Chapter 9. Point Estimation

Let $\{X_1, \cdots, X_n\}$ be a random sample from a population $F(\cdot, \boldsymbol{\theta})$, where the form of F is known but the parameter $\boldsymbol{\theta}$ is unknown, and it has p components. Often we may specify $\boldsymbol{\theta} \in \Theta$, where Θ is called the parameter space.

For $N(\mu, \sigma^2)$, $p = 2$ and $\Theta = (-\infty, \infty) \times (0, \infty)$. For $Poisson(\lambda)$, $p = 1$ and $\Theta = (0, \infty)$.

Goal: to find a (point) estimator for $\boldsymbol{\theta}$.

9.1 Method of Moments Estimation

Let $\mu_k \equiv \mu_k(\boldsymbol{\theta}) = E(X_1^k)$ denote the k-th moment of the population, $k = 1, 2, \cdots$. Then μ_k depends on unknown parameter $\boldsymbol{\theta}$, as everything else on

the distribution $F(\cdot, \boldsymbol{\theta})$ are known. Denote the k-th sample moment by

$$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k = \frac{1}{n}(X_1^k + \cdots + X_n^k).$$

The **MM estimator** $\widehat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ is the solution of the p equations

$$\mu_1(\widehat{\boldsymbol{\theta}}) = M_1,\ \mu_2(\widehat{\boldsymbol{\theta}}) = M_2,\ \cdots,\ \mu_p(\widehat{\boldsymbol{\theta}}) = M_p.$$

Example 1. Let $\{X_1, \cdots, X_n\}$ be a sample from a population with mean μ and variance $\sigma^2 < \infty$. Find the MM estimator for (μ, σ^2).

There are two unknown parameters. Let

$$\mu = \mu_1 = M_1, \qquad \mu_2 = M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

This gives us $\widehat{\mu} = M_1 = \bar{X}$. Since $\sigma^2 = \mu_2 - \mu_1^2$,

$$\widehat{\sigma}^2 = M_2 - M_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Note. $E(\widehat{\sigma}^2) = E(X_1^2) - E(\bar{X}^2) = \sigma^2 + \mu^2 - (\sigma^2/n + \mu^2) = \frac{n-1}{n}\sigma^2$. We call $E(\widehat{\sigma}^2) - \sigma^2 = -\sigma^2/n$ the estimation bias. The **sample variance**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is a more frequently used estimator for σ^2, and it has zero-bias.

Theorem 1. Under some mild regularity conditions, the MME $\widehat{\boldsymbol{\theta}}$ is a consistent estimator in the sense that as $n \to \infty$, $\widehat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}$, i.e.

$$P\{||\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}|| > \epsilon\} \to 0 \qquad \text{for any } \epsilon > 0.$$

Further it is asymptotically normal, i.e. $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges in distribution to a p-dimensional normal distribution.

9.2. Maximum likelihood estimation

9.2.1 Likelihood

Likelihood is one of the most fundamental concepts 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 **not a density or probability function**! It provides

a rational degree of belief, or
an order of preferences

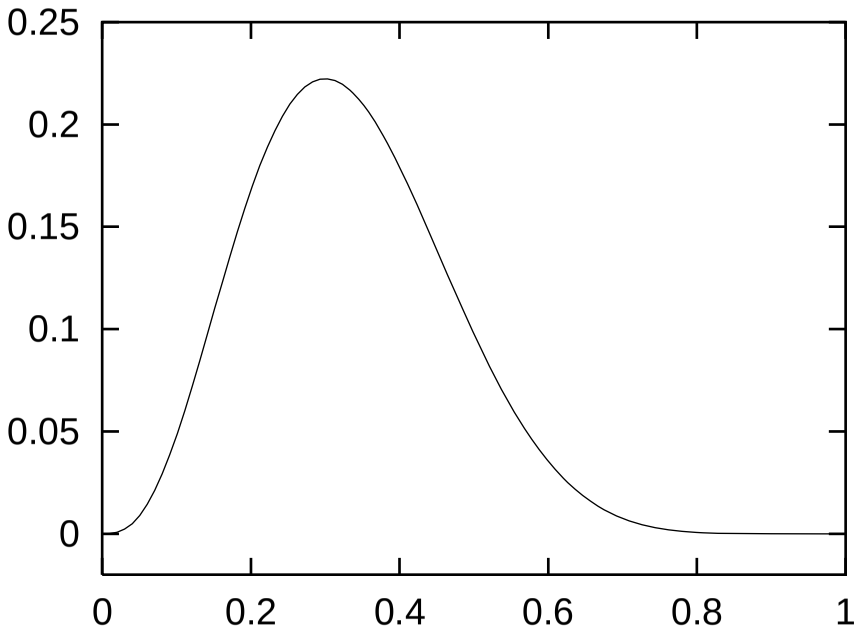
on possible values of the parameter $\boldsymbol{\theta}$. This can be seen more clearly in the simple example on next slide.

