\left[\log f(\boldsymbol{X}_n|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n)) - nE_{G(z)}\left[\log f(Z|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n))\right]\right], \tag{3.75}$$

where the expectation $E_{G(\boldsymbol{x}_n)}$ is taken with respect to the joint distribution, $\prod_{\alpha=1}^{n} G(x_\alpha) = G(\boldsymbol{x}_n)$, of the sample $\boldsymbol{X}_n$, and $E_{G(z)}$ is the expectation on the true distribution $G(z)$. We see that the general form of the information criterion can be constructed by evaluating the bias and correcting for the bias of the log-likelihood as follows:

$$\begin{aligned}\mathrm{IC}(\boldsymbol{X}_n;\hat{G}) &= -2(\text{log-likelihood of statistical model} - \text{bias estimator})\\ &= -2\sum_{\alpha=1}^{n} \log f(X_\alpha|\hat{\boldsymbol{\theta}}) + 2\left\{\text{estimator for } b(G)\right\}. \end{aligned} \tag{3.76}$$

In general, the bias $b(G)$ can take various forms depending on the relationship between the true distribution generating the data and the specified model and on the method employed to construct a statistical model. In the following, we derive an information criterion for evaluating statistical models constructed by the maximum likelihood method.

3.4.3 Derivation of Bias of the Log-Likelihood

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is given as the p-dimensional parameter $\boldsymbol{\theta}$ that maximizes the log-likelihood function $\ell(\boldsymbol{\theta})= \sum_{\alpha=1}^{n} \log f(X_\alpha|\boldsymbol{\theta})$ or by solving the likelihood equation

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{\alpha=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_\alpha|\boldsymbol{\theta}) = \mathbf{0}. \tag{3.77}$$

Further, by taking the expectation, we obtain

$$E_{G(\boldsymbol{x}_n)}\left[\sum_{\alpha=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_\alpha|\boldsymbol{\theta})\right] = nE_{G(z)}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f(Z|\boldsymbol{\theta})\right]. \tag{3.78}$$

Therefore, for a continuous model, if $\boldsymbol{\theta}_0$ is a solution of the equation

$$E_{G(z)}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f(Z|\boldsymbol{\theta})\right] = \int g(z)\frac{\partial}{\partial \boldsymbol{\theta}} \log f(z|\boldsymbol{\theta})dz = \mathbf{0}, \tag{3.79}$$

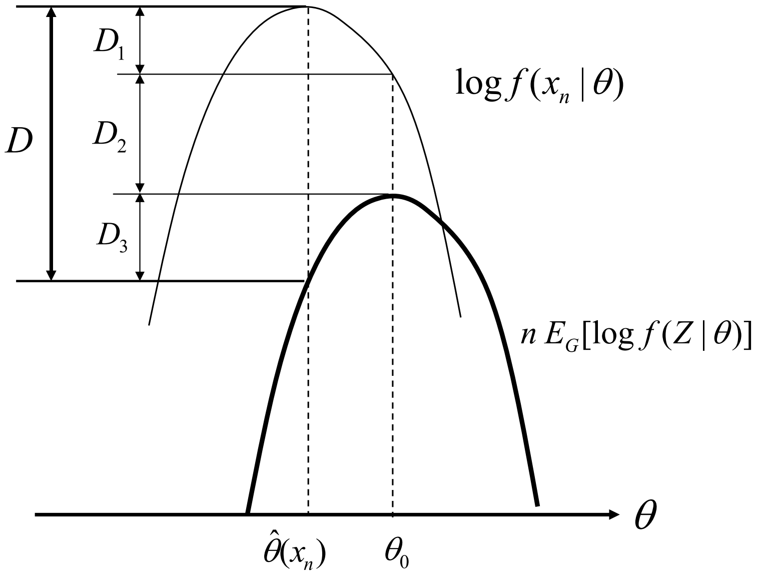

Fig. 3.7. Decomposition of the bias term.

it can be shown that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_0$ when $n \to +\infty$. For a discrete model, see (3.17).

Using the above results, we now evaluate the bias

$$b(G) = E_{G(\boldsymbol{x}_n)}\left[\log f(\boldsymbol{X}_n|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n)) - nE_{G(z)}\left[\log f(Z|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n))\right]\right] \tag{3.80}$$

when the expected log-likelihood is estimated using the log-likelihood of the statistical model. To this end, we first decompose the bias as follows (Figure 3.7):

$$\begin{aligned}
&E_{G(\boldsymbol{x}_n)}\left[\log f(\boldsymbol{X}_n|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n)) - nE_{G(z)}\left[\log f(Z|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n))\right]\right]\\
&\quad = E_{G(\boldsymbol{x}_n)}\left[\log f(\boldsymbol{X}_n|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n)) - \log f(\boldsymbol{X}_n|\boldsymbol{\theta}_0)\right]\\
&\qquad + E_{G(\boldsymbol{x}_n)}\left[\log f(\boldsymbol{X}_n|\boldsymbol{\theta}_0) - nE_{G(z)}\left[\log f(Z|\boldsymbol{\theta}_0)\right]\right]\\
&\qquad + E_{G(\boldsymbol{x}_n)}\left[nE_{G(z)}\left[\log f(Z|\boldsymbol{\theta}_0)\right] - nE_{G(z)}\left[\log f(Z|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n))\right]\right]\\
&\quad = D_1 + D_2 + D_3.
\end{aligned} \tag{3.81}$$

Notice that $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{X}_n)$ depends on the sample $\boldsymbol{X}_n$. In the next step, we calculate separately the three expectations D_1, D_2, and D_3.

(1) Calculation of D_2 The easiest case is the evaluation of D_2, which does not contain an estimator. It can easily be seen that

$$\begin{aligned} D_2 &= E_{G(\boldsymbol{x}_n)}\Big[\log f(\boldsymbol{X}_n|\boldsymbol{\theta}_0) - nE_{G(z)}\big[\log f(Z|\boldsymbol{\theta}_0)\big]\Big] \\ &= E_{G(\boldsymbol{x}_n)}\left[\sum_{\alpha=1}^{n}\log f(X_\alpha|\boldsymbol{\theta}_0)\right] - nE_{G(z)}\Big[\log f(Z|\boldsymbol{\theta}_0)\Big] \\ &= 0. \end{aligned} \tag{3.82}$$

This implies that in Figure 3.7, although D_2 varies randomly depending on the data, its expectation becomes 0.

