Chapter 4

The Maximum Likelihood Estimator

4.1 The Maximum likelihood estimator

As illustrated in the exponential family of distributions, discussed above, the maximum likelihood estimator of θ_0 (the true parameter) is defined as

$$\hat{\theta}_T = \arg\max_{\theta\in\Theta} L_T(\underline{X};\theta) = \arg\max_{\theta\in\Theta} \mathcal{L}_T(\theta).$$

Often we find that $\frac{\partial \mathcal{L}_T(\theta)}{\partial\theta}\rfloor_{\theta=\hat{\theta}_T} = 0$, hence solution can be obtained by solving the derivative of the log likelihood (often called the *score function*). However, if θ_0 lies on or close to the boundary of the parameter space this will not necessarily be true.

Below we consider the sampling properties of $\hat{\theta}_T$ when the true parameter θ_0 lies in the interior of the parameter space Θ.

We note that the likelihood is invariant to transformations of the data. For example if X has the density $f(\cdot;\theta)$ and we define the transformed random variable $Z = g(X)$, where the function g has an inverse (its a 1-1 transformation), then it is easy to show that the density of Z is $f(g^{-1}(z);\theta)\frac{\partial g^{-1}(z)}{\partial z}$. Therefore the likelihood of $\{Z_t = g(X_t)\}$ is

$$\prod_{t=1}^{T} f(g^{-1}(Z_t);\theta)\frac{\partial g^{-1}(z)}{\partial z}\rfloor_{z=Z_t} = \prod_{t=1}^{T} f(X_t;\theta)\frac{\partial g^{-1}(z)}{\partial z}\rfloor_{z=Z_t}.$$

Hence it is proportional to the likelihood of $\{X_t\}$ and the maximum of the likelihood of $\{Z_t = g(X_t)\}$ is the same as the maximum of the likelihood of $\{X_t\}$.

4.1.1 Evaluating the MLE

Examples

Example 4.1.1 *$\{X_t\}$ are iid random variables, which follow a Normal (Gaussian) distribution $\mathcal{N}(\mu, \sigma^2)$. The likelihood is proportional to*

$$\mathcal{L}_T(\underline{X}; \mu, \sigma^2) \quad = \quad -T \log \sigma - \frac{1}{2\sigma^2} \sum_{t=1}^{T} (X_t - \mu)^2.$$

Maximising the above with respect to μ and σ^2 gives $\hat{\mu}_T = \bar{X}$ and $\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^{T} (X_t - \bar{X})^2$.

Example 4.1.2 *Question:*

$\{X_t\}$ are iid random variables, which follow a Weibull distribution, which has the density

$$\frac{\alpha y^{\alpha-1}}{\theta^\alpha} \exp(-(y/\theta)^\alpha) \qquad \theta, \alpha > 0.$$

Suppose that α is <u>known</u>, but θ is <u>unknown</u> (and we need to estimate it). What is the maximum likelihood estimator of θ?

Solution:

The log-likelihood (of interest) is proportional to

$$\begin{aligned} \mathcal{L}_T(\underline{X}; \theta) \quad &= \quad \sum_{t=1}^{T} \Big(\log \alpha + (\alpha - 1) \log Y_t - \alpha \log \theta - \big(\frac{Y_t}{\theta}\big)^\alpha \Big) \\ &\propto \quad \sum_{t=1}^{T} \Big(- \alpha \log \theta - \big(\frac{Y_t}{\theta}\big)^\alpha . \Big). \end{aligned}$$

The derivative of the log-likelihood wrt to θ is

$$\frac{\partial \mathcal{L}_T}{\partial \theta} = -\frac{T\alpha}{\theta} + \frac{\alpha}{\theta^{\alpha+1}} \sum_{t=1}^{T} Y_t^\alpha = 0.$$

Solving the above gives $\hat{\theta}_T = (\frac{1}{T} \sum_{t=1}^{T} Y_t^\alpha)^{1/\alpha}$.

Example 4.1.3 *Notice that if α is given, an explicit solution for the maximum of the likelihood, in the above example, can be obtained. Consider instead the maximum of the likelihood with respect to α and θ, ie.*

$$\arg\max_{\theta,\alpha} \sum_{t=1}^{T} \Big(\log \alpha + (\alpha - 1) \log Y_t - \alpha \log \theta - \big(\frac{Y_t}{\theta}\big)^\alpha \Big).$$

The derivative of the likelihood is

$$\begin{aligned}\frac{\partial \mathcal{L}_T}{\partial \theta} &= -\frac{T\alpha}{\theta} + \frac{\alpha}{\theta^{\alpha+1}}\sum_{t=1}^{T} Y_t^{\alpha} = 0 \\ \frac{\partial \mathcal{L}_T}{\partial \alpha} &= \frac{T}{\alpha} - \sum_{t=1}^{T}\log Y_t - T\log\theta - \frac{T\alpha}{\theta} + \sum_{t=1}^{T}\log(\frac{Y_t}{\theta}) \times (\frac{Y_t}{\theta})^{\alpha} = 0.\end{aligned}$$

It is clear that an explicit expression to the solution of the above does not exist and we need to find alternative methods for finding a solution. Below we shall describe numerical routines which can be used in the maximisation. In special cases, one can use other methods, such as the Profile likelihood (we cover this later on).

Numerical Routines

In an ideal world to maximise a likelihood, we would consider the derivative of the likelihood and solve it ($\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta}\rfloor_{\theta=\hat{\theta}_T} = 0$), and an explicit expression would exist for this solution. In reality this rarely happens (as we illustrated in the section above).