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 2. 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 a $x = 3$ when $n = 10$ is shown in the Figure below.


0.25
0.2
0.15
0.1
0.05
0
0
0.2
0.4
0.6
0.8
1

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 **likely** 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 $\mathbf{X} = (X_1, \cdots, X_n)'$. 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 **log-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.

9.2.3 Maximum likelihood estimator (MLE)

The MLE is by far the most popular estimator.

Definition 2 — MLE
A *Maximum Likelihood Estimator* (MLE), $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{X}) \in \Theta$, of parameter $\boldsymbol{\theta}$ is an estimator satisfying

> $L(\widehat{\boldsymbol{\theta}}; \mathbf{X}) \geq L(\boldsymbol{\theta}; \mathbf{X})$ for all $\boldsymbol{\theta} \in \Theta$, or equivalently $l(\widehat{\boldsymbol{\theta}}; \mathbf{X}) \geq l(\boldsymbol{\theta}; \mathbf{X})$ 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, $\widehat{\boldsymbol{\theta}}$ **is 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 3. 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

$$l(\mu, \sigma^2) = -n \log \sqrt{2\pi} - n/2 \log \sigma^2 - \sum_1^n (Y_i - \mu)^2/(2\sigma^2)$$

$$= \quad -n \log \sqrt{2\pi} - n/2 \log \sigma^2 - \sum_1^n (Y_i - \bar{Y})^2/(2\sigma^2) - n(\bar{Y} - \mu)^2/(2\sigma^2).$$

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 **unbiased**.

The **profile 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 $b > 0$ are constants. Then $L(b) \geq L(x)$ for all $x > 0$.

Example 4. Let $X_1, \cdots, X_n$ be i.i.d. Bernoulli(π). Then

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

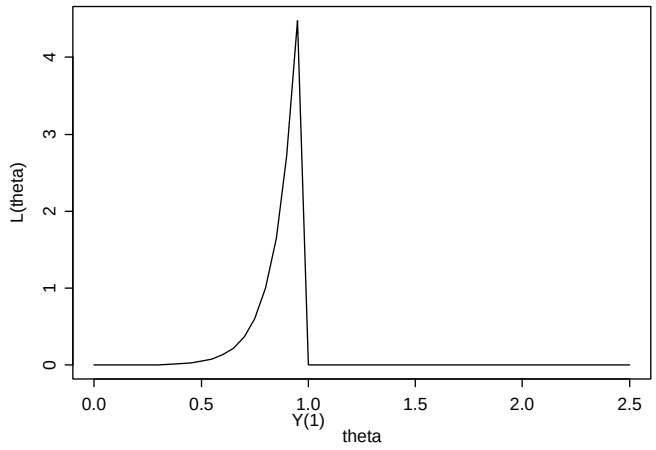
$$l(\pi) = n\bar{X} \log \pi + n(1 - \bar{X}) \log(1 - \pi).$$

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

Example 5. 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 $\boldsymbol{\psi} = g(\boldsymbol{\theta})$. 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 $\boldsymbol{\psi}$ is $\widehat{\boldsymbol{\psi}} = g(\widehat{\boldsymbol{\theta}})$.

If $\boldsymbol{\psi} = g(\boldsymbol{\theta})$ is a 1-1 transform and $\widehat{\boldsymbol{\psi}}$ is the MLE for $\boldsymbol{\psi}$, $\widehat{\boldsymbol{\theta}} \equiv g^{-1}(\widehat{\boldsymbol{\psi}})$ is the MLE for $\boldsymbol{\theta}$.

