2. MAXIMUM LIKELIHOOD ESTIMATION

2.1 Likelihood

Likelihood is one of the most fundamental concept in all types of statistical inference.

Definition 1 Suppose that $\mathbf{X}$ has density function or probability function $f(\mathbf{x};\boldsymbol{\theta})$. We have observed $\mathbf{X}=\mathbf{x}$. Then the likelihood function with observation $\mathbf{x}$ is defined as

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta};\mathbf{x}) = f(\mathbf{x};\boldsymbol{\theta}).$$

Density/probability function: a function of $\mathbf{x}$, specifying the distribution of random variable $\mathbf{X}$

Likelihood: a function of $\boldsymbol{\theta}$, reflecting information on $\boldsymbol{\theta}$ contained in observation $\mathbf{x}$

Note. A likelihood function represents the uncertainty on a unknown non-random constant $\boldsymbol{\theta}$, and it is **n**ot a density or probability function! It provides

a rational degree of belief, or
an order of preferences

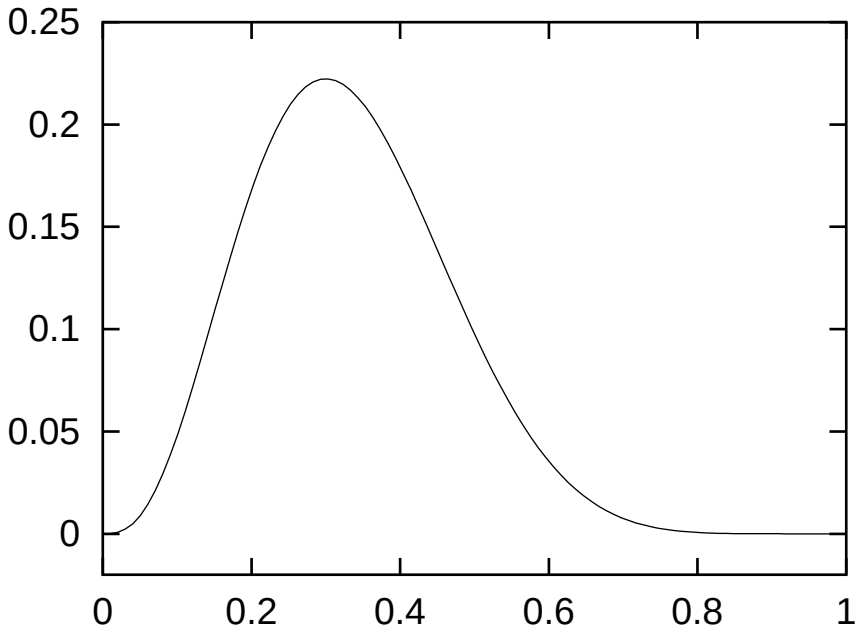
on possible values of the parameter $\boldsymbol{\theta}$. This can be seen more clearly in the simple example on next slide.

In fact, a likelihood function is often defined up to a positive constant — the constant here refers to a quantity independent of $\boldsymbol{\theta}$. But it may depending on $\mathbf{x}$. (Note $\mathbf{x}$ is a given constant.)

Example 1. Suppose that x is the number of successes from a known number n of independent trials with unknown probability of success π. The probability function, and so the likelihood function is

$$L(\pi) = f(x;\pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x}.$$

The likelihood function $L(\pi;x)$ can be graphed as a function of π. It changes shape for different values of x. A likelihood function for $x=3$ when $n=10$ is shown in the Figure below.



Notice that the likelihood function shown above is *not* a density function. It does not have an area of 1 below it.

We use the likelihood function to compare the plausibility of different possible parameter values. For instance, the likelihood is much larger for $\pi = 0.3$ than for $\pi = 0.8$, that is the data $x = 3$ have a greater probability of being observed if $\pi = 0.3$ than if $\pi = 0.8$. This makes $\pi = 0.3$ much more **l**ikely as the true value for π than 0.8.

Note. In the above argument, we do not need to calculate exact probabilities under different values of $\boldsymbol{\theta}$. Only the order of those quantities matters!

Let $X_1, \cdots, X_n$ be i.i.d. with pdf $f(\cdot, \boldsymbol{\theta})$. Write $\mathbf{X} = (X_1, \cdots, X_n)^\tau$. Then the likelihood function is

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{X}) = \prod_{i=1}^{n} f(X_i, \boldsymbol{\theta}),$$

which is a product of n terms. Then the **l**og-likelihood function is

$$l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}; \mathbf{X}) \equiv \log\{L(\boldsymbol{\theta}; \mathbf{X})\} = \sum_{i=1}^{n} \log\{f(X_i, \boldsymbol{\theta})\},$$

which is a sum of n terms.

This explains why log-likelihood functions are often used with independent observations.

Definition 2. *Likelihood Principle*

For two observed values $\mathbf{x}$ and $\mathbf{y}$, if

$$L(\boldsymbol{\theta}; \mathbf{x}) \propto L(\boldsymbol{\theta}; \mathbf{y}),$$

the inferences for $\boldsymbol{\theta}$ based on $\mathbf{x}$ and $\mathbf{y}$ should be the same.

2.2 Sufficiency and Data Compression

If we start with n observations and construct from them a k-dimensional statistic, if $k < n$ we expect that in general some information in the sample about $\boldsymbol{\theta}$ would be lost. The statistics is less complicated than the original sample because it is in a smaller dimension. It reduces the data, and in general will reduce the information about $\boldsymbol{\theta}$. A simple example may help to make clear what is happening.