Usually, we will be unable to obtain an explicit expression for the MLE. In such cases, one has to do the maximisation using alternative, numerical methods. Typically it is relative straightforward to maximise the likelihood of random variables which belong to the exponential family (numerical algorithms sometimes have to be used, but they tend to be fast and attain the maximum of the likelihood - not just the local maximum). However, the story becomes more complicated even if we consider mixtures of exponential family distributions - these do not belong to the exponential family, and can be difficult to maximise using conventional numerical routines. We give an example of such a distribution here. Let us suppose that $\{X_t\}$ are iid random variables which follow the classical normal mixture distribution

$$f(y;\theta) = pf_1(y;\theta_1) + (1-p)f_2(y;\theta_2),$$

where f_1 is the density of the normal with mean μ_1 and variance σ_1^2 and f_2 is the density of the normal with mean μ_2 and variance σ_2^2. The log likelihood is

$$\mathcal{L}_T(\underline{Y};\theta) = \sum_{t=1}^{T}\log\left(p\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp(-\frac{1}{2\sigma_1^2}(X_t-\mu_1)^2) + (1-p)\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp(-\frac{1}{2\sigma_2^2}(X_t-\mu_2)^2)\right).$$

Studying the above it is clear there does not explicit solution to the maximum. Hence one needs to use a numerical algorithm to maximise the above likelihood.

We discuss a few such methods below.

The Newton Raphson Routine The Newton-Raphson routine is the standard method to numerically maximise the likelihood, this can often be done automatically in `R` by using the `R` functions `optim` or `nlm`. To apply Newton-Raphson, we have to assume that the derivative of the likelihood exists (this is not always the case - think about the ℓ_1-norm based estimators!) and the minimum lies inside the parameter space such that $\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta}\rfloor_{\theta=\hat{\theta}_T} = 0$. We choose an initial value θ_1 and apply the routine

$$\theta_n = \theta_{n-1} + \left(\frac{\partial^2 \mathcal{L}_T(\theta)}{\partial \theta^2}\rfloor_{\theta_{n-1}}\right)^{-1} \frac{\partial \mathcal{L}_T(\theta_{n-1})}{\partial \theta}\rfloor_{\theta_{n-1}}.$$

Where this routine comes from will be clear by using the Taylor expansion of $\frac{\partial \mathcal{L}_T(\theta_{n-1})}{\partial \theta}$ about θ_0 (see Section 4.1.3). If the likelihood has just one global maximum and no local maximums (hence it is convex), then it is quite easy to maximise. If on the other hand, the likelihood has a few local maximums and the initial value θ_1 is not chosen close enough to the true maximum, then the routine may converge to a local maximum (not good!). In this case it may be a good idea to do the routine several times for several different initial values $\theta_1^{(i)}$. For each convergence value $\hat{\theta}_T^{(i)}$ evaluate the likelihood $\mathcal{L}_T(\hat{\theta}_T^{(i)})$ and select the value which gives the largest likelihood. It is best to avoid these problems by starting with an informed choice of initial value.

Implementing without any thought a Newton-Rapshon routine can lead to estimators which take an incredibly long time to converge. If one carefully considers the likelihood one can shorten the convergence time by rewriting the likelihood and using faster methods (often based on the Newton-Raphson).

Iterative least squares This is a method that we shall describe later when we consider Generalised linear models. As the name suggests the algorithm has to be interated, however at each step weighted least squares is implemented (see later in the course).

The EM-algorithm This is done by the introduction of dummy variables, which leads to a new 'unobserved' likelihood which can easily be maximised. In fact one the simplest methods of maximising the likelihood of mixture distributions is to use the EM-algorithm.

We cover this later in the course.

See Example 4.23 on page 117 in Davison (2002).

The likelihood for dependent data

We mention that the likelihood for dependent data can also be constructed (though often the estimation and the asymptotic properties can be a lot harder to derive). Using Bayes rule (ie.

$P(A_1, A_2, \ldots, A_T) = P(A_1) \prod_{i=2}^{T} P(A_i|A_{i-1}, \ldots, A_1))$ we have

$$L_T(\underline{X}; \theta) = f(X_1; \theta) \prod_{t=2}^{T} f(X_t|X_{t-1}, \ldots, X_1; \theta).$$

Under certain conditions on $\{X_t\}$ the structure above $\prod_{t=2}^{T} f(X_t|X_{t-1}, \ldots, X_1; \theta)$ can be simplified. For example if X_t were Markovian then we have X_t conditioned on the past on depends the most recent past observation, ie. $f(X_t|X_{t-1}, \ldots, X_1; \theta) = f(X_t|X_{t-1}; \theta)$ in this case the above likelihood reduces to

$$L_T(\underline{X}; \theta) = f(X_1; \theta) \prod_{t=2}^{n} f(X_t|X_{t-1}; \theta). \tag{4.1}$$

Example 4.1.4 *A lot of the material we will cover in this class will be for independent observations. However likelihood methods can also work for dependent observations too. Consider the $AR(1)$ time series*

$$X_t = aX_{t-1} + \varepsilon_t,$$

where ε_t are iid random variables with mean zero. We will assume that $|a| < 1$.

We see from the above that the observation X_{t-1} as a linear influence on the next observation and it is Markovian, that it given X_{t-1}, the random variable X_{t-2} has no influence on X_t (to see this consider the distribution function $P(X_t \leq x|X_{t-1}, X_{t-2})$). Therefore by using (4.1) the likelihood of $\{X_t\}_t$ is

$$L_T(\underline{X}; a) = f(X_1; a) \prod_{t=2}^{T} f_\varepsilon(X_t - aX_{t-1}), \tag{4.2}$$

where f_ε is the density of ε and $f(X_1; a)$ is the marginal density of X_1. This means the likelihood of $\{X_t\}$ only depends on f_ε and the marginal density of X_t. We use $\hat{a}_T = \arg\max L_T(\underline{X}; a)$ as the mle estimator of a.