(2) Calculation of D_3 First, we write

$$\eta(\hat{\boldsymbol{\theta}}) := E_{G(z)}\Big[\log f(Z|\hat{\boldsymbol{\theta}})\Big]. \tag{3.83}$$

By performing a Taylor series expansion of $\eta(\hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0$ given as a solution to (3.79), we obtain

$$\begin{aligned} \eta(\hat{\boldsymbol{\theta}}) &= \eta(\boldsymbol{\theta}_0) + \sum_{i=1}^{p}(\hat{\theta}_i - \theta_i^{(0)})\frac{\partial \eta(\boldsymbol{\theta}_0)}{\partial \theta_i} \\ &\quad + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}(\hat{\theta}_i - \theta_i^{(0)})(\hat{\theta}_j - \theta_j^{(0)})\frac{\partial^2 \eta(\boldsymbol{\theta}_0)}{\partial \theta_i \partial \theta_j} + \cdots, \end{aligned} \tag{3.84}$$

where $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_p)^T$ and $\boldsymbol{\theta}_0 = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)})^T$. Here, by virtue of the fact that $\boldsymbol{\theta}_0$ is a solution of (3.79), it holds that

$$\frac{\partial \eta(\boldsymbol{\theta}_0)}{\partial \theta_i} = E_{G(z)}\left[\frac{\partial}{\partial \theta_i}\log f(Z|\boldsymbol{\theta})\bigg|_{\boldsymbol{\theta}_0}\right] = 0, \quad i = 1, 2, \ldots, p, \tag{3.85}$$

where $|_{\boldsymbol{\theta}_0}$ is the value of the partial derivative at the point $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

Therefore, (3.84) can be approximated as

$$\eta(\hat{\boldsymbol{\theta}}) = \eta(\boldsymbol{\theta}_0) - \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T J(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \tag{3.86}$$

where $J(\boldsymbol{\theta}_0)$ is the $p \times p$ matrix given by

$$J(\boldsymbol{\theta}_0) = -E_{G(z)}\left[\frac{\partial^2 \log f(Z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}_0}\right] = -\int g(z)\frac{\partial^2 \log f(z|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}_0} dz \tag{3.87}$$

such that its $(a,b)^{th}$ element is given by

$$j_{ab} = -E_{G(z)}\left[\left.\frac{\partial^2 \log f(Z|\boldsymbol{\theta})}{\partial\theta_a\partial\theta_b}\right|_{\boldsymbol{\theta}_0}\right] = -\int g(z)\left.\frac{\partial^2 \log f(z|\boldsymbol{\theta})}{\partial\theta_a\partial\theta_b}\right|_{\boldsymbol{\theta}_0} dz. \tag{3.88}$$

Then, because D_3 is the expectation of $\eta(\boldsymbol{\theta}_0) - \eta(\hat{\boldsymbol{\theta}})$ with respect to $G(\boldsymbol{x}_n)$, we obtain approximately

$$\begin{aligned} D_3 &= E_{G(\boldsymbol{x}_n)}\left[nE_{G(z)}\left[\log f(Z|\boldsymbol{\theta}_0)\right] - nE_{G(z)}\left[\log f(Z|\hat{\boldsymbol{\theta}})\right]\right] \\ &= \frac{n}{2}E_{G(\boldsymbol{x}_n)}\left[(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T J(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right] \\ &= \frac{n}{2}E_{G(\boldsymbol{x}_n)}\left[\text{tr}\left\{J(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T\right\}\right] \\ &= \frac{n}{2}\text{tr}\left\{J(\boldsymbol{\theta}_0)E_{G(\boldsymbol{x}_n)}\left[(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T\right]\right\}. \end{aligned} \tag{3.89}$$

By substituting the (asymptotic) variance covariance matrix [see (3.58)]

$$E_{G(\boldsymbol{x}_n)}\left[(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^T\right] = \frac{1}{n}J(\boldsymbol{\theta}_0)^{-1}I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1} \tag{3.90}$$

of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ into (3.89), we have

$$D_3 = \frac{1}{2}\text{tr}\left\{I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}\right\}, \tag{3.91}$$

where $J(\boldsymbol{\theta}_0)$ is given in (3.87) and $I(\boldsymbol{\theta}_0)$ is the $p \times p$ matrix given by

$$\begin{aligned} I(\boldsymbol{\theta}_0) &= E_{G(z)}\left[\left.\frac{\partial \log f(Z|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \log f(Z|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_0}\right] \\ &= \int g(z)\left.\frac{\partial \log f(z|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \log f(z|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}\right|_{\boldsymbol{\theta}_0} dz. \end{aligned} \tag{3.92}$$

All that remains to do be done now is to calculate D_1.

(3) Calculation of D_1 By writing $\ell(\boldsymbol{\theta}) = \log f(\boldsymbol{X}_n|\boldsymbol{\theta})$ and by applying a Taylor series expansion around the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, we obtain

$$\ell(\boldsymbol{\theta}) = \ell(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\frac{\partial\ell(\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}} + \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\frac{\partial^2\ell(\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) + \cdots. \tag{3.93}$$

Here, the quantity $\hat{\boldsymbol{\theta}}$ satisfies the equation $\partial\ell(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta} = \mathbf{0}$ by virtue of the maximum likelihood estimator given as a solution of the likelihood equation $\partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}$.

We see that the quantity

$$\frac{1}{n}\frac{\partial^2 \ell(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \frac{1}{n}\frac{\partial^2 \log f(\boldsymbol{X}_n|\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \tag{3.94}$$

converges in probability to $J(\boldsymbol{\theta}_0)$ in (3.87) when n tends to infinity. This can be derived from the fact that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ and from the result of (3.63), which was obtained based on the law of large numbers. Using these results, we obtain the approximation

$$\ell(\boldsymbol{\theta}_0) - \ell(\hat{\boldsymbol{\theta}}) \approx -\frac{n}{2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T J(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) \tag{3.95}$$

for (3.93). Based on this result and the asymptotic variance covariance matrix (3.90) of the maximum likelihood estimator, D_1 can be calculated approximately as follows:

$$\begin{aligned} D_1 &= E_{G(\boldsymbol{x}_n)}\left[\log f(\boldsymbol{X}_n|\hat{\boldsymbol{\theta}}(\boldsymbol{X}_n)) - \log f(\boldsymbol{X}_n|\boldsymbol{\theta}_0)\right] \\ &= \frac{n}{2}E_{G(\boldsymbol{x}_n)}\left[(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T J(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})\right] \\ &= \frac{n}{2}E_{G(\boldsymbol{x}_n)}\left[\mathrm{tr}\left\{J(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T\right\}\right] \\ &= \frac{n}{2}\mathrm{tr}\left\{J(\boldsymbol{\theta}_0)E_{G(\boldsymbol{x}_n)}[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T]\right\} \\ &= \frac{1}{2}\mathrm{tr}\left\{I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}\right\}. \end{aligned} \tag{3.96}$$

Therefore, combining (3.82), (3.91), and (3.96), the bias resulting from the estimation of the expected log-likelihood using the log-likelihood of the model is asymptotically obtained as

$$\begin{aligned} b(G) &= D_1 + D_2 + D_3 \\ &= \frac{1}{2}\mathrm{tr}\left\{I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}\right\} + 0 + \frac{1}{2}\mathrm{tr}\left\{I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}\right\} \\ &= \mathrm{tr}\left\{I(\boldsymbol{\theta}_0)J(\boldsymbol{\theta}_0)^{-1}\right\}, \end{aligned} \tag{3.97}$$

where $I(\boldsymbol{\theta}_0)$ and $J(\boldsymbol{\theta}_0)$ are respectively given in (3.92) and (3.87).

