One can also show that the EM algorithm always converges to a local or global maximum, or at least to a saddlepoint of the log-likelihood. However, the convergence can be quite slow; typically, more iterations are required than for the Newton–Raphson algorithm. Another disadvantage is that the algorithm does not automatically give the observed Fisher information. Of course, this can be calculated after convergence if the second derivative of the log-likelihood $l(\theta; x)$ of the observed data x is available.

2.4 Quadratic Approximation of the Log-Likelihood Function

An important approximation of the log-likelihood function is based on a quadratic function. To do so, we apply a Taylor approximation of second order (cf. Appendix B.2.3) around the MLE $\hat{\theta}_{\mathrm{ML}}$:

$$\begin{aligned} l(\theta) &\approx l(\hat{\theta}_{\mathrm{ML}}) + \frac{dl(\hat{\theta}_{\mathrm{ML}})}{d\theta}(\theta - \hat{\theta}_{\mathrm{ML}}) + \frac{1}{2}\frac{d^2 l(\hat{\theta}_{\mathrm{ML}})}{d\theta^2}(\theta - \hat{\theta}_{\mathrm{ML}})^2 \\ &= l(\hat{\theta}_{\mathrm{ML}}) + S(\hat{\theta}_{\mathrm{ML}})(\theta - \hat{\theta}_{\mathrm{ML}}) - \frac{1}{2} \cdot I(\hat{\theta}_{\mathrm{ML}})(\theta - \hat{\theta}_{\mathrm{ML}})^2. \end{aligned}$$

Due to $S(\hat{\theta}_{\mathrm{ML}}) = 0$, the quadratic approximation of the relative log-likelihood is

$$\tilde{l}(\theta) = l(\theta) - l(\hat{\theta}_{\mathrm{ML}}) \approx -\frac{1}{2} \cdot I(\hat{\theta}_{\mathrm{ML}})(\theta - \hat{\theta}_{\mathrm{ML}})^2. \tag{2.17}$$

Example 2.15 (Poisson model) Assume that we have one observation $x = 11$ from a Poisson distribution $\mathrm{Po}(e\lambda)$ with known offset $e = 3.04$ and unknown parameter λ. The MLE of λ is $\hat{\lambda}_{\mathrm{ML}} = x/e = 3.62$, cf. Example 2.4. The observed Fisher information turns out to be $I(\hat{\lambda}_{\mathrm{ML}}) = x/\hat{\lambda}^2_{\mathrm{ML}}$, so that the quadratic approximation of the relative log-likelihood is

$$\tilde{l}(\lambda) \approx -\frac{1}{2}\frac{x}{\hat{\lambda}^2_{\mathrm{ML}}}(\lambda - \hat{\lambda}_{\mathrm{ML}})^2.$$

Figure 2.7 displays $\tilde{l}(\lambda)$ and its quadratic approximation. ■

Example 2.16 (Normal model) Let $X_{1:n}$ denote a random sample from a normal distribution $\mathrm{N}(\mu, \sigma^2)$ with unknown mean μ and known variance σ^2. We know from Example 2.9 that

$$\begin{aligned} l(\mu) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \\ &= -\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right\}, \end{aligned}$$

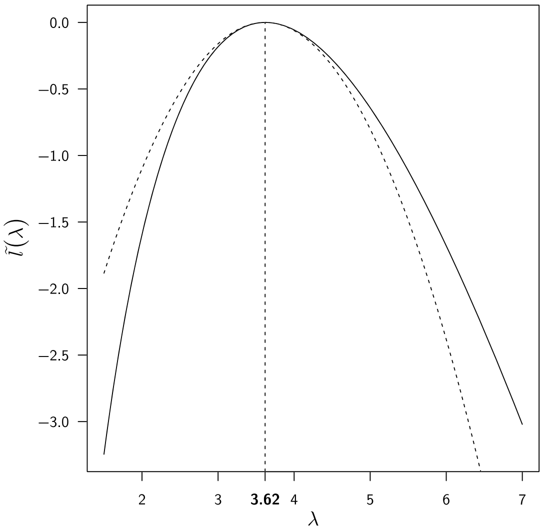


Fig. 2.7 Relative log-likelihood $\tilde{l}(\lambda)$ and its quadratic approximation (*dashed line*) for a single observation $x = 11$ from a Poisson distribution with mean $e\lambda$ and known offset $e = 3.04$

$$l(\hat{\mu}_{\mathrm{ML}}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \text{and hence}$$

$$\tilde{l}(\mu) = l(\mu) - l(\hat{\mu}_{\mathrm{ML}}) = -\frac{n}{2\sigma^2}(\bar{x} - \mu)^2,$$

but we also have

$$-\frac{1}{2} \cdot I(\hat{\mu}_{\mathrm{ML}})(\mu - \hat{\mu}_{\mathrm{ML}})^2 = -\frac{n}{2\sigma^2}(\mu - \bar{x})^2.$$

Both sides of Eq. (2.17) are hence identical, so the quadratic approximation is here exact. ■

Under certain regularity conditions, which we will not discuss here, it can be shown that a quadratic approximation of the log-likelihood improves with increasing sample size. The following example illustrates this phenomenon in the binomial model.