Example 2 Based on a sample $\{X_1, \cdots, X_n\}$, consider a sequence of statistics of increasing dimensions:

$$\begin{aligned}
T_1(\mathbf{X}) &= \left(\sum_{i=1}^{n} X_i\right) \\
T_2(\mathbf{X}) &= \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right) \\
T_3(\mathbf{X}) &= \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i^3\right) \\
&\cdots \\
T_{n-1}(\mathbf{X}) &= \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i^3, \ldots, \sum_{i=1}^{n} X_i^{n-1}\right) \\
T_n(\mathbf{X}) &= \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i^3, \ldots, \sum_{i=1}^{n} X_i^n\right)
\end{aligned}$$

Here we see that T_1 is very simple and in 1 dimension, whereas T_n is in n dimensions. Actually, from T_n one can reconstruct the original observations $X_1, X_2, \ldots, X_n$, except for their ordering. So T_n contains all

the information in the sample about $\boldsymbol{\theta}$. But except in special cases T_1 will not include all the information about $\boldsymbol{\theta}$ in the sample.

If a statistic contains all the information in a sample about $\boldsymbol{\theta}$ we shall say that it is a **sufficient statistic** for $\boldsymbol{\theta}$.

Example 3 If X_1, X_2 are independent observations from $N(\mu, 1)$ then the statistic $X_1 + X_2$ contains all the information in the sample about μ. This is intuitive, because knowing $X_1 + X_2$ and $X_1 - X_2$ is equivalent to having both observations X_1, X_2, but $X_1 - X_2$ has a $N(0, 2)$ distribution and so contains no information about μ. This conclusion is made more secure by the independence of X_1+X_2 and X_1-X_2.

Now we make the idea of sufficiency more precise with one of the great ideas from the work of R.A.Fisher. We seek to define an idea of 'sufficiency' such that a statistic T is 'sufficient' for $\boldsymbol{\theta}$ if it contains all the information about $\boldsymbol{\theta}$ that was in the whole set of observations.

Definition 3. *Sufficient Statistic*
Suppose $\mathbf{X} \sim f(\cdot, \boldsymbol{\theta})$. $\mathbf{T}(\mathbf{X})$ is said to be a **s**ufficient statistic for $\boldsymbol{\theta}$ if the conditional distribution of the sample $\mathbf{X}$ given $\mathbf{T}(\mathbf{X})$ does not depend on $\boldsymbol{\theta}$.

The likelihood principle implies the sufficiency principle which indicates that we only need to use sufficient statistics in inference.

Definition 4. *Sufficiency Principle*
All sufficient statistics based on $\mathbf{X}$ should lead to the same inferences for $\boldsymbol{\theta}$.

Theorem 1. *Factorisation Criterion*
Let $\mathbf{X} \sim f(\mathbf{x}, \boldsymbol{\theta})$. Then $\mathbf{T} = \mathbf{T}(\mathbf{X})$ is a sufficient statistic for $\boldsymbol{\theta}$ iff

$$f(\mathbf{x}; \boldsymbol{\theta}) = g(\mathbf{T}(\mathbf{x}), \boldsymbol{\theta}) h(\mathbf{x}).$$

We will not prove this result. To see how a proof might go, note that it is always possible, but not necessary, to take $h(\mathbf{x})$ as the conditional density of $\mathbf{X}$ given $\mathbf{T}(\mathbf{X})$ and $g(\mathbf{T}(\mathbf{x}), \boldsymbol{\theta})$ as the density of $\mathbf{T}(\mathbf{X})$.

Example 4. Suppose that we have a random sample $\mathbf{X} = (X_1, \ldots, X_n)^\tau$ from $N(\mu, 1)$. Then the joint density is

$$f_{\mathbf{X}}(\mathbf{X}; \boldsymbol{\theta}) = \frac{1}{(\sqrt{2\pi})^n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^2\right]$$

$$= \exp\left[-\frac{1}{2}n(\bar{X} - \mu)^2\right] \frac{1}{(\sqrt{2\pi})^n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right].$$

So we can take $T(\mathbf{X}) = \bar{X}$ and

$$g(\bar{X}, \boldsymbol{\theta}) = \exp\left[-\frac{1}{2}n(\bar{X} - \mu)^2\right],$$

and

$$h(\mathbf{X}) = \frac{1}{(\sqrt{2\pi})^n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right].$$

Two samples with the same sample mean should lead to exactly the same inferences about μ if the sufficiency principle holds.

Example 5. Here is another example where a sufficient statistic is of the same dimension as the sample size, but is not the whole sample. Let $\mathbf{X} = (X_1, \cdots, X_n)$ be a sample from $U(-\boldsymbol{\theta}, \boldsymbol{\theta})$ and $\Theta = R^+$. Then the absolute values of the observations $\mid X_1 \mid, \mid X_2 \mid, \ldots, \mid X_n \mid$ are from $U(0, \boldsymbol{\theta})$. One loses the information on the signs of the observations, but retains full information on $\boldsymbol{\theta}$. Here the statistic $\mathbf{T}(X) = (\mid X_1 \mid, \mid X_2 \mid, \ldots, \mid X_n \mid)$ is a sufficient statistic for $\boldsymbol{\theta}$.

In fact the conditional distribution of $\mathbf{X}$ given $\mathbf{T}(\mathbf{X}) = (T_1, T_2, \ldots, T_n) = (\mid X_1 \mid, \mid X_2 \mid, \ldots, \mid X_n \mid)$ is the distribution where independently for all i, $X_i = T_i$ with probability 0.5 and $X_i = -T_i$ with probability 0.5. This conditional distribution does not depend on $\boldsymbol{\theta}$, so the definition is satisfied.