Often we ignore the term $f(X_1; a)$ (because this is often hard to know - try and figure it out - its relatively easy in the Gaussian case) and consider what is called the conditional likelihood

$$Q_T(\underline{X}; a) = \prod_{t=2}^{T} f_\varepsilon(X_t - aX_{t-1}). \tag{4.3}$$

$\tilde{a}_T = \arg\max L_T(\underline{X}; a)$ *as the quasi-mle estimator of a.*

Exercise: What is the quasi-likelihood proportional to in the case that $\{\varepsilon_t\}$ are Gaussian random variables with mean zero. It should be mentioned that often the conditional likelihood is derived as if the errors $\{\varepsilon_t\}$ are Gaussian - even if they are not. This is often called the quasi or pseudo likelihood.

4.1.2 A quick review of the central limit theorem

In this section we will not endeavour to proof the central limit theorem (which is usually based on showing that the characteristic function - a close cousin of the moment generating function - of the average converges to the characteristic function of the normal distribution). However, we will recall the general statement of the CLT and generalisations of it. The purpose of this section is not to lumber you with unnecessary mathematics but to help you understand when an estimator is close to normal (or not).

Lemma 4.1.1 (The famous CLT) *Let us suppose that $\{X_t\}$ are iid random variables, let $\mu = \mathbb{E}(X_t) < \infty$ and $\sigma^2 = var(X_t) < \infty$. Define $\bar{X} = \frac{1}{T}\sum_{t=1}^{T} X_t$. Then we have*

$$\sqrt{T}(\bar{X} - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

alternatively $(\bar{X} - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \frac{\sigma^2}{T})$.

What this means that if we have a large enough sample size and plotted the histogram of several replications of the average, this should be close to normal.

Remark 4.1.1 *(i) The above lemma appears to be 'restricted' to just averages. However, it can be used in several different contexts. Averages arise in several different situations. It is not just restricted to the average of the observations. By judicious algebraic manipulations, one can show that several estimators can be rewritten as an average (or approximately as an average). At first appearance, the MLE of the Weibull parameters given in Example 4.1.3) does not look like an average, however, in the section we will consider the general maximum likelihood estimators, and show that they can be rewritten as an average hence the CLT applies to them too.*

(ii) The CLT can be extended in several ways.

- *(a) To random variables whose variance are not all the same (ie. indepedent but identically distributed random variables).*
- *(b) Dependent random variables (so long as the dependency 'decays' in some way).*
- *(c) To not just averages but weighted averages too (so long as the weight depends in certain way). However, the weights should be 'distributed well' over all the random variables. Ie. suppose that $\{X_t\}$ are iid random variables. Then it is clear that $\frac{1}{10}\sum_{t=1}^{10} X_t$ will never be normal (unless $\{X_t\}$ is normal - observe 10 is fixed!), but it seems plausible that $\frac{1}{n}\sum_{t=1}^{n} \sin(2\pi t/12) X_t$ is normal (despite this not being the sum of iid random variables).*

- *There exists several theorems which one can use to prove normality. But really the take home message is, look at your estimator and see whether asymptotic normality it looks plausible - you could even check it through simulations.*

Example 4.1.5 (Some problem cases) *One should think a little before blindly applying the CLT. Suppose that the iid random variables $\{X_t\}$ follow a t-distribution with 2 degrees of freedom, ie. the density function is*

$$f(x) = \frac{\Gamma(3/2)}{\sqrt{2\pi}}(1+\frac{x^2}{2})^{-3/2}.$$

Let $\bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$ denote the sample mean. It is well known that the mean of the t-distribution with two degrees of freedom exists, but the variance does not (it is too thick tailed). Thus, the assumptions required for the CLT to hold are violated and $\bar{X}$ is not normally distributed (in fact it follows a stable law distribution). Intuitively this is clear, recall that the chance of outliers for a t-distribution with a small number of degrees of freedom, if large. This makes it impossible that even averages should be 'well behaved' (there is a large chance that an average could also be too large or too small).

To see why the variance is infinite, study the form of t-distribution (with two degrees). For the variance to be finite, the tails of the distribution should converge to zero fast enough (in other words the probability of outliers should not be too large). See that the tails of the t-distribution (for large x) behaves like $f(x) \sim Cx^{-3}$ (make a plot in Maple to check), thus the second moment $\mathbb{E}(X^2) \geq \int_M^\infty Cx^{-3}x^2dx = \int_M^\infty Cx^{-1}dx$ (for some C and M), is clearly not finite! This argument can be made precise.

4.1.3 The Taylor series expansion - the statisticians tool

The Taylor series is used all over the place in statistics and you should be completely fluent with using it. It can be used to prove consistency of an estimator, normality (based on the assumption that averages converge to a normal distribution), obtaining the limiting variance of an estimator etc. We start by demonstrating its use for the log likelihood.