(4) Estimation of bias Because the bias depends on the unknown probability distribution G that generated the data through $I(\boldsymbol{\theta}_0)$ and $J(\boldsymbol{\theta}_0)$, the bias must be estimated based on observed data. Let $\hat{I}$ and $\hat{J}$ be the consistent estimators of $I(\boldsymbol{\theta}_0)$ and $J(\boldsymbol{\theta}_0)$. In this case, we obtain an estimator of the bias $b(G)$ using

$$\hat{b} = \mathrm{tr}(\hat{I}\hat{J}^{-1}). \tag{3.98}$$

Thus, if we determine the asymptotic bias of the log-likelihood as an estimator of the expected log-likelihood of a statistical model, then the information criterion

$$\begin{aligned}\mathrm{TIC} &= -2\left\{\sum_{\alpha=1}^{n}\log f(X_\alpha|\hat{\boldsymbol{\theta}}) - \mathrm{tr}(\hat{I}\hat{J}^{-1})\right\}\\ &= -2\sum_{\alpha=1}^{n}\log f(X_\alpha|\hat{\boldsymbol{\theta}}) + 2\mathrm{tr}(\hat{I}\hat{J}^{-1}) \end{aligned} \tag{3.99}$$

is derived by correcting the bias of the log-likelihood of the model in the form shown in (3.76). This information criterion, which was investigated by Takeuchi (1976) and Stone (1977), is referred to as the "TIC."

Notice that the matrices $I(\boldsymbol{\theta}_0)$ and $J(\boldsymbol{\theta}_0)$ can be estimated by replacing the unknown probability distribution $G(z)$ or $g(z)$ by an empirical distribution function $\hat{G}(z)$ or $\hat{g}(z)$ based on the observed data as follows:

$$I(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{\alpha=1}^{n}\frac{\partial \log f(x_\alpha|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial \log f(x_\alpha|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\hat{\boldsymbol{\theta}}}, \tag{3.100}$$

$$J(\hat{\boldsymbol{\theta}}) = -\frac{1}{n}\sum_{\alpha=1}^{n}\frac{\partial^2 \log f(x_\alpha|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\bigg|_{\hat{\boldsymbol{\theta}}}. \tag{3.101}$$

The $(i,j)^{th}$ elements of these matrices are

$$I_{ij}(\hat{G}) = \frac{1}{n}\sum_{\alpha=1}^{n}\frac{\partial \log f(X_\alpha|\boldsymbol{\theta})}{\partial \theta_i}\frac{\partial \log f(X_\alpha|\boldsymbol{\theta})}{\partial \theta_j}\bigg|_{\hat{\boldsymbol{\theta}}}, \tag{3.102}$$

$$J_{ij}(\hat{G}) = -\frac{1}{n}\sum_{\alpha=1}^{n}\frac{\partial^2 \log f(X_\alpha|\boldsymbol{\theta})}{\partial \theta_i\partial \theta_j}\bigg|_{\hat{\boldsymbol{\theta}}}, \tag{3.103}$$

respectively.

3.4.4 Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) has played a significant role in solving problems in a wide variety of fields as a model selection criterion for analyzing actual data. The AIC is defined by

$$\mathrm{AIC} = -2(\text{maximum log-likelihood}) + 2(\text{number of free parameters}). \tag{3.104}$$

The number of free parameters in a model refers to the dimensions of the parameter vector $\boldsymbol{\theta}$ contained in the specified model $f(x|\boldsymbol{\theta})$.

The AIC is an evaluation criterion for the badness of the model whose parameters are estimated by the maximum likelihood method, and it indicates that the bias of the log-likelihood (3.80) approximately becomes the "number of free parameters contained in the model." The bias is derived under the assumption that the true distribution $g(x)$ is contained in the specified parametric model $\{f(x|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subset R^p\}$, that is, there exists a $\boldsymbol{\theta}_0 \in \Theta$ such that the equality $g(x) = f(x|\boldsymbol{\theta}_0)$ holds.

Let us now assume that the parametric model is $\{f(x|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta \subset R^p\}$ and that the true distribution $g(x)$ can be expressed as $g(x) = f(x|\boldsymbol{\theta}_0)$ for properly specified $\boldsymbol{\theta}_0 \in \Theta$. Under this assumption, the equality $I(\boldsymbol{\theta}_0) = J(\boldsymbol{\theta}_0)$ holds for the $p \times p$ matrix $J(\boldsymbol{\theta}_0)$ given in (3.87) and the $p \times p$ matrix $I(\boldsymbol{\theta}_0)$ given in (3.92), as stated in Remark 2 of Subsection 3.3.5. Therefore, the bias (3.97) of the log-likelihood is asymptotically given by

$$\begin{aligned} E_{G(\boldsymbol{x}_n)} & \left[\sum_{\alpha=1}^{n} \log f(X_\alpha|\hat{\boldsymbol{\theta}}) - nE_{G(z)} \log f(Z|\hat{\boldsymbol{\theta}}) \right] \\ &= \mathrm{tr}\left\{ I(\boldsymbol{\theta}_0) J(\boldsymbol{\theta}_0)^{-1} \right\} = \mathrm{tr}(I_p) = p, \end{aligned} \tag{3.105}$$

where I_p is the identity matrix of dimension p. Hence, the AIC

$$\mathrm{AIC} = -2 \sum_{\alpha=1}^{n} \log f(X_\alpha \mid \hat{\boldsymbol{\theta}}) + 2p \tag{3.106}$$

can be obtained by correcting the asymptotic bias p of the log-likelihood.

The AIC does not require any analytical derivation of the bias correction terms for individual problems and does not depend on the unknown probability distribution G, which removes fluctuations due to the estimation of the bias. Further, Akaike (1974) states that if the true distribution that generated the data exists near the specified parametric model, the bias associated with the log-likelihood of the model based on the maximum likelihood method can be approximated by the number of parameters. These attributes make the AIC a highly flexible technique from a practical standpoint.

Findley and Wei (2002) provided a derivation of AIC and its asymptotic properties for the case of vector time series regression model [see also Findley (1985), Bhansali (1986)]. Burnham and Anderson (2002) provided a nice review and explanation of the use of AIC in the model selection and evaluation problems [see also Linhart and Zucchini (1986), Sakamoto et al. (1986), Bozdogan (1987), Kitagawa and Gersch (1996), Akaike and Kitagawa (1998), McQuarrie and Tsai (1998), and Konishi (1999, 2002)]. Burnham and Anderson (2002) also discussed modeling philosophy and perspectives on model selection from an information-theoretic point of view, focusing on the AIC.