Example 6. (Continuing Example) Another sufficient statistic of $\boldsymbol{\theta}$ is $\max_{1 \le i \le n} |X_i|$. To this end, note that

$$f_{X_i}(x_i) = \frac{1}{2}\boldsymbol{\theta} I_{(-\boldsymbol{\theta}, \boldsymbol{\theta})}(x_i) = \begin{cases} \frac{1}{2\boldsymbol{\theta}} & |x_i| < \boldsymbol{\theta}, \\ 0 & \text{otherwise.} \end{cases}$$

Hence

$$f_{\mathbf{X}}(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

$$= \begin{cases} \frac{1}{(2\boldsymbol{\theta})^n} & \max_{1 \le i \le n} |x_i| < \boldsymbol{\theta}, \\ 0 & \text{otherwise} \end{cases}$$

$$= \frac{1}{(2\boldsymbol{\theta})^n} I(\max_{1 \le i \le n} |x_i| < \boldsymbol{\theta}).$$

Now the conclusion follows from the factorisation criterion.

How do you estimate $\boldsymbol{\theta}$?

Example 7. Suppose that the $X_1, \cdots, X_n$ are iid $U(0, \boldsymbol{\theta})$, where $\boldsymbol{\theta} > 0$. Then

$$f(x_i; \boldsymbol{\theta}) = \begin{cases} \frac{1}{\boldsymbol{\theta}} & x_i \in [0, \boldsymbol{\theta}] \\ 0 & \text{otherwise} \end{cases}$$

which can be written

$$f(x_i; \boldsymbol{\theta}) = \frac{1}{\boldsymbol{\theta}} \times I_{\{x_i \le \boldsymbol{\theta}\}} I_{\{x_i \ge 0\}}.$$

The joint density is

$$f(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{\boldsymbol{\theta}^n} \prod_{i=1}^{n} I_{\{x_i \le \boldsymbol{\theta}\}} \prod_{i=1}^{n} I_{\{x_i \ge 0\}}$$

$$= \frac{1}{\boldsymbol{\theta}^n} I_{\{\max_i x_i \le \boldsymbol{\theta}\}} I_{\{\min_i x_i \ge 0\}}$$

$$= g\left(T(\mathbf{x}); \boldsymbol{\theta}\right) \times h(\mathbf{x})$$

where $T(\mathbf{x}) = \max_i x_i$. In this case the (at first sight) rather surprising result is that $\max_i X_i$ is a sufficient statistic for $\boldsymbol{\theta}$.

2.3 Maximum likelihood estimator (MLE)

The MLE is by far the most popular method for deriving estimators.

Definition 5 — **M**LE
A *Maximum Likelihood Estimator* (MLE), $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{X}) \in \Theta$, of parameter $\boldsymbol{\theta}$ is an estimator satisfying

$$L(\widehat{\boldsymbol{\theta}}; \mathbf{X}) \geq L(\boldsymbol{\theta}; \mathbf{X}) \text{ for all } \boldsymbol{\theta} \in \Theta, \text{ or equivalently } l(\widehat{\boldsymbol{\theta}}; \mathbf{X}) \geq l(\boldsymbol{\theta}; \mathbf{X}) \text{ for all } \boldsymbol{\theta} \in \Theta.$$

Obviously, a maximum likelihood estimator is the most plausible value for $\boldsymbol{\theta}$ as judged by the likelihood function. In many cases where Θ is continuous and the maximum does not occur at a boundary of Θ, $\widehat{\boldsymbol{\theta}}$ **i**s often the solution of the equation

$$s(\boldsymbol{\theta}; \mathbf{X}) = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}; \mathbf{X}) = 0.$$

We call $s(\boldsymbol{\theta}) \equiv s(\boldsymbol{\theta}; \mathbf{X})$ **a** score function.

Example 8. Suppose that $Y_1, Y_2, \ldots, Y_n$ is a random sample from $N(\mu, \sigma^2)$ where neither μ or σ^2 is known. Then we can find the maximum likelihood estimator from the log-likelihood

$$\begin{aligned} l(\mu, \sigma^2) &= -n \log \sqrt{2\pi} - n/2 \log \sigma^2 - \sum_1^n (Y_i - \mu)^2/(2\sigma^2) \\ &= -n \log \sqrt{2\pi} - n/2 \log \sigma^2 - \sum_1^n (Y_i - \bar{Y})^2/(2\sigma^2) \\ &- n(\bar{Y} - \mu)^2/(2\sigma^2). \end{aligned}$$

This is maximised by choosing $\mu = \bar{Y}$, so $\widehat{\mu} = \bar{Y}$ is the MLE for μ. It is easy to see

$$E(\widehat{\mu}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \mu.$$

Such a estimator is called **u**nbiased.

Note. The above MLE for μ is equivalent to minimise

$$\sum_{i=1}^n (Y_i - \mu)^2.$$

Therefore $\widehat{\mu}$ is also the **l**east squares estimator (LSE). Note that LSE is derived from a simple empirical/geometric rule, which makes **no use** of the underlying distribution.

The **p**rofile log-likelihood remaining is

$$l(\widehat{\mu}, \sigma^2) = -n \log \sqrt{2\pi} + (n/2)(\log \sigma^{-2} - \widehat{\sigma}^2 \sigma^{-2}),$$

where $\widehat{\sigma}^2 = \sum_1^n (Y_i - \bar{Y})^2/n$. By the lemma below, the MLE for σ^2 is $\widehat{\sigma}^2$. Note that the MLE of σ^2 is *biased* since

$$E(\widehat{\sigma}^2) = (1 - 1/n)\sigma^2 \neq \sigma^2.$$

Lemma. Define $L(x) = \log(x^{-1}) - b/x$, where $n \geq 1$ and $b > 0$ are constants. Then $L(b) \geq L(x)$ for all $x > 0$.