We recall that the mean value (in the univariate case) states that

$$f(x) = f(x_0) + (x-x_0)f'(\bar{x}_1) \qquad f(x) = f(x_0) + (x-x_0)f'(x_0) + \frac{(x-x_0)^2}{2}f''(\bar{x}_2),$$

where $\bar{x}_1$ and $\bar{x}_2$ both lie between x and x_0. In the case that f is a multivariate function, then we have

$$\begin{aligned} f(\underline{x}) &= f(\underline{x}_0) + (\underline{x}-\underline{x}_0)\nabla f(\underline{x})\rfloor_{\underline{x}=\underline{\bar{x}}_1} \\ f(\underline{x}) &= f(\underline{x}_0) + (\underline{x}-\underline{x}_0)'\nabla f(\underline{x})\rfloor_{\underline{x}=\underline{x}_0} + \frac{1}{2}(\underline{x}-\underline{x}_0)'\nabla^2 f(\underline{x})\rfloor_{\underline{x}=\underline{\bar{x}}_2}(\underline{x}-\underline{x}_0), \end{aligned}$$

where $\bar{\underline{x}}_1$ and $\bar{\underline{x}}_2$ both lie between $\underline{x}$ and $\underline{x}_0$. In the case that $f(\underline{x})$ is a vector, then the mean value theorem does not directly work. Strictly speaking we cannot say that

$$\underline{f}(\underline{x}) \quad = \quad \underline{f}(\underline{x}_0) + (\underline{x} - \underline{x}_0)' \nabla \underline{f}(\underline{x}) \rfloor_{\underline{x}=\bar{\underline{x}}_1},$$

where $\bar{\underline{x}}_1$ lies between $\underline{x}$ and $\underline{x}_0$. However, it is quite straightforward to overcome this inconvience. The mean value theorem does hold pointwise, for every element of the vector $\underline{f}(\underline{x}) = (f_1(\underline{x}), \ldots, f_d(\underline{x}))$, ie. for every $1 \leq i \leq d$ we have

$$f_i(\underline{x}) \quad = \quad f_i(\underline{x}_0) + (\underline{x} - \underline{x}_0) \nabla f_i(\underline{x}) \rfloor_{\underline{x}=\bar{\underline{x}}_i},$$

where $\bar{\underline{x}}_i$ lies between $\underline{x}$ and $\underline{x}_0$. Thus if $\nabla f_i(\underline{x}) \rfloor_{\underline{x}=\bar{\underline{x}}_i} \rightarrow \nabla f_i(\underline{x}) \rfloor_{\underline{x}=\underline{x}_0}$, we do have that

$$\underline{f}(\underline{x}) \quad \approx \quad \underline{f}(\underline{x}_0) + (\underline{x} - \underline{x}_0)' \nabla \underline{f}(\underline{x}) \rfloor_{\underline{x}=\underline{x}_0}.$$

We use the above below.

- Application 1 (An expression for $\mathcal{L}_T(\hat{\theta}_T) - \mathcal{L}_T(\theta_0)$ in terms of $(\hat{\theta}_T - \theta_0)$):

  The expansion of $\mathcal{L}_T(\hat{\theta}_T)$ about θ_0 (the true parameter)

  $$\mathcal{L}_T(\theta_0) - \mathcal{L}_T(\hat{\theta}_T) = \frac{\partial \mathcal{L}_T(\theta)}{\partial \theta} \rfloor_{\hat{\theta}_T} (\theta_0 - \hat{\theta}_T) + \frac{1}{2} (\theta_0 - \hat{\theta}_T)' \frac{\partial^2 \mathcal{L}_T(\theta)}{\partial \theta^2} \rfloor_{\bar{\theta}_T} (\theta_0 - \hat{\theta}_T)$$

  where $\bar{\theta}_T$ lies between θ_0 and $\hat{\theta}_T$. If $\hat{\theta}_T$ lies in the interior of the parameter space (this is an extremely important assumption here) then $\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta} \rfloor_{\hat{\theta}_T} = 0$. Moreover, if it can be shown that $|\hat{\theta}_T - \theta_0| \stackrel{\mathcal{P}}{\rightarrow} 0$ (we show this in the section below), then under certain conditions on $\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta}$ (such as the existence of the third derivative etc.) it can be shown that $\frac{\partial^2 \mathcal{L}_T(\theta)}{\partial \theta^2} \rfloor_{\bar{\theta}_T} \stackrel{\mathcal{P}}{\rightarrow} \mathbb{E}(\frac{\partial^2 \mathcal{L}_T(\theta)}{\partial \theta^2} \rfloor_{\theta_0}) = I(\theta_0)$. Hence the above is roughly

  $$2(\mathcal{L}_T(\hat{\theta}_T) - \mathcal{L}_T(\theta_0)) \approx (\hat{\theta}_T - \theta_0)' I(\theta_0) (\hat{\theta}_T - \theta_0)$$

  Note that in many of the derivations below we will use that

  $$\frac{\partial^2 \mathcal{L}_T(\theta)}{\partial \theta^2} \rfloor_{\bar{\theta}_T} \stackrel{\mathcal{P}}{\rightarrow} \mathbb{E}(\frac{\partial^2 \mathcal{L}_T(\theta)}{\partial \theta^2} \rfloor_{\theta_0}) = -I(\theta_0).$$

  But it should be noted that this only true if (i) $|\hat{\theta}_T - \theta_0| \stackrel{\mathcal{P}}{\rightarrow} 0$ and (ii) $\frac{\partial^2 \mathcal{L}_T(\theta)}{\partial \theta^2}$ converges uniformly to $\mathbb{E}(\frac{\partial^2 \mathcal{L}_T(\theta)}{\partial \theta^2} \rfloor_{\theta_0})$.

  We consider below another closely related application.