Example 10 (TIC for normal model) We assume a normal distribution for the model

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}. \tag{3.107}$$

We start by deriving TIC in (3.99) for any $g(x)$. Given n observations $\{x_1, x_2, \ldots, x_n\}$ that are generated from the true distribution $g(x)$, the statistical model is given by

$$f(x|\hat{\mu}, \hat{\sigma}^2) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left\{ -\frac{(x-\hat{\mu})^2}{2\hat{\sigma}^2} \right\}, \tag{3.108}$$

with the maximum likelihood estimators $\hat{\mu} = n^{-1}\sum_{\alpha=1}^{n} x_\alpha$ and $\hat{\sigma}^2 = n^{-1}\sum_{\alpha=1}^{n} (x_\alpha - \hat{\mu})^2$. Therefore, the bias associated with the estimation of the expected log-likelihood using the log-likelihood of the model,

$$E_G\left[\frac{1}{n}\sum_{\alpha=1}^{n} \log f(X_\alpha|\hat{\mu}, \hat{\sigma}^2) - \int g(z) \log f(z|\hat{\mu}, \hat{\sigma}^2) dz \right], \tag{3.109}$$

can be calculated using the matrix $I(\boldsymbol{\theta})$ of (3.92) and the matrix $J(\boldsymbol{\theta})$ of (3.87). This involves performing the following calculations:

For the log-likelihood function

$$\log f(x|\boldsymbol{\theta}) = -\frac{1}{2}\log\bigl(2\pi\sigma^2\bigr) - \frac{(x-\mu)^2}{2\sigma^2},$$

the expected value is obtained by

$$E_G[\log f(x|\theta)] = -\frac{1}{2}\log(2\pi\sigma^2) - \sigma^2(G) + \frac{(\mu - \mu(G))^2}{\sigma^2},$$

where $\mu(G)$ and $\sigma^2(G)$ are the mean and the variance of the true distribution $g(x)$, respectively. Therefore, the "true" parameters of the model are given by $\theta_0 = (\mu(G), \sigma^2(G))$.

The partial derivatives with respect to μ and σ^2 are

$$\begin{aligned}
&\frac{\partial}{\partial\mu}\log f(x|\boldsymbol{\theta}) = \frac{x-\mu}{\sigma^2}, \quad \frac{\partial}{\partial\sigma^2}\log f(x|\boldsymbol{\theta}) = -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4}, \\
&\frac{\partial^2}{\partial\mu^2}\log f(x|\boldsymbol{\theta}) = -\frac{1}{\sigma^2}, \quad \frac{\partial^2}{\partial\mu\partial\sigma^2}\log f(x|\boldsymbol{\theta}) = -\frac{x-\mu}{\sigma^4}, \\
&\frac{\partial^2}{(\partial\sigma^2)^2}\log f(x|\boldsymbol{\theta}) = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}.
\end{aligned}$$

Then the 2×2 matrices $I(\boldsymbol{\theta}_0)$ and $J(\boldsymbol{\theta}_0)$ are given by

$$\begin{aligned}
J(\boldsymbol{\theta}) &= -\begin{bmatrix} E_G\left[\frac{\partial^2}{\partial\mu^2}\log f(X|\boldsymbol{\theta})\right] & E_G\left[\frac{\partial^2}{\partial\sigma^2\partial\mu}\log f(X|\boldsymbol{\theta})\right] \\ E_G\left[\frac{\partial^2}{\partial\mu\partial\sigma^2}\log f(X|\boldsymbol{\theta})\right] & E_G\left[\frac{\partial^2}{(\partial\sigma^2)^2}\log f(X|\boldsymbol{\theta})\right] \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\sigma^2} & \frac{E_G[X-\mu]}{\sigma^4} \, \frac{E_G(X-\mu)^2}{\sigma^6} \\ \frac{E_G[X-\mu]}{\sigma^4} & \frac{E_G[(X-\mu)^2]}{\sigma^6} - \frac{1}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix},
\end{aligned}$$

$$I(\boldsymbol{\theta}) = E_G\left[\begin{pmatrix} \dfrac{X-\mu}{\sigma^2} \\ -\dfrac{1}{2\sigma^2}+\dfrac{(X-\mu)^2}{2\sigma^4} \end{pmatrix}\left(\frac{X-\mu}{\sigma^2}, -\frac{1}{2\sigma^2}+\frac{(X-\mu)^2}{2\sigma^4}\right)\right]$$

$$= E_G\begin{bmatrix} \dfrac{(X-\mu)^2}{\sigma^4} & -\dfrac{X-\mu}{2\sigma^4}+\dfrac{(X-\mu)^3}{2\sigma^6} \\ -\dfrac{X-\mu}{2\sigma^4}+\dfrac{(X-\mu)^3}{2\sigma^6} & \dfrac{1}{4\sigma^4}-\dfrac{(X-\mu)^2}{4\sigma^6}+\dfrac{(X-\mu)^4}{4\sigma^8} \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{1}{\sigma^2} & \dfrac{\mu_3}{2\sigma^6} \\ \dfrac{\mu_3}{2\sigma^6} & \dfrac{\mu_4}{4\sigma^8}-\dfrac{1}{4\sigma^4} \end{bmatrix},$$

where $\mu_j = E_G[(X-\mu)^j]$ $(j=1,2,\ldots)$ is the jth-order centralized moment of the true distribution $g(x)$. We note here that, in general, $I(\boldsymbol{\theta}_0) \neq J(\boldsymbol{\theta}_0)$.

From the above preparation, the bias correction term can be calculated as follows:

$$I(\boldsymbol{\theta})J(\boldsymbol{\theta})^{-1} = \begin{bmatrix} \dfrac{1}{\sigma^2} & \dfrac{\mu_3}{2\sigma^6} \\ \dfrac{\mu_3}{2\sigma^6} & \dfrac{\mu_4}{4\sigma^8}-\dfrac{1}{4\sigma^4} \end{bmatrix}\begin{bmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \dfrac{\mu_3}{\sigma^2} \\ \dfrac{\mu_3}{2\sigma^4} & \dfrac{\mu_4}{2\sigma^4}-\dfrac{1}{2} \end{bmatrix}.$$

Therefore,

$$\mathrm{tr}\left\{I(\boldsymbol{\theta})J(\boldsymbol{\theta})^{-1}\right\} = 1+\frac{\mu_4}{2\sigma^4}-\frac{1}{2} = \frac{1}{2}\left(1+\frac{\mu_4}{\sigma^4}\right).$$

This result is generally not equal to the number of parameters, i.e. two in this case. However, if there exists a $\boldsymbol{\theta}_0$ that satisfies $f(x|\boldsymbol{\theta}_0) = g(x)$, then $g(x)$ is a normal distribution, and we have $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$. Hence, it follows that

$$\frac{1}{2}+\frac{\mu_4}{2\sigma^4} = \frac{1}{2}+\frac{3\sigma^4}{2\sigma^4} = \frac{1}{2}+\frac{3}{2} = 2.$$

Given the data, the estimator for the bias is obtained using

$$\frac{1}{n}\mathrm{tr}(\hat{I}\hat{J}^{-1}) = \frac{1}{n}\left\{\frac{1}{2}+\frac{\hat{\mu}_4}{2\hat{\sigma}^4}\right\}, \tag{3.110}$$

where $\hat{\sigma}^2 = n^{-1}\sum_{\alpha=1}^{n}(x_\alpha-\overline{x})^2$ and $\hat{\mu}_4 = n^{-1}\sum_{\alpha=1}^{n}(x_\alpha-\overline{x})^4$. Consequently, the information criteria TIC and AIC are given by the following formulas, respectively:

$$\mathrm{TIC} = -2\sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\mu},\hat{\sigma}^2) + 2\left(\frac{1}{2} + \frac{\hat{\mu}_4}{2\hat{\sigma}^4}\right), \tag{3.111}$$

$$\mathrm{AIC} = -2\sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\mu},\hat{\sigma}^2) + 2 \times 2, \tag{3.112}$$

where the maximum log-likelihood is given by

$$\sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\mu},\hat{\sigma}^2) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{n}{2}.$$

Table 3.3. Change of the bias correction term $\frac{1}{2}(1+\hat{\mu}_4/\hat{\sigma}^4)$ of the TIC when the true distribution is assumed to be a mixed normal distribution ($\xi_1 = \xi_2 = 0, \sigma_1^2 = 1, \sigma_2^2 = 3$); ε denotes the mixing ratio and n is the number of observations. The mean and standard deviation of the estimated bias correction term for each value of ε and n are shown.