9.2.4 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

$$\dot{s}(\boldsymbol{\theta}) = \ddot{l}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \theta \partial \theta'} l(\boldsymbol{\theta}) = \left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right).$$

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

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

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

Example 6. Let $X_1, \cdots, X_n$ be a sample from Cauchy distribution with PDF

$$f(x, \theta) = \frac{1}{\pi\{1 + (x - \theta)^2\}},$$

where θ is the location parameter. The log-likelihood is

$$l(\theta) = -\sum_{i=1}^{n} \log\{1 + (X_i - \theta)^2\} - n \log \pi.$$

The MLE is the solution of $s(\widehat{\theta}) = 0$, where

$$s(\theta) = 2 \sum_{i=1}^{n} \frac{X_i - \theta}{1 + (X_i - \theta)^2}.$$

Since $s(\theta) = 0$ does not admit an explicit solution, we adopt a Newton-Raphson scheme: $\theta_{k+1} = \theta_k - s(\theta_k)/\dot{s}(\theta_k)$, where

$$\dot{s}(\theta) = 2 \sum_{i=1}^{n} \frac{(X_i - \theta)^2 - 1}{\{1 + (X_i - \theta)^2\}^2}.$$

The *R*-function below implements the above scheme.

```
cauchyMLE <- function(n, theta, init, Tiny) {
x <- rcauchy(n, theta) # x is a sample
i <- 0   # No. of iterations
theta0 <- init +10*Tiny
theta1 <- init
while(abs(theta1-theta0)>Tiny) {
    theta0 <- theta1
    x2 <- x-theta0
    x22 <- x2*x2
    t1 <- mean(x2/(x22+1))          #  s(theta0)/(2n)
    t2 <- mean((x22-1)/(x22+1)^2)  #  derivation of s(theta0)/(2n)
    theta1 <- theta0 - t1/t2
    i <- i+1
    cat(i, "iteration:", theta1, "\n")  # print out iteration values
```

```
}
cat("MLE:", theta1, "No. of iterations:", i, "\n")
}
```

By calling `cauchyMLE(100,10,11.12,0.01)`, we excute the above iterative algorithm as follows:

```
> source("cauchyMLE.r")
> cauchyMLE(100, 10, 11.12, 0.01)
1 iteration: 9.594835
2 iteration: 10.08787
3 iteration: 10.0752
4 iteration: 10.07521

MLE: 10.07521 No. of iterations: 4
```

Note the initial value is important: the iterations will not converge if $\theta_0 < 8.75$ or $\theta_0 > 11.2$ on my PC.

Choosing a good initial value is always important. For this example, the PDF is symmetric around θ, it makes sense to consider either the sample mean or sample median as an initial estimate. However $E(X_1)$ is not well-defined, so the sample mean may not be a good estimator θ. Thus we may use the sample median as the initial value for our algorithm.

The Fisher Scoring method: replace $\dot{s}(\widehat{\boldsymbol{\theta}}_k)$ in (1) by $E_{\boldsymbol{\theta}}\{\dot{s}(\boldsymbol{\theta})\}$ under $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_k$. So the algorithm is now

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - [E\{\dot{s}(\boldsymbol{\theta}_k)\}]^{-1} s(\boldsymbol{\theta}_k) = \boldsymbol{\theta}_k + \{\mathcal{I}(\boldsymbol{\theta}_k)\}^{-1} s(\boldsymbol{\theta}_k),$$

where

$$\mathcal{I}(\boldsymbol{\theta}) = \mathrm{Var}\{s(\boldsymbol{\theta})\} = -E\{\dot{s}(\boldsymbol{\theta})\}$$

is the Fisher information.