Example 9. Let $X_1, \cdots, X_n$ be i.i.d. Bernoulli(π). Then

$$L(\pi) = \prod_{i=1}^n \pi^{X_i} (1 - \pi)^{1 - X_i} = \pi^{n\bar{X}} (1 - \pi)^{n(1 - \bar{X})}.$$

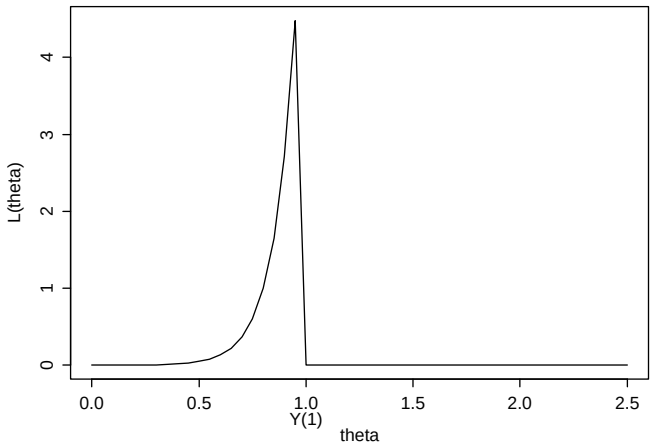
$$l(\pi) = n\bar{X} \log \pi + n(1 - \bar{X}) \log(1 - \pi).$$

Let $s(\pi) = \frac{\partial}{\partial \pi} l(\pi) = 0$, leading to $\widehat{\pi} = \bar{X}$.

Example 10. Suppose that $Y_1, Y_2, \ldots, Y_n$ is a random sample from an exponential distribution with density function $e^{-(y-\theta)}$ for $y \geq \theta$. This is the usual exponential distribution shifted to start at θ. The Likelihood is

$$L(\theta; \mathbf{Y}) = e^{-n(\bar{Y}-\theta)} I_{\{(\theta,\infty)\}}(Y_{(1)}),$$

where $Y_{(1)}$ is the smallest observation. This likelihood is zero for $\theta > Y_{(1)}$ and increases in θ for $\theta \leq Y_{(1)}$. So the MLE $\widehat{\theta} = Y_{(1)}$, which is a boundary maximum.



Invariance property of MLEs

Suppose $\mathbf{X} \sim f(\mathbf{x}, \boldsymbol{\theta})$, and $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$ is a 1-1 transform. Let $\widehat{\boldsymbol{\theta}}$ be the MLE for $\boldsymbol{\theta}$, i.e.

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} f(\mathbf{X}, \boldsymbol{\theta}).$$

It is obvious to see that the MLE for $\boldsymbol{\mu}$ is

$$\widehat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\widehat{\boldsymbol{\theta}}).$$

2.4 Numerical computation of MLEs

In modern statistical applications, it is typically difficult to find explicit analytic forms for the maximum likelihood estimators. These estimators are found more often by iterative procedures built into computer software. An iterative scheme starts with some guess at the MLE and then steadily improves it with each iteration. The estimator is considered found when it has become numerically stable. Sometimes the iterative procedures become trapped at a local maximum which is not a global maximum. There may be a very large number of parameters in a model, which makes such local entrapment much more common.

Newton-Raphson Scheme

Suppose that the log-likelihood function function $l(\boldsymbol{\theta})$ is sufficiently smooth. Then

$$s(\widehat{\boldsymbol{\theta}}) = 0,$$

where $\widehat{\boldsymbol{\theta}}$ is the MLE and $s(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta} l(\boldsymbol{\theta})$ is the score function. Let

$$\dot{s}(\boldsymbol{\theta}) = \ddot{l}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \theta \partial \theta^\tau} l(\boldsymbol{\theta}).$$

Suppose $\widehat{\boldsymbol{\theta}}$ is close to the true value $\boldsymbol{\theta}_0$. By a simple Taylor expansion,

$$s(\boldsymbol{\theta}_0) = \dot{s}(\boldsymbol{\theta}_0)(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}) + o_p(||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||).$$

This leads to the approximation

$$\widehat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}_0 - \{\dot{s}(\boldsymbol{\theta}_0)\}^{-1} s(\boldsymbol{\theta}_0).$$

Since θ_0 is unknown, we use iterative estimators

$$\widehat{\boldsymbol{\theta}}_{k+1} = \widehat{\boldsymbol{\theta}}_k - \{\dot{s}(\widehat{\boldsymbol{\theta}}_k)\}^{-1} s(\widehat{\boldsymbol{\theta}}_k) \tag{1}$$

for $k = 1, 2, \cdots$, where $\widehat{\boldsymbol{\theta}}_0$ is a prescribed initial value. We define $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_j$ if $\widehat{\boldsymbol{\theta}}_j$ and $\widehat{\boldsymbol{\theta}}_{j-1}$ differ by a small amount.

Fisher Score method: replace $\dot{s}(\widehat{\boldsymbol{\theta}}_k)$ in (1) by $E_{\boldsymbol{\theta}}\{\dot{s}(\boldsymbol{\theta})\}$ under $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_k$.

Like most iterative algorithms, the choice of appropriate **i**nitial values is important to ensure the convergence to right limits. In practice multiple initial values are often used.

The differences between the Newton-Raphson and score methods are subtle. We make observations below

- The convergence of the Newton-Raphson algorithm is often faster when both algorithms converge
- The radius of convergence for the score method is often larger, making the choice of initial values less important for the score method.

2.5 EM algorithms

Goal: to find the MLE $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{Y})$ for $\boldsymbol{\theta}$ from the likelihood based on data $\mathbf{Y}$:

$$L(\boldsymbol{\theta}; \mathbf{Y}) = f_{\mathbf{Y}}(\mathbf{Y}, \boldsymbol{\theta}),$$

while the '*complete*' data $\mathbf{X}^\tau = (\mathbf{Y}^\tau, \mathbf{Z}^\tau)$ contain a '*missing*' component $\mathbf{Z}$. The likelihood based on the complete data is

$$L(\boldsymbol{\theta}; \mathbf{X}) = f_{\mathbf{X}}(\mathbf{X}, \boldsymbol{\theta}).$$

EM (*E*xpectation and *M*aximisation) **a**lgorithms

- *E-step*: compute the conditional expectation

$$Q(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\mathbf{Y}, \boldsymbol{\theta}_0) \equiv E\{\log L(\boldsymbol{\theta}; \mathbf{X})|\mathbf{Y}, \boldsymbol{\theta}_0\}$$

- *M-step*: maximise $Q(\boldsymbol{\theta})$ to give an updated value $\boldsymbol{\theta}_1$

then go to the E-step using $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_1$, and keep iterating until convergence. The limit of $\boldsymbol{\theta}_0$ is taken as $\widehat{\boldsymbol{\theta}}(\mathbf{Y})$.