ε	$n = 25$	$n = 100$	$n = 400$	$n = 1600$
0.00	1.89 (0.37)	1.97 (0.23)	1.99 (0.12)	2.00 (0.06)
0.01	2.03 (0.71)	2.40 (1.25)	2.67 (1.11)	2.78 (0.71)
0.02	2.14 (0.83)	2.73 (1.53)	3.18 (1.38)	3.33 (0.81)
0.05	2.44 (1.13)	3.45 (1.78)	4.02 (1.35)	4.24 (0.80)
0.10	2.74 (1.24)	3.87 (1.56)	4.42 (1.09)	4.60 (0.60)
0.15	2.87 (1.18)	3.96 (1.34)	4.38 (0.89)	4.49 (0.46)
0.20	2.91 (1.09)	3.84 (1.12)	4.16 (0.69)	4.24 (0.37)
0.30	2.85 (0.94)	3.48 (0.82)	3.67 (0.48)	3.73 (0.25)
0.40	2.68 (0.80)	3.14 (0.65)	3.26 (0.37)	3.29 (0.19)
0.50	2.52 (0.69)	2.84 (0.50)	2.92 (0.28)	2.95 (0.15)
0.60	2.37 (0.60)	2.61 (0.44)	2.67 (0.24)	2.68 (0.12)
0.70	2.22 (0.53)	2.40 (0.36)	2.45 (0.20)	2.46 (0.10)
0.80	2.10 (0.47)	2.23 (0.30)	2.27 (0.16)	2.28 (0.08)
0.90	1.98 (0.41)	2.09 (0.26)	2.12 (0.14)	2.12 (0.07)
1.00	1.88 (0.36)	1.97 (0.23)	1.99 (0.12)	2.00 (0.06)

Example 11 (TIC for normal model versus mixture of two normal distributions) Let us assume that the true distribution generating data is a mixture of two normal distributions

$$g(x) = (1-\varepsilon)\phi(x|\xi_1,\sigma_1^2) + \varepsilon\phi(x|\xi_2,\sigma_2^2) \qquad (0 \le \varepsilon \le 1), \tag{3.113}$$

where $\phi(x|\xi_i,\sigma_i^2)$ $(i = 1, 2)$ is the probability density function of the normal distribution with mean ξ_i and variance σ_i^2. We assume the normal model

$N(\mu, \sigma^2)$ for the model. Table 3.3 shows the mean and the standard deviation of 10,000 simulation runs of the TIC bias correction term $\frac{1}{2}(1 + \hat{\mu}_4/\hat{\sigma}^4)$ in (3.111), which were obtained by varying the mixing ratio and the number of observations in a mixed normal distribution. When n is small and ε is equal to either 0 or 1, the result is smaller than the bias correction term 2 of the AIC. The bias correction term is maximized when the value of ε is in the neighborhood of 0.1 to 0.2. Notice that in the region in which the correction term in the TIC is large, the standard deviation is also large.

Table 3.4. Estimated bias correction terms of TIC and their standard deviations when normal distribution models are fitted to simulated data from the t-distribution.

df	$n = 25$	$n = 100$	$n = 400$	$n = 1,600$
∞	1.89 (0.37)	1.98 (0.23)	2.00 (0.12)	2.00 (0.06)
9	2.12 (0.62)	2.42 (0.69)	2.54 (0.52)	2.58 (0.34)
8	2.17 (0.66)	2.51 (0.82)	2.67 (0.86)	2.73 (0.63)
7	2.21 (0.72)	2.64 (0.99)	2.85 (1.05)	2.95 (0.91)
6	2.29 (0.81)	2.85 (1.43)	3.20 (1.81)	3.36 (1.46)
5	2.43 (1.00)	3.21 (1.96)	3.87 (3.21)	4.28 (4.12)
4	2.67 (1.23)	3.94 (3.01)	5.49 (6.37)	7.46 (15.96)
3	3.06 (1.62)	5.72 (5.38)	10.45 (14.71)	19.79 (41.12)
2	4.01 (2.32)	10.54 (9.39)	30.88 (35.67)	101.32 (138.74)
1	6.64 (3.17)	25.27 (13.94)	100.14 (56.91)	404.12 (232.06)

Example 12 (TIC for normal model versus t-distribution) Table 3.4 shows the means and the standard deviations of the estimated bias correction term of the TIC, $\frac{1}{2}(1 + \hat{\mu}_4/\hat{\sigma}^4)$ in (3.111), when it is assumed that the true distribution is the t-distribution with degrees of freedom df,

$$g(x|df) = \frac{\Gamma\left(\frac{df+1}{2}\right)}{\sqrt{df\pi}\,\Gamma\left(\frac{df}{2}\right)}\left(1 + \frac{x^2}{df}\right)^{-\frac{1}{2}(df+1)}, \tag{3.114}$$

which were obtained by repeating 10,000 simulation runs. Four data lengths (n= 25, 100, 400, and 1,600) and 10 different values for the degrees of freedom [1 to 9 and the normal distribution ($df = \infty$)] were examined.

When the degrees of freedom df is small and the number of observations is large, the results differ significantly from the correction term 2 of the AIC. Notice that in this case, the standard deviation is also extremely large, exceeding the value of the bias in some cases.

Example 13 (Polynomial regression models) Assume that the following 20 observations, (x, y), are observed in experiments (Figure 3.8):

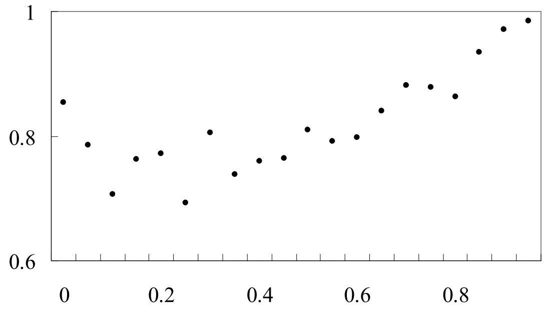

Fig. 3.8. Twenty observations used for polynomial regression models.

(0.00, 0.854),	(0.05, 0.786),	(0.10, 0.706),	(0.15, 0.763),	(0.20, 0.772),
(0.25, 0.693),	(0.30, 0.805),	(0.35, 0.739),	(0.40, 0.760),	(0.45, 0.764),
(0.50, 0.810),	(0.55, 0.791),	(0.60, 0.798),	(0.65, 0.841),	(0.70, 0.882),
(0.75, 0.879),	(0.80, 0.863),	(0.85, 0.934),	(0.90, 0.971),	(0.95, 0.985).