- Application 2 (An expression for $(\hat{\theta}_T - \theta_0)$ in terms of $\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta}\rfloor_{\theta_0}$):

  The expansion of the p-dimension vector $\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta}\rfloor_{\hat{\theta}_T}$ pointwise about θ_0 (the true parameter) gives (for $1 \leq i \leq d$)

$$\frac{\partial \mathcal{L}_{i,T}(\theta)}{\partial \theta}\rfloor_{\hat{\theta}_T} = \frac{\partial \mathcal{L}_{i,T}(\theta)}{\partial \theta}\rfloor_{\theta_0} + \frac{\partial \mathcal{L}_{i,T}(\theta)}{\partial \theta}\rfloor_{\bar{\theta}_T}(\hat{\theta}_T - \theta_0).$$

  Now by using the same argument as in Application 1 we have

$$\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta}\rfloor_{\theta_0} \approx I(\theta_0)(\hat{\theta}_T - \theta_0).$$

  We mention that $U_T(\theta_0) = \frac{\partial \mathcal{L}_T(\theta)}{\partial \theta}\rfloor_{\theta_0}$ is often called the *score or U statistic*. And we see that the asymptotic sampling properties of U_T determine the sampling properties of $(\hat{\theta}_T - \theta_0)$.

Example 4.1.6 (The Weibull) *Evaluate the second derivative of the likelihood given in Example 4.1.3, take the expection on this, $I(\theta, \alpha) = \mathbb{E}(\nabla^2 \mathcal{L}_T)$ (we use the ∇ to denote the second derivative with respect to the parameters α and θ). Exercise: Evaluate $I(\theta, \alpha)$.*

Application 2 implies that the maximum likelihood estimators $\hat{\theta}_T$ and $\hat{\alpha}_T$ (recalling that no explicit expression for them exists) can be written as

$$\begin{pmatrix} \hat{\theta}_T - \theta \\ \hat{\alpha}_T - \alpha \end{pmatrix} \approx I(\theta, \alpha)^{-1} \begin{pmatrix} \sum_{t=1}^T \left(-\frac{\alpha}{\theta} + \frac{\alpha}{\theta^{\alpha+1}} Y_t^\alpha \right) \\ \sum_{t=1}^T \left(\frac{1}{\alpha} - \log Y_t - \log \theta - \frac{\alpha}{\theta} + \log(\frac{Y_t}{\theta}) \times (\frac{Y_t}{\theta})^\alpha \right) \end{pmatrix}$$

4.1.4 Sampling properties of the maximum likelihood estimator

See also Section 4.4.2 (p118), Davison (2002). These proofs will not be examined, but you should have some idea why Theorem 4.1.2 is true.

We have shown that under certain conditions the maximum likelihood estimator can often be the minimum variance unbiased estimator (for example, in the case of exponential family of distributions). However, for finite samples, the mle may not attain the C-R lower bound. Hence for finite sample $\text{var}(\hat{\theta}_T) > I(\theta)^{-1}$. However, it can be shown that asymptotically the variance of the mle attains the mle lower bound. In other words, for large samples, the variance of the mle is close to the Cramer-Rao bound. We will prove the result in the case that ℓ_T is the log likelihood of independent, identically distributed random variables. The proof can be generalised to the case of non-identically distributed random variables.

We first state sufficient conditions for this to be true.

Assumption 4.1.1 *[Regularity Conditions 2] Let $\{X_t\}$ be iid random variables with density $f(X; \theta)$.*

(i) Suppose the conditions in Assumption 1.1.1 hold.

(ii) **Almost sure uniform convergence** *(This is optional)*

For every $\varepsilon > 0$ there exists a δ such that

$$P\Big(\lim_{T\to\infty}\sup_{|\theta_1-\theta_2|>\delta}\big|\frac{1}{T}\mathcal{L}_T(\underline{X};\theta) - \mathbb{E}(\mathcal{L}_T(\theta))\big| > \varepsilon\Big) \to 0.$$

We mention that directly verifying uniform convergence can be difficult. However, it can be established by showing that the parameter space is compact, point wise convergence of the likelihood to its expectation and almost sure equicontinuity in probability.

(iii) **Model identifiability**

For every $\theta \in \Theta$, there does not exist another $\tilde{\theta} \in \Theta$ such that $f(x;\theta) = f(x;\tilde{\theta})$ for all x.

(iv) The parameter space Θ is finite and compact.

(v) $\sup \mathbb{E}|\frac{1}{T}\frac{\partial \mathcal{L}_T(\underline{X};\theta)}{\partial\theta}| < \infty$.

We require Assumption 4.1.1(ii,iii) to show consistency and Assumptions 1.1.1 and 4.1.1(iii-v) to show asymptotic normality.

Theorem 4.1.1 *Supppose Assumption 4.1.1(ii,iii) holds. Let θ_0 be the true parameter and $\hat{\theta}_T$ be the mle. Then we have $\hat{\theta}_T \overset{a.s.}{\to} \theta_0$ (consistency).*

PROOF. To prove the result we first need to show that the expectation of the maximum likelihood is maximum at the true parameter and that this is the unique maximum. In other words we need to show that $\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta)) - \mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0)) \leq 0$ for all $\theta \in \Theta$. To do this, we have

$$\begin{aligned}\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta)) - \mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0)) &= \int \log\frac{f(x;\theta)}{f(x;\theta_0)} f(x;\theta_0)dx \\ &= \mathbb{E}\big(\log\frac{f(X;\theta)}{f(X;\theta_0)}\big).\end{aligned}$$

Now by using Jensen's inequality we have

$$\mathbb{E}\big(\log\frac{f(X;\theta)}{f(X;\theta_0)}\big) \leq \log\mathbb{E}\big(\frac{f(X;\theta)}{f(X;\theta_0)}\big) = \log\int f(x;\theta)dx = 0.$$

Thus giving $\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta)) - \mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0)) \leq 0$. To prove that $\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta)) - \mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0)) = 0$ only when θ_0 we note that identifiability assumption in Assumption 4.1.1(iii), which means that $f(x;\theta) = f(x;\theta_0)$ only when θ_0 and no other function of f gives equality.