Example 11. The genetic example from (Rao 1973, p.396) assumes that the phenotype data

$$\begin{aligned} \mathbf{Y} &= (Y_1, Y_2, Y_3, Y_4)^\tau \sim \\ &\quad \text{Multinomial}\big(4; \frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\big), \end{aligned}$$

where $\theta \in (0,1)$. Then log-likelihood is

$$l(\theta, \mathbf{Y}) = Y_1 \log(2+\theta) + (Y_2 + Y_3)\log(1-\theta) + Y_4 \log\theta + C,$$

which **d**oes not yield a closed form $\widehat{\theta}$.

Now we treat $\mathbf{Y}$ as incomplete data from $\mathbf{X} = (X_1, \ldots, X_5)^\tau$ with multinomial probabilities

$$(\frac{1}{2},\ \frac{\theta}{4},\ \frac{1-\theta}{4},\ \frac{1-\theta}{4},\ \frac{\theta}{4}).$$

Then

$$Y_1 = X_1 + X_2, \quad Y_i = X_{i+1} \quad \text{for } i = 2,3,4.$$

The log-likelihood of based on $\mathbf{X}$ is

$$l(\theta, \mathbf{X}) = (X_2 + X_5)\log\theta + (X_3 + X_4)\log(1-\theta) + C,$$

which **r**eadily yields

$$\widehat{\theta}(\mathbf{X}) = \frac{X_2 + X_5}{X_2 + X_3 + X_4 + X_5}.$$

Now the E-step is to find

$$\begin{aligned} Q(\theta) &= E\{l(\theta, \mathbf{X})|\mathbf{Y}, \theta_0\} \\ &= \log\theta E(X_2 + X_5|\mathbf{Y}, \theta_0) + \log(1-\theta)E(X_3 + X_4|\mathbf{Y}, \theta_0) \\ &= (\widehat{X}_2 + Y_4)\log\theta + (Y_2 + Y_3)\log(1-\theta), \end{aligned}$$

where $\widehat{X}_2 = E(X_2|\mathbf{Y}, \theta_0)$. Since the conditional distribution of X_2 given $Y_1 (= X_1 + X_2)$ is a binomial distribution with $n = Y_1$ and

$$p = \frac{\theta_0/4}{1/2 + \theta_0/4} = \frac{\theta_0}{2+\theta_0}.$$

Hence

$$\widehat{X}_2 = np = \frac{Y_1\theta_0}{2+\theta_0}. \tag{2}$$

The M-step leads to

$$\theta_1 = \frac{\widehat{X}_2 + Y_4}{\widehat{X}_2 + Y_4 + Y_2 + Y_3}. \tag{3}$$

For $Y = (125, 18, 20, 34)$ and the **initial value**

$$\theta_0 = 4 \times 34/(125 + 18 + 20 + 34)$$

which is a relative frequency estimate, the first 5 iterations between (2) and (3) are 0.690, 0.635, 0.628, 0.627 and 0.627, giving the MLE $\widehat{\theta} = 0.627$.

What can be said about *convergence properties* of the EM algorithm?

Let $\boldsymbol{\theta}_0$ be an arbitrary initial value and $\boldsymbol{\theta}_1$ be the updated value obtained from applying the iteration once. Then it can be shown that

$$L(\boldsymbol{\theta}_1; \mathbf{Y}) \geq L(\boldsymbol{\theta}_0; \mathbf{Y}).$$

Unfortunately, it does not imply that the iterations will always lead to the MLE eventually.

It is important to choose appropriate initial values to ensure the algorithm converges to the MLE. (This is also true for both Gaussian-Raphson and score methods!) In practice, it is a good idea to use **a** variety of initial values.

Further discussion on the convergence of EM algorithms, see Wu (1983) *Annals of Statistics, Vol.11, pp.95-103* and §12.4 of Pawitan (2001).

General comments on EM algorithms:

- The EM algorithm is a general procedure for computing MLEs. It is not really a numerical algorithm. The calculation of M-Step typically involves other numerical algorithms such as Newton-Raphson and the score methods.
- It can be applied when some data are *genuinely* missing. It can also be applied when the missing information is merely a concept based on which we transform a difficult optimisation problem into a sequence of easier problems; see Example 11 above. This is particularly relevant when $\widehat{\boldsymbol{\theta}}(\mathbf{Y})$ is difficult to calculate while $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ is easier to obtain.
- The convergence of EM algorithm may be very slow, depending on the amount of missing information. It is often slower than the Newton-Raphson algorithm.

Example 12. Mixture distributions
Let $Y_1, \cdots, Y_n$ be i.i.d. with a mixture pdf

$$f(y) = \sum_{j=1}^{k} \alpha_j f_j(y, \boldsymbol{\lambda}_j),$$

where f_j are pdfs, $\alpha_j \geq 0$ and $\sum_j \alpha_j = 1$. The parameter $\boldsymbol{\theta}$ contains all α_j and $\boldsymbol{\lambda}_j$. This model becomes important because

1. it represents heterogeneous data well since each f_j represents one heterogeneous component, and
2. it provides very good approximations to a large class of distributions.