A polynomial regression model is then fitted to these 20 observations; specifically, to the following model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2). \tag{3.115}$$

Here we write $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p, \sigma^2)^T$ and when data $\{(y_\alpha, x_\alpha), \alpha = 1, \ldots, n\}$ are given, the log-likelihood function can be written as

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{\alpha=1}^{n}\Big(y_\alpha - \sum_{j=0}^{p}\beta_j x_\alpha^j\Big)^2. \tag{3.116}$$

Therefore, the maximum likelihood estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ for the coefficients can be obtained by minimizing the following term:

$$\sum_{\alpha=1}^{n}\Big(y_\alpha - \sum_{j=0}^{p}\beta_j x_\alpha^j\Big)^2. \tag{3.117}$$

In addition, the maximum likelihood estimator of the error variance is given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{\alpha=1}^{n}\Big(y_\alpha - \sum_{j=0}^{p}\hat{\beta}_j x_\alpha^j\Big)^2. \tag{3.118}$$

By substituting this expression into (3.116), we obtain the maximum log-likelihood

$$\ell(\hat{\boldsymbol{\theta}}) = -\frac{n}{2}\log\big(2\pi\hat{\sigma}^2\big) - \frac{n}{2}. \tag{3.119}$$

Further, because the number of parameters contained in this model is $p + 2$, that is, for $\beta_0, \beta_1, \ldots, \beta_p$ and σ^2, the AIC for evaluating the p^{th} order polynomial regression model is given by

Table 3.5. Results of estimating polynomial regression models.

Order	$\hat{\sigma}^2$	Log-Likelihood	AIC	AIC Difference
—	0.678301	−24.50	50.99	126.49
0	0.006229	22.41	−40.81	34.68
1	0.002587	31.19	−56.38	19.11
2	0.000922	41.51	−75.03	0.47
3	0.000833	42.52	−75.04	0.46
4	0.000737	43.75	−75.50	—
5	0.000688	44.44	−74.89	0.61
6	0.000650	45.00	−74.00	1.49
7	0.000622	45.45	−72.89	2.61
8	0.000607	45.69	−71.38	4.12
9	0.000599	45.83	−69.66	5.84

$$\mathrm{AIC}_p = n(\log 2\pi + 1) + n \log \hat{\sigma}^2 + 2(p+2). \tag{3.120}$$

Table 3.5 summarizes the results obtained by fitting polynomials up to order nine to this set of data. As the order increases, the residual variance reduces, and the log-likelihood increases monotonically. The AIC attains a minimum at $p = 4$, and the model

$$\begin{aligned} y_j &= 0.835 - 1.068x_j + 3.716x_j^2 - 4.573x_j^3 + 2.141x_j^4 + \varepsilon_j, \\ \varepsilon_j &\sim N(0, 0.737 \times 10^{-3}), \end{aligned} \tag{3.121}$$

is selected as the best model.

In order to demonstrate the importance of order selection in a regression model, Figure 3.9 shows the results of running Monte Carlo experiments. Using different random numbers, 20 observations were generated according to (3.115), and using the data, 2nd-, 4th-, and 9th-order polynomials were estimated. Figure 3.9 shows the 10 regression curves that were obtained by repeating these operations 10 times, along with the “true” regression polynomial that was used for generating the data. In the case of the 2nd-order polynomial regression model, while the width of the fluctuations is small, the low order of the polynomial results in a large bias in the regression curves. For the 4th-order polynomial, the 10 estimated values cover the true regression polynomial. By contrast, for the 9th-order polynomial, although the true regression polynomial is covered, the large fluctuations indicate that the estimated values are highly unstable.

Example 14 (Factor analysis model) Suppose that $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ is an observable random vector with mean vector $\boldsymbol{\mu}$ and variance covariance matrix Σ. The factor analysis model is

$$\boldsymbol{x} = \boldsymbol{\mu} + L\boldsymbol{f} + \boldsymbol{\varepsilon}, \tag{3.122}$$

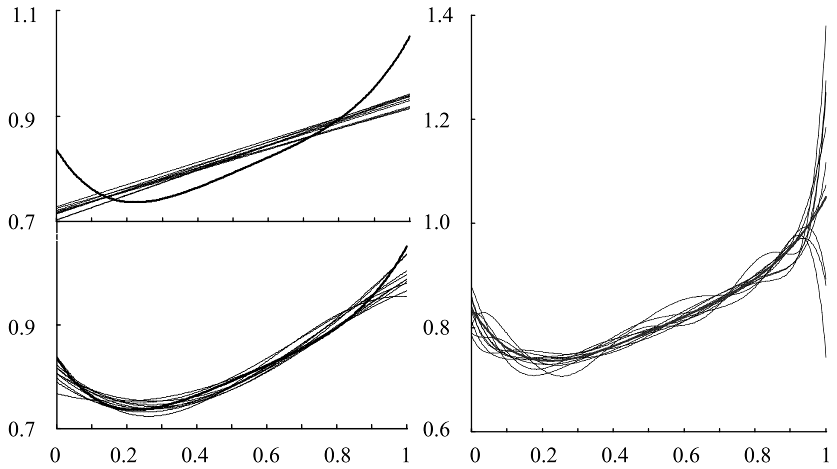

Fig. 3.9. Fluctuations in estimated polynomials for (3.115). Upper left: $p = 2$; lower-left: $p = 4$; right: $p = 9$.

where L is a $p \times m$ matrix of factor loadings, and $\boldsymbol{f} = (f_1, \ldots, f_m)^T$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^T$ are unobservable random vectors. The elements of $\boldsymbol{f}$ are called common factors while the elements of $\boldsymbol{\varepsilon}$ are referred to as specific or unique factors. It is assumed that

$$\begin{aligned} E[\boldsymbol{f}] &= \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{f}) = E[\boldsymbol{f}\boldsymbol{f}^T] = I_m, \\ E[\boldsymbol{\varepsilon}] &= \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \Psi = \mathrm{diag}[\psi_1, \cdots, \psi_p], \\ \mathrm{Cov}(\boldsymbol{f}, \boldsymbol{\varepsilon}) &= E[\boldsymbol{f}\boldsymbol{\varepsilon}^T] = 0, \end{aligned} \tag{3.123}$$

where I_m is the identity matrix of order m and Ψ is a $p \times p$ diagonal matrix with i^{th} diagonal element ψ_i (> 0). It then follows from (3.122) and (3.123) that Σ can be expressed as

$$\Sigma = LL^T + \Psi. \tag{3.124}$$

Assume that the common factors $\boldsymbol{f}$ and the specific factors $\boldsymbol{\varepsilon}$ are normally distributed. Let $\overline{\boldsymbol{x}}$ and S be, respectively, the sample mean vector and sample covariance matrix based on a set of n observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ on $\boldsymbol{x}$. It is known [see, for example, Lawley and Maxwell (1971) and Anderson (2003)] that the maximum likelihood estimates, $\hat{L}$ and $\hat{\Psi}$, of the matrix L of factor loadings and the covariance matrix Ψ of specific factors are obtained by minimizing the discrepancy function

$$Q(L, \Psi) = \log|\Sigma| - \log|S| + \mathrm{tr}\left(\Sigma^{-1}S\right) - p, \tag{3.125}$$

subject to the condition that $L^T\Psi^{-1}L$ is a diagonal matrix. Then, the AIC is defined by

$$\mathrm{AIC} = n\left\{p\log(2\pi) + \log|\hat{\Sigma}| + \mathrm{tr}\left(\hat{\Sigma}^{-1}S\right)\right\} + 2\left\{p(m+1) - \frac{1}{2}m(m-1)\right\}, \tag{3.126}$$

where $\hat{\Sigma} = \hat{L}\hat{L}^T + \hat{\Psi}$.