Example 2.17 (Binomial model) Figure 2.8 displays the relative log-likelihood of the success probability π in a binomial model with sample size $n = 10, 50, 200, 1000$. The observed datum x has been fixed at $x = 8, 40, 160, 800$ such that the MLE of π is $\hat{\pi}_{\mathrm{ML}} = 0.8$ in all four cases. We see that the quadratic approximation of the relative log-likelihood improves with increasing sample size n. The two functions are nearly indistinguishable for $n = 1000$. ■

The advantage of the quadratic approximation of the relative log-likelihood lies in the fact that we only need to know the MLE $\hat{\theta}_{\mathrm{ML}}$ and the observed Fisher information $I(\hat{\theta}_{\mathrm{ML}})$, no matter what the actual log-likelihood looks like. However, in certain pathological cases the approximation may remain poor even if the sample size increases.

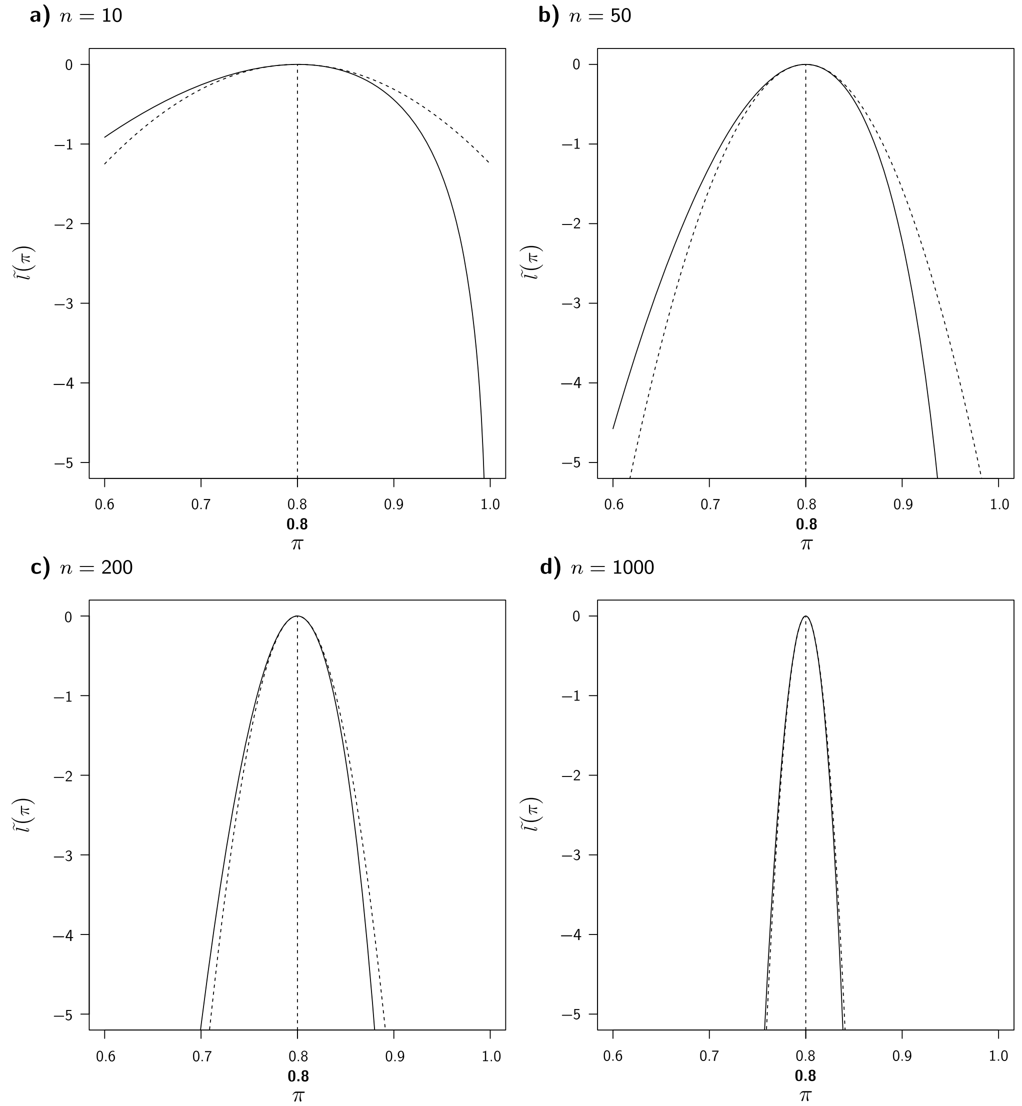


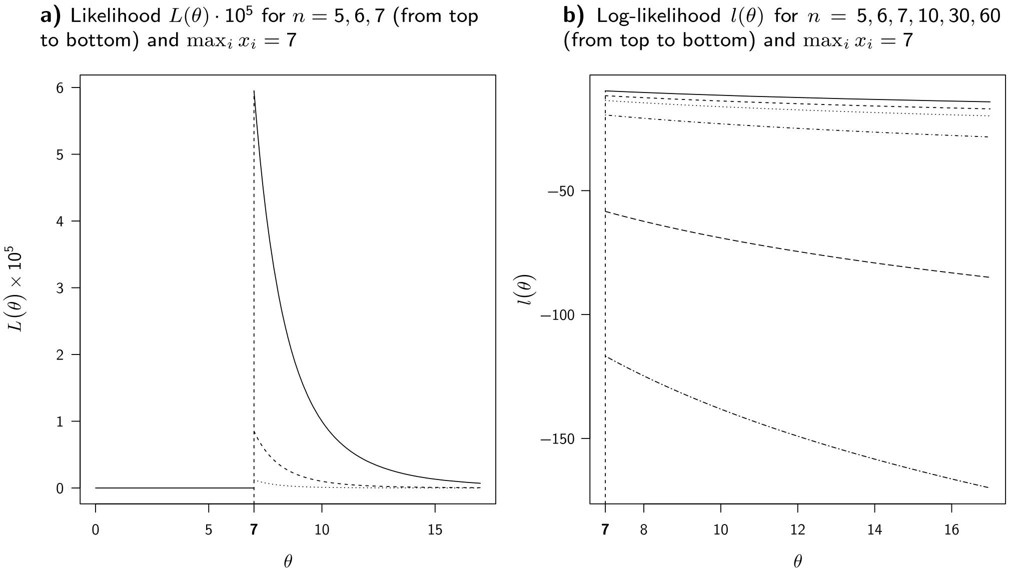
Fig. 2.8 Quadratic approximation (*dashed line*) of the relative log-likelihood (*solid line*) of the success probability π in a binomial model

Example 2.18 (Uniform model) Let $X_{1:n}$ denote a random sample from a continuous uniform distribution $\mathrm{U}(0, \theta)$ with unknown upper limit $\theta \in \mathbb{R}^+$. The density function of the uniform distribution is

$$f(x; \theta) = \frac{1}{\theta} \mathsf{I}_{[0,\theta)}(x)$$

with *indicator function* $\mathsf{I}_A(x)$ equal to one if $x \in A$ and zero otherwise. The likelihood function of θ is

$$L(\theta) = \begin{cases} \prod_{i=1}^{n} f(x_i; \theta) = \theta^{-n} & \text{for } \theta \geq \max_i(x_i), \\ 0 & \text{otherwise,} \end{cases}$$