The likelihood based on $\mathbf{Y} = (Y_1, \cdots, Y_n)^\tau$ is

$$L(\boldsymbol{\theta}; \mathbf{Y}) = \prod_{i=1}^{n} \Big\{ \sum_{j=1}^{k} \alpha_j f_j(Y_j, \boldsymbol{\lambda}_j) \Big\}.$$

Maximising this likelihood is difficult, due to the presence of the summations, which reflect the fact that we are typically lacking in knowledge of which component any particular sample value comes from. This is the missing information!

Let $\mathbf{X}_i^\tau = (Y_i, \mathbf{Z}_i^\tau)$, where $\mathbf{Z}_i = (Z_{i1}, \cdots, Z_{ik})^\tau$ is a $k \times 1$ vector with a 1 in the position corresponding to the component of the mixture that Y_i comes from, and 0 elsewhere.

Let $\mathbf{e}_j$ be the $k \times 1$ vector with the j-th component 1 and all the other components 0. The joint probability-density function for $(\mathbf{Z}_1, Y_1)$ is

$$\begin{aligned} & P\{Z_1 = \mathbf{e}_\ell,\ Y_1 \in [y_1, y_1 + dy_1)\}/dy_1 \\ = \ & P(Z_1 = \mathbf{e}_\ell)\, P\{Y_1 \in [y_1, y_1 + dy_1) | Z_1 = \mathbf{e}_\ell\}/dy \\ = \ & \alpha_\ell f_\ell(y_1, \boldsymbol{\lambda}_\ell) \ = \ \prod_{j=1}^{k} \alpha_j^{e_{\ell j}} f_j(y_1, \boldsymbol{\lambda}_j)^{e_{\ell j}}, \end{aligned}$$

where $\mathbf{e}_\ell = (e_{\ell 1}, \cdots, e_{\ell k})^\tau$.

Let $\mathbf{X}^\tau = (\mathbf{X}_1^\tau, \cdots, \mathbf{X}_n^\tau)$.

$$L(\boldsymbol{\theta}; \mathbf{X}) = \prod_{i=1}^{n} \prod_{j=1}^{k} \alpha_j^{Z_{ij}} f_j(Y_i, \boldsymbol{\lambda}_j)^{Z_{ij}},$$

$$l(\boldsymbol{\theta};\mathbf{X}) = \sum_{i=1}^{n} \mathbf{Z}_i^\tau \left\{ \begin{pmatrix} \log\alpha_1 \\ \vdots \\ \log\alpha_k \end{pmatrix} + \begin{pmatrix} \log f_1(Y_i,\boldsymbol{\lambda}_1) \\ \vdots \\ \log f_k(Y_i,\boldsymbol{\lambda}_k) \end{pmatrix} \right\}.$$

Note that

$$E(\mathbf{Z}_i|\mathbf{Y},\boldsymbol{\theta}_0) = \Big(\frac{\alpha_1^0 f_1(Y_i|\boldsymbol{\lambda}_1^0)}{\sum_{\ell=1}^k \alpha_\ell^0 f_\ell(Y_i|\boldsymbol{\lambda}_\ell^0)}, \cdots, \frac{\alpha_k^0 f_k(Y_i|\boldsymbol{\lambda}_k^0)}{\sum_{\ell=1}^k \alpha_\ell^0 f_\ell(Y_i|\boldsymbol{\lambda}_\ell^0)}\Big)^\tau,$$

which are constants as far as the M-step is concerned.

Now the E-step implies that

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{n} E(\mathbf{Z}_i^\tau|\mathbf{Y},\boldsymbol{\theta}_0) \left\{ \begin{pmatrix} \log\alpha_1 \\ \vdots \\ \log\alpha_k \end{pmatrix} + \begin{pmatrix} \log f_1(Y_i,\boldsymbol{\lambda}_1) \\ \vdots \\ \log f_k(Y_i,\boldsymbol{\lambda}_k) \end{pmatrix} \right\},$$

and the M-step requires to maximise this function, which is much easier than minimizing $l(\theta,\mathbf{Y})$ directly for, for example, normal pdf f_j's.

2.6 Evaluating estimation

To measure the accuracy of an MLE or, more general, any estimation procedure, we need to define some measures for the goodness (or badness) of an estimator.

Let $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{X})$ be an estimator of $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_o$ be the (unknown) *true value* of $\boldsymbol{\theta}$. Note that

(i) exact estimation error $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o$ is unknown, and
(ii) $\widehat{\boldsymbol{\theta}}$ is a random variable

we have to gauge the error

(i) in terms of a probability average, and
(ii) for all possible values of $\boldsymbol{\theta}_o \in \Theta$.

Let $P_{\boldsymbol{\theta}}$, $E_{\boldsymbol{\theta}}$ and $\mathrm{Var}_{\boldsymbol{\theta}}$ denote the probability distribution, expectation and variance under $\boldsymbol{\theta}_o = \boldsymbol{\theta}$.

Bias: $\mathrm{Bias}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta}$

Variance: $\mathrm{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})$

Standard error: $\{\mathrm{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})\}^{1/2}$

Mean square error (MSE): $E_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2$

Mean absolute error (MAE): $E_{\boldsymbol{\theta}}|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}|$

Note that

- standard error is a meaningful measure of accuracy for unbiased estimators only, and
- MSE (it its squared-root) should be used in general as

$$\mathrm{MSE}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) = \{\mathrm{Bias}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}})\}^2 + \mathrm{Var}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}).$$

Ideally we would seek for the estimator which minimises MSE or MAE **f**or all $\boldsymbol{\theta} \in \Theta$ over all possible candidate estimators. Unfortunately such a global optimum rarely exists. However if we confine to some subclass of estimators, the MLE is often optimal or asymptotically optimal.

The MSE is most frequently used largely due to is technical tractability while the MAE leads to estimators which is more robust against outliers in observations.