The use of the AIC in the factor analysis model was considered by Akaike (1973, 1987). Ichikawa and Konishi (1999) derived the TIC for a covariance structure analysis model and investigated the performance of three information criteria, namely the AIC, the TIC, and the bootstrap information criteria (introduced in Chapter 8). The use of AIC-type criteria for selecting variables in principal component, canonical correlation, and discriminant analyses was discussed, in relation to the likelihood ratio tests, by Fujikoshi (1985) and Siotani et al. (1985, Chapter 13).

3.5 Properties of MAICE

The estimators and models selected by minimizing the AIC are referred to as MAICE (minimum AIC estimators). In this section, we discuss several topics related to the properties of MAICE.

3.5.1 Finite Correction of the Information Criterion

In Section 3.4, we derived the AIC for general statistical models estimated using the maximum likelihood method. In contrast, information criterion for particular models such as normal distribution models can be derived directly and analytically by calculating the bias, without having to resort to asymptotic theories such as the Taylor series expansion or the asymptotic normality. Let us first consider a simple normal distribution model, $N(\mu, \sigma^2)$.

Since the logarithm of the probability density function is

$$\log f(x|\mu, \sigma^2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2},$$

the log-likelihood of the model based on the data, $\boldsymbol{x}_n = \{x_1, x_2, \ldots, x_n\}$, is given by

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{\alpha=1}^{n}(x_\alpha - \mu)^2.$$

By substituting the maximum likelihood estimators

$$\hat{\mu} = \frac{1}{n}\sum_{\alpha=1}^{n}x_\alpha, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{\alpha=1}^{n}(x_\alpha - \hat{\mu})^2,$$

into this expression, we obtain the maximum log-likelihood

$$\ell(\hat{\mu}, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{n}{2}.$$

If the data set is obtained from the same normal distribution $N(\mu, \sigma^2)$, then the expected log-likelihood is given by

$$E_G\left[\log f(Z|\hat{\mu}, \hat{\sigma}^2)\right] = -\frac{1}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\left\{\sigma^2 + (\mu - \hat{\mu})^2\right\},$$

where $G(z)$ is the distribution function of the normal distribution $N(\mu, \sigma^2)$. Therefore, the difference between the two quantities is

$$\ell(\hat{\mu}, \hat{\sigma^2}) - nE_G\left[\log f(Z|\hat{\mu}, \hat{\sigma^2})\right] = \frac{n}{2\hat{\sigma}^2}\left\{\sigma^2 + (\mu - \hat{\mu})^2\right\} - \frac{n}{2}.$$

By taking the expectation with respect to the joint distribution of n observations distributed as the normal distribution $N(\mu, \sigma^2)$, and using

$$E_G\left[\frac{\sigma^2}{\hat{\sigma}^2(\boldsymbol{x}_n)}\right] = \frac{n}{n-3}, \quad E_G\left[\{\mu - \hat{\mu}(\boldsymbol{x}_n)\}^2\right] = \frac{\sigma^2}{n},$$

we obtain the bias correction term for a finite sample as

$$b(G) = \frac{n}{2}\frac{n}{(n-3)\sigma^2}\left(\sigma^2 + \frac{\sigma^2}{n}\right) - \frac{n}{2} = \frac{2n}{n-3}. \tag{3.127}$$

Here we used the fact that for a χ^2 random variable with degrees of freedom r, χ^2_r, we have $E[1/\chi^2_r] = 1/(r-2)$. Therefore, the information criterion (IC) for the normal distribution model is given by

$$\mathrm{IC} = -2\ell(\hat{\mu}, \hat{\sigma}^2) + \frac{4n}{n-3}. \tag{3.128}$$

Table 3.6 shows changes in this bias term $b(G)$ with respect to several values of n. This table shows that $b(G)$ approaches the correction term 2 of the AIC as the number of observations increases.

Table 3.6. Changes of the bias $b(G)$ for normal distribution model as the number of the observations increases.

n	4	6	8	12	18	25	50	100
$b(G)$	8.0	4.0	3.2	2.7	2.4	2.3	2.1	2.1

The topic of a finite correction of the AIC for more general Gaussian linear regression models will be discussed in Subsection 7.2.2.

3.5.2 Distribution of Orders Selected by AIC

Let us consider the problem of order selection in an autoregressive model

$$y_n = \sum_{j=1}^{m} a_j y_{n-j} + \varepsilon_n, \quad \varepsilon_n \sim N(0,\sigma^2). \tag{3.129}$$

In this case, an asymptotic distribution of the number of orders is obtained when the number of orders is selected using the AIC minimization method [Shibata (1976)]. We now define p_j and q_j $(j = 1,\ldots,M)$ by setting $\alpha_i = \Pr(\chi_i^2 > 2i)$, $p_0 = q_0 = 1$, with respect to the χ^2-variate with i degrees of freedom according to the following equations:

$$p_j = \sum \left\{ \prod_{i=1}^{j} \frac{1}{r_i!} \left(\frac{\alpha_i}{i} \right)^{r_i} \right\}, \tag{3.130}$$

$$q_j = \sum \left\{ \prod_{i=1}^{j} \frac{1}{r_i!} \left(\frac{1-\alpha_i}{i} \right)^{r_i} \right\}, \tag{3.131}$$

where $\sum$ is the sum of all combinations of $(r_1,\ldots,r_j)$ that satisfy the equation $r_1 + 2r_2 + \cdots + nr_j = j$. In this case, according to Shibata (1976), if the AR model with order m_0 is the true model, and if the order $0 \le m \le M$ of the AR model is selected using the AIC, then the asymptotic distribution of $\hat{m}$ can be obtained as

$$\lim_{n\to+\infty} \Pr(\hat{m} = m) = \begin{cases} p_{m-m_0}\, q_{M-m} & \text{for } m_0 \le m \le M, \\ 0 & \text{for } \quad m < m_0. \end{cases} \tag{3.132}$$

This result shows that the probability of selecting the true order using the minimum AIC procedure is not unity even as $n \to +\infty$. In other words, the order selection using the AIC is not consistent. At the same time, since the distribution of the selected order has an asymptotic distribution, the result indicates that it will not spread as n increases.

In general, under the assumptions that the true model is of finite dimension and it is included in the class of candidate models, a criterion that identifies the correct model asymptotically with probability one is said to be consistent. The consistency has been investigated by Shibata (1976, 1981), Nishii (1984), Findley (1985), etc. A review of consistency on model selection criteria was provided by Rao and Wu (2001) and Burnham and Anderson (2002, Section 6.3).