Example 6 (continue). It can be shown that

$$E\{\dot{s}(\theta)\} = \frac{2n}{\pi} \int_{-\infty}^{\infty} \frac{(x-\theta)^2 - 1}{\{1 + (x-\theta)^2\}^3} dx = -n/2.$$

Hence the Fisher scoring method is

$$\theta_{k+1} = \theta_k + \frac{4}{n} \sum_{i=1}^{n} \frac{X_i - \theta_k}{1 + (X_i - \theta_k)^2}.$$

```
cauchyMLEscoring <- function(n, theta, init, Tiny) {
   # Fisher scoring MLE for Cauchy(theta)
   # n is sample size, Tiny is the tolerance limit controls
   # iteration estimates, init is the initial value used iteration
x <- rcauchy(n, theta)
i <- 0   # No. of iterations
theta0 <- init +10*Tiny
theta1 <- init
while(abs(theta1-theta0)>Tiny) {
    theta0 <- theta1
    x2 <- x-theta0
    t1 <- mean(x2/(x2*x2+1))         # s(theta0)/(2n)
    theta1 <- theta0 + 4*t1
    i <- i+1
    cat(i, "iteration:", theta1, "\n")  # print out iteration values
```

```
}
cat("\n", "MLE:", theta1, "No. of iterations:", i, "\n")
}
```

Calling `cauchyMLEscoring(100, 10, 15, 0.1)` yields

```
MLE: 10.14528 No.  of iterations:  7
```

Note now that the range of valid initial values is much bigger.

Like most iterative algorithms, the choice of appropriate **initial 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 Fisher scoring 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 Fisher scoring method is often larger, making the choice of initial values less important for the scoring method.

9.2.5 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 $\mathbf{X}' = (\mathbf{Y}', \mathbf{Z}')$ 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) **algorithms**

- *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 7. The genetic example from (Rao 1973, p.396) assumes that the phenotype data

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

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 **does not yield a closed form** $\widehat{\theta}$.

Now we treat $\mathbf{Y}$ as incomplete data from $\mathbf{X} = (X_1, \ldots, X_5)'$ 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 **X** is

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

which **readily 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

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

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 7 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.

Example 8. 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 $\mathbf{Y} = (Y_1, \cdots, Y_n)'$ is

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

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 $\mathbf{X}_i' = (Y_i, \mathbf{Z}_i')$, where $\mathbf{Z}_i = (Z_{i1}, \cdots, Z_{ik})'$ 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 $\mathbf{e}_\ell = (e_{\ell 1}, \cdots, e_{\ell k})'$. Let $\mathbf{X}' = (\mathbf{X}_1', \cdots, \mathbf{X}_n')$.

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

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

Note that

$$\begin{aligned} E(\mathbf{Z}_i | \mathbf{Y}, \boldsymbol{\theta}_0) &= \left(\frac{\alpha_1^0 f_1(Y_i, \boldsymbol{\lambda}_1^0)}{\sum_{\ell=1}^{k} \alpha_\ell^0 f_\ell(Y_i, \boldsymbol{\lambda}_\ell^0)}, \cdots, \frac{\alpha_k^0 f_k(Y_i, \boldsymbol{\lambda}_k^0)}{\sum_{\ell=1}^{k} \alpha_\ell^0 f_\ell(Y_i, \boldsymbol{\lambda}_\ell^0)} \right)' \\ &\equiv \left(b_1(Y_i, \boldsymbol{\theta}_0), \cdots, b_k(Y_i, \boldsymbol{\theta}_0) \right)', \end{aligned}$$

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

Now the E-step implies that

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

The M-step requires to maximise $Q(\boldsymbol{\theta})$, i.e.

$$\boldsymbol{\theta}_1 = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}).$$

Note that $\{\boldsymbol{\lambda}_j^1\}$ and $\{\alpha_j^1\}$ can be evaluated separately, which is much easier than minimizing $I(\boldsymbol{\theta}, \mathbf{Y})$ directly. For example,

$$\alpha_j^1 = \frac{\sum_{i=1}^{k} b_j(Y_i, \boldsymbol{\theta}_0)}{\sum_{\ell=1}^{k} \sum_{i=1}^{k} b_\ell(Y_i, \boldsymbol{\theta}_0)}, \qquad j = 1, \cdots, k.$$

When f_j is a normal PDF, $\boldsymbol{\lambda}_j^1$ admits an explicit formula.

Let $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_1$, keep iterating between E-step and M-step until two successive values of $\boldsymbol{\theta}_1$ differ by a small amount.

9.3 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 $\text{Var}_{\boldsymbol{\theta}}$ denote the probability distribution, expectation and variance under $\boldsymbol{\theta}_o = \boldsymbol{\theta}$.