Fisher Information

Suppose $\mathbf{X} \sim f(\mathbf{x}, \boldsymbol{\theta})$. The score function is

$$s(\boldsymbol{\theta}) = \dot{l}(\boldsymbol{\theta}; \mathbf{X}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log\{f(\mathbf{X}, \boldsymbol{\theta})\}.$$

We assume certain regularity conditions so that we can take derivatives under the integral sign.

Mean of $s(\boldsymbol{\theta})$:

$$\begin{aligned} E_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta})\} &= \int s(\boldsymbol{\theta}) f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} \\ &= \int \frac{\partial}{\partial \boldsymbol{\theta}} \log\{f(\mathbf{x}, \boldsymbol{\theta})\} f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} \\ &= \int \frac{\partial}{\partial \boldsymbol{\theta}} f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \int f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} \; = \; 0. \end{aligned}$$

Variance of $s(\boldsymbol{\theta})$ — **F**isher information matrix:

$$\begin{aligned} \mathcal{I}(\boldsymbol{\theta}) &= \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta})\} \\ &= E_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta}) s(\boldsymbol{\theta})^{\tau}\} \\ &= -E_{\boldsymbol{\theta}}\big[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\tau}} \log\{f(\mathbf{X}, \boldsymbol{\theta})\}\big], \end{aligned}$$

because

$$\begin{aligned} & E_{\boldsymbol{\theta}}\big\{\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\tau}} l(\boldsymbol{\theta})\big\} \\ = & \int \frac{\ddot{L}(\boldsymbol{\theta}) L(\boldsymbol{\theta}) - \dot{L}(\boldsymbol{\theta}) \dot{L}(\boldsymbol{\theta})^{\tau}}{L(\theta)} d\mathbf{x} \\ = & \int \ddot{L}(\boldsymbol{\theta}) d\mathbf{x} - \int \frac{\dot{L}(\boldsymbol{\theta}) \dot{L}(\boldsymbol{\theta})^{\tau}}{L(\theta)} d\mathbf{x} \\ = & -\int \frac{\dot{L}(\boldsymbol{\theta}) \dot{L}(\boldsymbol{\theta})^{\tau}}{L(\theta)} d\mathbf{x} \\ = & -\int s(\boldsymbol{\theta}) s(\boldsymbol{\theta})^{\tau} f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} \\ = & -E_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta}) s(\boldsymbol{\theta})^{\tau}\}. \end{aligned}$$

Fisher information $\mathcal{I}(\boldsymbol{\theta})$ measures **t**he information on $\boldsymbol{\theta}$ contained in data $\mathbf{X}$. Further if $\mathbf{X} = (X_1, \cdots, X_n)^{\tau}$, and $X_1, \cdots, X_n$ are independent,

$$\mathcal{I}(\boldsymbol{\theta}) = \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}) = \sum_{j=1}^{n} \mathcal{I}_{X_j}(\boldsymbol{\theta}).$$

For $\boldsymbol{\theta} = \theta$ is a scalar, the Fisher information is

$$\mathcal{I}(\theta) = E_\theta\{s(\theta)^2\} = -E_\theta\{\ddot{l}(\theta)\}.$$

Theorem 2. *(Cramér-Rao inequality)*
Let $\mathbf{X} \sim f(\cdot, \theta)$ which satisfying some regularity conditions. Let $T = T(\mathbf{X})$ be a statistic with $g(\theta) = E_\theta(T)$. Then for any $\theta \in \Theta$,

$$\mathrm{Var}_\theta(T) \geq \{\dot{g}(\theta)\}^2/\mathcal{I}(\theta).$$

The Cramér-Rao inequality specifies **a** lower bound for any *unbiased estimator* for the parameter $g(\theta)$. When the equality holds, T is the **m**inimum variance unbiased estimator (MVUE) of $g(\theta)$.

An MVUE is a member in the class of all unbiased estimators with the minimum variance. *Not all MVUEs can attain the Cramér-Rao lower bounds.*

Example 13. Let $X_1, \cdots, X_n$ be a sample from $N(\mu, \sigma^2)$. We consider estimators for μ, treating σ^2 as known. The score function (for one observation) is

$$\begin{aligned} s(\mu; X_1) &= \frac{\partial}{\partial \mu} \log[e^{-\frac{1}{2\sigma^2}(X_1-\mu)^2}/\sqrt{2\pi\sigma^2}] \\ = \quad & \frac{\partial}{\partial \mu}[-\frac{1}{2\sigma^2}(X_1-\mu)^2] = (X_1-\mu)/\sigma^2. \end{aligned}$$

Note $\ddot{l}(\mu) = \dot{s}(\mu) = -\sigma^{-2}$. Hence the Fisher information based on a single observation is $\mathcal{I}_{X_1}(\mu) = \sigma^{-2}$. Therefore

$$\mathcal{I}(\mu) = \mathcal{I}_{X_1,\cdots,X_n}(\mu) = n/\sigma^2.$$

For any unbiased estimator $\widehat{\mu}$ for μ, it holds that

$$\mathrm{Var}_\mu(\widehat{\mu}) \geq \sigma^2/n,$$

which is the variance of $\bar{X}$. Hence $\bar{X}$ is the MVUE for μ.

Asymptotic properties of MLEs

Let $X_1, \cdots, X_n$ be i.i.d. with pdf $f(\cdot, \boldsymbol{\theta})$. Write

$$l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}; \mathbf{X}) = \sum_{j=1}^{n} \log f(X_j, \boldsymbol{\theta}).$$

Let $\widehat{\boldsymbol{\theta}}$ be the MLE which maximises $l(\boldsymbol{\theta})$. Suppose f fulfils certain regularity conditions.

(a) **Consistency**.
The MLE is consistent in the sense that as $n \to \infty$,

$$P_{\boldsymbol{\theta}}\{||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}|| > \varepsilon\} \to 0$$

for any $\varepsilon > 0$.