Hence $\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta))$ is uniquely maximum at θ_0. Finally, we need to show that $\hat{\theta}_T \xrightarrow{\mathcal{P}} \theta_0$. By Assumption 4.1.1(ii) (and also the LLN) we have that for all $\theta \in \Theta$ that $\frac{1}{T}\mathcal{L}_T(\underline{X};\theta) \xrightarrow{\text{a.s.}} \ell(\theta)$. Therefore, for every mle $\hat{\theta}_T$ we have

$$\frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0) \leq \frac{1}{T}\mathcal{L}_T(\underline{X};\hat{\theta}_T) \xrightarrow{\text{a.s.}} \mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\hat{\theta}_T)) \leq \mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0)) \tag{4.4}$$

To bound $|\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0)) - \frac{1}{T}\mathcal{L}_T(\underline{X};\hat{\theta}_T)|$ we note that

$$\begin{aligned}\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0)) - \frac{1}{T}\mathcal{L}_T(\underline{X};\hat{\theta}_T) = \big\{\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0)) - \frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0)\big\} + \\ \big\{\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\hat{\theta}_T)) - \frac{1}{T}\mathcal{L}_T(\underline{X};\hat{\theta}_T)\big\} + \big\{\frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0) - \mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\hat{\theta}_T))\big\}.\end{aligned}$$

Now by using (4.4) we have

$$\begin{aligned}\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0)) - \frac{1}{T}\mathcal{L}_T(\underline{X};\hat{\theta}_T) \leq \big\{\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0)) - \frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0)\big\} + \\ \big\{\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\hat{\theta}_T)) - \frac{1}{T}\mathcal{L}_T(\underline{X};\hat{\theta}_T))\big\} + \big\{\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\hat{\theta}_T)) - \frac{1}{T}\mathcal{L}_T(\underline{X};\hat{\theta}_T))\big\}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0))) - \frac{1}{T}\mathcal{L}_T(\underline{X};\hat{\theta}_T) \geq \big\{\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0))) - \frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0)\big\} + \\ \big\{\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\hat{\theta}_T)) - \frac{1}{T}\mathcal{L}_T(\underline{X};\hat{\theta}_T)\big\} + \big\{\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0)) - \frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0)\big\}.\end{aligned}$$

Therefore, under Assumption 4.1.1(ii) we have

$$|\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta_0)) - \frac{1}{T}\mathcal{L}_T(\underline{X};\hat{\theta}_T)| \leq 3\sup_{\theta\in\Theta}|\mathbb{E}(\frac{1}{T}\mathcal{L}_T(\underline{X};\theta)) - \frac{1}{T}\mathcal{L}_T(\underline{X};\theta)| \xrightarrow{\text{a.s.}} 0.$$

Since $\mathcal{L}_T(\theta)$ has a unique minimum this implies $\hat{\theta}_T \xrightarrow{\text{a.s.}} \theta_0$. □

Hence we have shown consistency of the mle. We now need to show asymptotic normality.

Theorem 4.1.2 *Suppose Assumption 4.1.1 is satisfied.*

(i) Then the score statistic is

$$\frac{1}{\sqrt{T}}\frac{\partial\mathcal{L}_T(\underline{X};\theta)}{\partial\theta}\rfloor_{\theta_0} \xrightarrow{\mathcal{D}} \mathcal{N}\bigg(0, \bigg\{\mathbb{E}\bigg(\frac{\partial\log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\bigg)^2\bigg\}\bigg). \tag{4.5}$$

(ii) Then the mle is

$$\sqrt{T}\big(\hat{\theta}_T - \theta_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}\bigg(0, \bigg\{\mathbb{E}\bigg(\frac{\partial\log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\bigg)^2\bigg\}^{-1}\bigg).$$

(iii) The log likelihood ratio is

$$2\bigg(\mathcal{L}_T(\underline{X};\hat{\theta}_T)-\mathcal{L}_T(\underline{X};\theta_0)\bigg)\xrightarrow{\mathcal{D}}\chi^2_p$$

PROOF. First we will prove (i). We recall because $\{X_t\}$ are iid random variables, then

$$\frac{1}{\sqrt{T}}\frac{\partial\mathcal{L}_T(\underline{X};\theta)}{\partial\theta}\rfloor_{\theta_0}=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\partial\log f(X_t;\theta)}{\partial\theta}\rfloor_{\theta_0}.$$

Hence $\frac{\partial\mathcal{L}_T(\underline{X};\theta)}{\partial\theta}\rfloor_{\theta_0}$ is the sum of iid random variables with mean zero and variance $\mathrm{var}(\frac{\partial\log f(X_t;\theta)}{\partial\theta}\rfloor_{\theta_0})$. Therefore, by the CLT for iid random variables we have (4.5).