Example 15 (Order selection in linear regression models) Figure 3.10 shows the distribution of the number of explanatory variables that are selected using the AIC for the case of an ordinary regression model

$$y_i = a_1 x_{i1} + \cdots + a_k x_{ik} + \varepsilon_i, \;\; \varepsilon_i \sim N(0,\sigma^2).$$

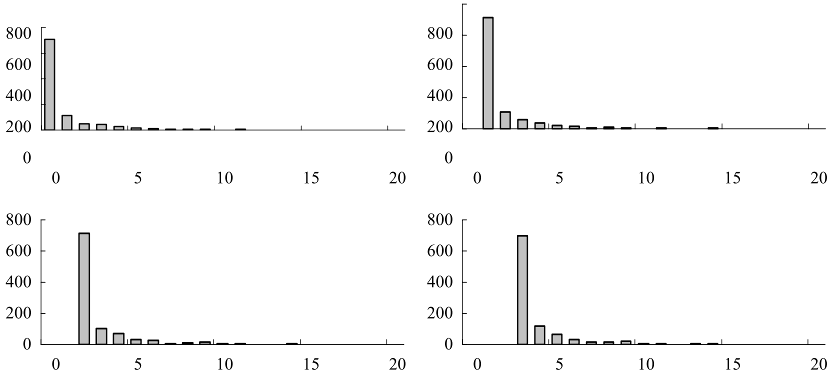

Fig. 3.10. Distributions of orders selected by AIC. The upper left, upper right, lower left, and lower right plots represent the cases in which the true order is 0, 1, 2, and 3, respectively.

It will be demonstrated by simulations that even for the ordinary regression case, we can obtain results that are qualitatively similar to those for the autoregression case.

For simplicity, we assume that x_{ij} $(j = 1, \dots, 20,\ i = 1, \dots, n)$ are orthonormal variables. We also assume that the true model that generates data is given by $\sigma^2 = 0.01$ and

$$a_j^* = \begin{cases} 0.7^j & \text{for } j = 1, \dots, k^*, \\ 0 & \text{for } j = k^* + 1, \dots, 20\,. \end{cases} \tag{3.133}$$

Figure 3.10 shows the distributions of orders obtained by generating data with $n = 400$ and by repeating 1,000 times the process of selecting orders using the AIC. The upper left plot represents the case in which the true order is defined as $k^* = 0$. Similarly, the upper right, lower left, and lower right plots represent the cases for which $k^* = 1, 2, 3$, respectively. These results also indicate that when the number of observations involved is relatively large (for example, $n = 400$) for both the regression model and autoregressive models, the probability with which the true order is obtained is approximately 0.7, which means that the order is overestimated with a probability of 0.3. In this distribution, varying the true order k^* only shifts the location of the maximum probability to the right, while only slightly modifying the shape of the distribution.

Figure 3.11 shows the results of examining changes in distribution as a function of the number of observations for the case $k^* = 1$. The graph on the left shows the case when $n = 100$, while the graph on the right shows the case when $n = 1,600$. The results suggest that when the true order is a finite number, the distribution of orders converges to a certain distribution when the size of n becomes large. Figure 3.12 shows the case for $k^* = 20$, in which

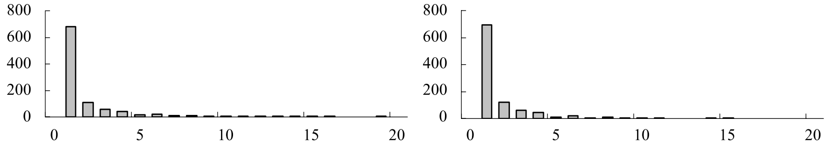

Fig. 3.11. Change in distribution of order selected by the AIC, for different number of observations. Left graph: $n = 100$; right graph: $n = 1,600$.

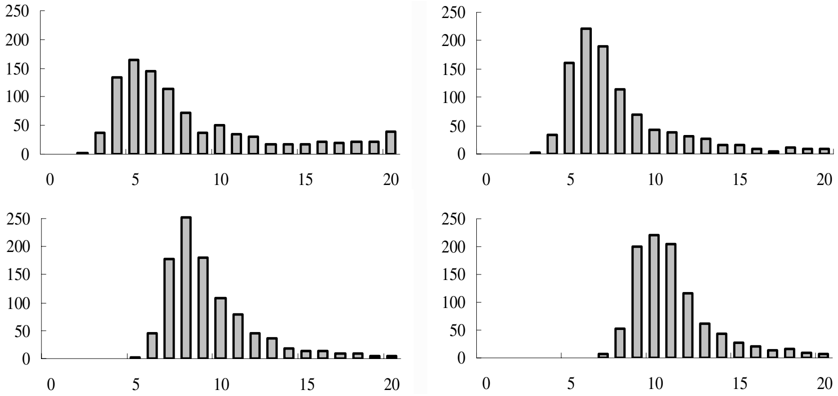

Fig. 3.12. Distributions of orders selected by AIC when the true coefficient decays with the order. The upper left, upper right, lower left, and lower right graphs represent the cases in which the number of observations is 50, 100, 400, and 1,600, respectively.

all of the coefficients are nonzero. The results indicate that the distribution's mode shifts to the right as the number of observations, n, increases and that when complex phenomena are approximated using a relatively simple model, the order selected by the AIC increases with the number of observations.

3.5.3 Discussion

Here we summarize several points regarding the selection of a model using the AIC. The AIC has been criticized because it does not yield a consistent estimator with respect to the selection of orders. Such an argument is frequently misunderstood, and we attempt to clarify these misunderstandings in the following.

(1) First, the objective of our modeling is to obtain a "good" model, rather than a "true" model. If one recalls that statistical models are approximations of complex systems toward certain objectives, the task of estimating the true order is obviously not an appropriate goal. A true model or order can be defined explicitly only in a limited number of situations, such as

when running simulation experiments. From the standpoint that a model is an approximation of a complex phenomenon, the true order can be infinitely large.

(2) Even if a true finite order exists, the order of a good model is not necessarily equal to the true order. In situations where there are only a small number of observations, considering the instability of the parameters being estimated, the AIC reveals the possibility that a higher prediction accuracy can be obtained using models having lower orders.

(3) Shibata's (1976) results described in the previous section indicate that if the true order is assumed, the asymptotic distribution of orders selected by the AIC can be a fixed distribution that is determined solely by the maximum order and the true order of a family of models. This indicates that the AIC does not provide a consistent estimator of orders. It should be noted, however, that when the true order is finite, the distribution of orders that is selected does not vary when the number of observations is increased. It should also be noted that in this case, even if a higher order is selected, when the number of observations is large, each coefficient estimate of a regressor with an order greater than the true order converges to the true value 0 and that a consistent estimator can be obtained as a model.

(4) Although the information criterion makes automatic model selection possible, it should be noted that the model evaluation criterion is a relative evaluation criterion. This means that selecting a model using an information criterion is only a selection from a family of models that we have specified. Therefore, the critical task for us is to set up more appropriate models by making use of knowledge regarding that object.