Consistency requires that an estimator converges to the parameter to be estimated. It is a very mild and modest condition that any reasonable estimator should fulfil. The consistency condition is often used to *rule out bad estimators*.

(b) **Asymptotic normality**
As $n \to \infty$,

$$n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{D} N\big(0,\ \{\mathcal{I}_{X_1}(\boldsymbol{\theta})\}^{-1}\big).$$

For large n, it holds **a**pproximately that

$$\widehat{\boldsymbol{\theta}} \sim N\big(\boldsymbol{\theta},\ \{\mathcal{I}_{X_1}(\boldsymbol{\theta})\}^{-1}/n\big).$$

Therefore *asymptotically* the MLE is unbiased and attains the Cramér-Rao lower bound. Any estimator fulfilling this condition is called **e**fficient.

An approximate standard error of the j-th component of $\widehat{\boldsymbol{\theta}}$ is the square-root of the (j,j)-th element of $\{\mathcal{I}_{X_1}(\boldsymbol{\theta})\}^{-1}$ divided by $n^{1/2}$.

2.7 Bootstrapping MSEs

An MSE provides a measure for the accuracy of the estimator. The standard errors of an MLE derived from the asymptotic normality typically depend on unknown parameters, which are replaced by their estimators in practice. Alternatively we may estimate the MSE by a simulation method called Bootstrap.

Let $X_1, \cdots, X_n$ be i.i.d. with pdf $f(\cdot, \theta)$. Let

$$\widehat{\theta} = T(X_1, \cdots, X_n)$$

be an estimator. The *goal* here is to estimate

$$\nu \equiv \{\mathrm{MSE}_{\theta_o}(\widehat{\theta})\}^{1/2},$$

where θ_o is the true value.

If we *knew* $f(\cdot, \theta_o)$ completely, ν is known in principle, and may be estimated easily via a repeated sampling as follows. We draw B independent samples of size n from $f(\cdot, \theta_o)$. For each sample, we calculate $\widehat{\theta}$, obtaining $\widehat{\theta}_1, \cdots, \widehat{\theta}_B$. Then the sample root-MSE

$$\Big\{\frac{1}{B}\sum_{b=1}^{B}(\widehat{\theta}_b - \theta_o)^2\Big\}^{1/2}$$

is a reasonable estimator for ν. By the LLN, this estimator converges to ν as $B \to \infty$.

The **b**asic idea of bootstrap is to adopt the above sampling procedure in the so-called *bootstrap world*: now the population is $f(\cdot, \widehat{\theta})$ which is **known**. We draw a sample denoted as $(X_1^*, \cdots, X_n^*)$ from this distribution. Define the *bootstrap version* of the estimator

$$\widehat{\theta}^* = T(X_1^*, \cdots, X_n^*).$$

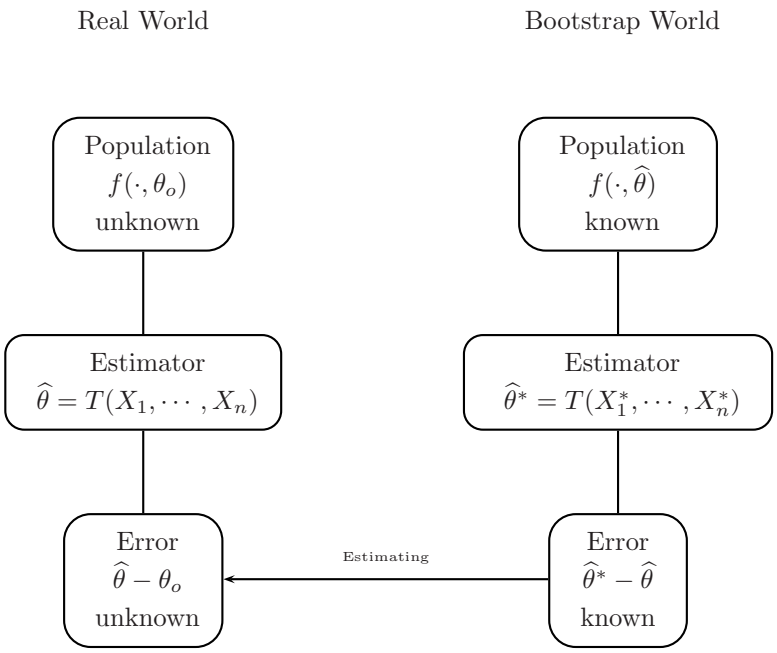
Then the quantity

$$\nu^* = \{\mathrm{MSE}_{\widehat{\theta}}(\widehat{\theta}^*)\}^{1/2}$$

is known in principle since the distribution $f(\cdot, \widehat{\theta})$ is completely known. Define ν^* as a **b**ootstrap estimator for ν.

In practice, we draw B sets samples from $f(\cdot, \widehat{\theta})$, forming B bootstrap versions of estimator $\widehat{\theta}_1^*, \cdots, \widehat{\theta}_B^*$. The ν^* is calculated as

$$\nu^* = \Big\{\frac{1}{B}\sum_{j=1}^{B}(\widehat{\theta}_j^* - \widehat{\theta})^2\Big\}^{1/2}.$$



Bootstrap is a powerful tool for statistical inference. It has different forms for different applications. The one introduced here is in the form of *parametric bootstrap*. The special monographs on this topic include:

- Hall, P. (1992). The bootstrap and Edgeworth expansion. Springer, New York.
- Efron, B. and Tibshirani, R.J. (1993). An introduction to the bootstrap. Chapman and Hall, New York.
- Shao, J. and Tu, D. (1995). The jackknife and bootstrap. Springer, New York.
- Davison, A.C. and Hinkley, D.V. (1997). Bootstrap methods and their application. Cambridge University Press, Cambridge.