We use (i) and Taylor (mean value) theorem to prove (ii). We first note that by the mean value theorem we have

$$\frac{1}{T}\frac{\partial\mathcal{L}_T(\underline{X};\theta)}{\partial\theta}\rfloor_{\hat{\theta}_T}=\frac{1}{T}\frac{\partial\mathcal{L}_T(\underline{X};\theta)}{\partial\theta}\rfloor_{\theta_0}+(\hat{\theta}_T-\theta_0)\frac{1}{T}\frac{\partial^2\mathcal{L}_T(\underline{X};\theta)}{\partial\theta^2}\rfloor_{\bar{\theta}_T}. \tag{4.6}$$

Now it can be shown because Θ has a compact support, $|\hat{\theta}_T-\theta_0|\xrightarrow{\text{a.s.}}0$ and the expectations of the third derivative of $\mathcal{L}_T$ is bounded that

$$\frac{1}{T}\frac{\partial^2\mathcal{L}_T(\underline{X};\theta)}{\partial\theta^2}\rfloor_{\bar{\theta}_T}\xrightarrow{\mathcal{P}}\frac{1}{T}\mathbb{E}\bigg(\frac{\partial^2\mathcal{L}_T(\underline{X};\theta)}{\partial\theta^2}\rfloor_{\theta_0}\bigg)=\mathbb{E}\bigg(\frac{\partial^2\log f(X;\theta)}{\partial\theta^2}\rfloor_{\theta_0}\bigg). \tag{4.7}$$

Substituting (4.7) into (4.6) gives

$$\begin{aligned}\sqrt{T}(\hat{\theta}_T-\theta_0) &= \bigg(\frac{1}{T}\frac{\partial^2\mathcal{L}_T(\underline{X};\theta)}{\partial\theta^2}\rfloor_{\bar{\theta}_T}\bigg)^{-1}\frac{1}{\sqrt{T}}\frac{\partial\mathcal{L}_T(\underline{X};\theta)}{\partial\theta}\rfloor_{\theta_0}\\ &= \mathbb{E}\bigg(\frac{1}{T}\frac{\partial^2\mathcal{L}_T(\underline{X};\theta)}{\partial\theta^2}\rfloor_{\theta_0}\bigg)^{-1}\frac{1}{\sqrt{T}}\frac{\partial\mathcal{L}_T(\underline{X};\theta)}{\partial\theta}\rfloor_{\theta_0}+o_p(1).\end{aligned}$$

We mention that the proof above is for univariate $\frac{\partial^2\mathcal{L}_T(\underline{X};\theta)}{\partial\theta^2}\rfloor_{\bar{\theta}_T}$, but by redo-ing the above steps pointwise it can easily be generalised to the multivariate case too. Hence by substituting the (4.5) into the above we have (ii). It is straightfoward to prove (iii) by using

$$2\bigg(\mathcal{L}_T(\underline{X};\hat{\theta}_T)-\mathcal{L}_T(\underline{X};\theta_0)\bigg)\approx(\hat{\theta}_T-\theta_0)'I(\theta_0)(\hat{\theta}_T-\theta_0)',$$

(i) and the result that if $X\sim\mathcal{N}(0,\Sigma)$, then $AX\sim\mathcal{N}(0,A'\Sigma A)$. □

Example 4.1.7 (The Weibull) *By using Example 4.1.6 we have*

$$\begin{pmatrix}\hat{\theta}_T-\theta\\ \hat{\alpha}_T-\alpha\end{pmatrix}\approx I(\theta,\alpha)^{-1}\begin{pmatrix}\sum_{t=1}^{T}\bigg(-\frac{\alpha}{\theta}+\frac{\alpha}{\theta^{\alpha+1}}Y_t^{\alpha}\bigg)\\ \sum_{t=1}^{T}\bigg(\frac{1}{\alpha}-\log Y_t-\log\theta-\frac{\alpha}{\theta}+\log(\frac{Y_t}{\theta})\times(\frac{Y_t}{\theta})^{\alpha}\bigg)\end{pmatrix}.$$

Now we observe that RHS consists of a sum iid random variables (this can be viewed as an average). Since the variance of this exists (you can show that it is $I(\theta,\alpha)$), the CLT can be applied and we have that

$$\begin{pmatrix} \hat{\theta}_T - \theta \\ \hat{\alpha}_T - \alpha \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, I(\theta,\alpha)^{-1}\right).$$

Remark 4.1.2 *(i) We recall that for iid random variables that the Fisher information for sample size T is*

$$I(\theta) = \mathbb{E}\left\{ \frac{\partial \log L_T(X;\theta)}{\partial\theta} \rfloor_{\theta_0} \right\}^2 = T\mathbb{E}\left(\frac{\partial \log f(X;\theta)}{\partial\theta} \rfloor_{\theta_0} \right)^2.$$

Hence comparing with the above theorem, we see that for iid random variables (so long as the regularity conditions are satisfied) the MLE, asympotitically, attains the Cramer-Rao bound even if for finite samples this may not be true.

Moreover, since

$$(\hat{\theta}_T - \theta_0) \approx I(\theta_0)^{-1} \frac{\partial \mathcal{L}_T(\theta)}{\partial\theta} \rfloor_{\theta_0} = (T^{-1} I(\theta_0))^{-1} \frac{1}{T} \frac{\partial \mathcal{L}_T(\theta)}{\partial\theta} \rfloor_{\theta_0},$$

and $var\big(\frac{1}{\sqrt{T}} \frac{\partial \mathcal{L}_T(\theta)}{\partial\theta} \rfloor_{\theta_0}\big) = \frac{1}{T} I(\theta_0)$, then it can be seen that $|\hat{\theta}_T - \theta_0| = O_p(T^{-1/2})$.

(ii) Under suitable conditions a similar result holds true for data which is not iid.

In summary, the MLE (under certain regularity conditions) tend to have the smallest variance, and for large samples, the variance is close to the lower bound, which is the Cramer-Rao bound.

In the case that Assumption 4.1.1 is satisfied, the MLE is said to be asymptotically efficient. This means for finite samples the MLE may not attain the C-R bound but asymptotically it will.