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

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

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

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

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 (approximately) unbiased estimators only, and

- MSE (or its squared-root) should be used in general as

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

Ideally we would seek for the estimator which minimises MSE or MAE **for 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 $s(\boldsymbol{\theta})$ — **Fisher information** matrix:

$$\mathcal{I}(\boldsymbol{\theta}) = \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta})\} = E_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta})s(\boldsymbol{\theta})'\} = -E_{\boldsymbol{\theta}}\Big[\frac{\partial^2}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\log\{f(\mathbf{X},\boldsymbol{\theta})\}\Big],$$

because

$$E_{\boldsymbol{\theta}}\Big\{\frac{\partial^2}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}l(\boldsymbol{\theta})\Big\} = \int \frac{\ddot{L}(\boldsymbol{\theta})L(\boldsymbol{\theta}) - \dot{L}(\boldsymbol{\theta})\dot{L}(\boldsymbol{\theta})'}{L(\theta)}d\mathbf{x} = \int \ddot{L}(\boldsymbol{\theta})d\mathbf{x} - \int \frac{\dot{L}(\boldsymbol{\theta})\dot{L}(\boldsymbol{\theta})'}{L(\theta)}d\mathbf{x}$$
$$= -\int \frac{\dot{L}(\boldsymbol{\theta})\dot{L}(\boldsymbol{\theta})'}{L(\theta)}d\mathbf{x} = -\int s(\boldsymbol{\theta})s(\boldsymbol{\theta})'f(\mathbf{x},\boldsymbol{\theta})d\mathbf{x} = -E_{\boldsymbol{\theta}}\{s(\boldsymbol{\theta})s(\boldsymbol{\theta})'\}.$$

Fisher information $\mathcal{I}(\boldsymbol{\theta})$ measures **the information on θ contained in data X**. Further if $\mathbf{X} = (X_1, \cdots, X_n)'$, and $X_1, \cdots, X_n$ are IID,

$$\mathcal{I}(\boldsymbol{\theta}) = \mathcal{I}_{\mathbf{X}}(\boldsymbol{\theta}) = \sum_{j=1}^{n}\mathcal{I}_{X_j}(\boldsymbol{\theta}) = n\mathcal{I}_{X_1},$$

i.e. the information is additive.

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 $\theta \in \Theta$,

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

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

Important case: For any unbiased estimator $\widehat{\theta} = \widehat{\theta}(\mathbf{X})$,

$$\text{Var}(\widehat{\theta}) \geq 1/\mathcal{I}(\theta).$$

Multivariate case: For any unbiased estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathbf{X})$, $\mathrm{Var}(\widehat{\boldsymbol{\theta}}) - \{\mathcal{I}(\boldsymbol{\theta})\}^{-1}$ is a non-negative definite matrix. Hence $\mathrm{Var}(\widehat{\theta}_j) \geq \mathcal{I}^{jj}(\boldsymbol{\theta})$, where $\widehat{\theta}_j$ is the j-th component of $\widehat{\boldsymbol{\theta}}$, and $\mathcal{I}^{jj}(\boldsymbol{\theta})$ is the (j,j)-th element of $\{\mathcal{I}(\boldsymbol{\theta})\}^{-1}$.

Example 9. 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

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

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 **approximately** 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 **efficient**.

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

Bootstrapping MSEs — Parametric bootstrap

An MSE provides a measure for the accuracy of the estimator. But it is not always feasible to derive an explicit expression for MSE. Furthermore it depends on the unknown parameter. Alternatively we may estimate the MSE by Bootstrapping.

Let $X_1, \cdots, X_n$ be i.i.d. with PDF $f(\cdot, \theta)$, where θ is a scalar. 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, v 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

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

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

The **basic idea of parametric 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

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

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

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 v^* is calculated as

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

Remark. The bootstrap methods introduced in Chapter 8 are in the category of *Nonparametric Bootstrap*. If we know the form of the underlying distribution, parametric bootstrap methods are typically more efficient.

Bootstrap is a powerful tool for statistical inference. It has different forms for different applications. The diagram above indicates that the basic idea of a *parametric bootstrap* method.