a) Likelihood $L(\theta) \cdot 10^5$ for $n = 5, 6, 7$ (from top to bottom) and $\max_i x_i = 7$

b) Log-likelihood $l(\theta)$ for $n = 5, 6, 7, 10, 30, 60$ (from top to bottom) and $\max_i x_i = 7$

Fig. 2.9 Likelihood and log-likelihood function for a random sample of different size n from a uniform distribution with unknown upper limit θ. Quadratic approximation of the log-likelihood is impossible even for large n

with MLE $\hat{\theta}_{\text{ML}} = \max_i(x_i)$, cf. Fig. 2.9a.

The derivatives of the log-likelihood function

$$l(\theta) = -n \log(\theta) \quad \text{for } \theta > \max_i(x_i)$$

are

$$S(\hat{\theta}_{\text{ML}}) = \frac{dl(\hat{\theta}_{\text{ML}})}{d\theta} \neq 0 \quad \text{and} \quad -I(\hat{\theta}_{\text{ML}}) = \frac{d^2 l(\hat{\theta}_{\text{ML}})}{d\theta^2} = \frac{n}{\hat{\theta}_{\text{ML}}^2} > 0,$$

so the log-likelihood $l(\theta)$ is not concave but convex, with negative (!) observed Fisher information, cf. Fig. 2.9b. It is obvious from Fig. 2.9b that a quadratic approximation to $l(\theta)$ will remain poor even if the sample size n increases. The reason for the irregular behaviour of the likelihood function is that the support of the uniform distribution depends on the unknown parameter θ. ■

2.5 Sufficiency

Under certain regularity conditions, a likelihood function can be well characterised by the MLE and the observed Fisher information. However, Example 2.18 illustrates that this is not always the case. An alternative characterisation of likelihood functions is in terms of *sufficient statistics*, a concept which we will introduce in the following. We will restrict our attention to random samples, but the description could be easily generalised if required.