(iii) A simple application of Theorem 4.1.2 is to the derivation of the distribution of $I(\theta_0)^{1/2}(\hat{\theta}_T - \theta_0)$. It is clear that by using Theorem 4.1.2 we have

$$I(\theta_0)^{1/2}(\hat{\theta}_T - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I_p)$$

(where I_p is the identity matrix) and

$$(\hat{\theta}_T - \theta_0)' I(\theta_0)(\hat{\theta}_T - \theta_0) \xrightarrow{\mathcal{D}} \chi^2_p.$$

(iv) Note that these results apply when θ_0 lies inside the parameter space Θ. As θ gets closer to the boundary of the parameter space

Remark 4.1.3 (Generalised estimating equations) *Closely related to the MLE are generalised estimating equations GEE, which are relate to the score statistic. These are estimators not based on maximising the likelihood but are related to equating the score statistic (derivative of the likelihood) to zero and solving for the unknown parameters. Often they are equivalent to the MLE but they can be adapted to be useful in themselves (and some adaptions will not be the derivative of a likelihood).*

4.1.5 The Fisher information

See also Section 4.3, Davison (2002).

Let us return to the Fisher information. We recall that undercertain regularity conditions an unbiased estimator, $\tilde{\theta}(\underline{X})$, of a parameter θ_0 is such that

$$\mathrm{var}(\tilde{\theta}(\underline{X})) \geq I(\theta_0)^{-1},$$

where

$$I(\theta) = \mathbb{E}\bigg(\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta}\bigg)^2 = \mathbb{E}\bigg(-\frac{\partial^2 \mathcal{L}_T(\theta)}{\partial \theta^2}\bigg).$$

is the Fisher information. Furthermore, under suitable regularity conditions, the MLE will asymptotically attain this bound. It is reasonable to ask, how one can interprete this bound.

(i) Situation 1. $I(\theta_0) = \mathbb{E}\bigg(-\frac{\partial^2 \mathcal{L}_T(\theta)}{\partial \theta^2}\rfloor_{\theta_0}\bigg)$ is large (hence variance of the mle will be small) then it means that the gradient of $\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta}$ is steep. Hence even for small deviations from θ_0, $\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta}$ is likely to be far from zero. This means the mle $\hat{\theta}_T$ is likely to be in a close neighbourhood of θ_0.

(ii) Situation 2. $I(\theta_0) = \mathbb{E}\bigg(-\frac{\partial^2 \mathcal{L}_T(\theta)}{\partial \theta^2}\rfloor_{\theta_0}\bigg)$ is small (hence variance of the mle will large). In this case the gradient of the likelihood $\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta}$ is flatter and hence $\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta} \approx 0$ for a large neighbourhood about the true parameter θ. Therefore the mle $\hat{\theta}_T$ can lie in a large neighbourhood of θ_0.

This is one explanation as to why $I(\theta)$ is called the Fisher information. It contains information on how close close any estimator of θ can be.

Look at the censoring example, Example 4.20, page 112, Davison (2002).

Chapter 5

Confidence Intervals

5.1 Confidence Intervals and testing

We first summarise the results in the previous section (which will be useful in this section). For convenience, we will assume that the likelihood is for iid random variables, whose density is $f(x;\theta_0)$ (though it is relatively simple to see how this can be generalised to general likelihoods - of not necessarily iid rvs). Let us suppose that θ_0 is the true parameter that we wish to estimate. Based on Theorem 4.1.2 we have

$$\sqrt{T}\big(\hat{\theta}_T - \theta_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}\bigg(0, \bigg\{\mathbb{E}\bigg(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\bigg)^2\bigg\}^{-1}\bigg), \tag{5.1}$$

$$\frac{1}{\sqrt{T}}\frac{\partial \mathcal{L}_T}{\partial\theta}\rfloor_{\theta=\theta_0} \xrightarrow{\mathcal{D}} \mathcal{N}\bigg(0, \bigg\{\mathbb{E}\bigg(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\bigg)^2\bigg\}\bigg) \tag{5.2}$$

and

$$2\big(\mathcal{L}_T(\hat{\theta}_T) - \mathcal{L}_T(\theta_0)\big) \xrightarrow{\mathcal{D}} \chi^2_p, \tag{5.3}$$

where p are the number of parameters in the vector θ. Using any of (5.1), (5.2) and (5.3) we can construct 95% CI for θ_0.

5.1.1 Constructing confidence intervals using the likelihood

See also Section 4.5, Davison (2002).

One the of main reasons that we show asymptotic normality of an estimator (it is usually not possible to derive normality for finite samples) is to construct confidence intervals (CIs) and to test.

In the case that θ_0 is a scaler (vector of dimension one), it is easy to use (5.1) to obtain

$$\sqrt{T}\Big\{\mathbb{E}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\Big)^2\Big\}^{1/2}\big(\hat{\theta}_T-\theta_0\big)\xrightarrow{\mathcal{D}}N(0,1). \tag{5.4}$$

Based on the above the 95% CI for θ_0 is

$$\Big[\hat{\theta}_T-\frac{1}{\sqrt{T}}\mathbb{E}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\Big)^2 z_{\alpha/2},\hat{\theta}_T+\frac{1}{\sqrt{T}}\mathbb{E}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\Big)^2 z_{\alpha/2}\Big].$$

The above, of course, requires an estimate of the (standardised) Fisher information $\mathbb{E}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\Big)^2=\mathbb{E}\Big(-\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\rfloor_{\theta_0}\Big)$ Usually, we evaluate the second derivative of $\frac{1}{T}\log L_T(\theta)=\frac{1}{T}\mathcal{L}_T(\theta)$ and replace θ with the estimator of θ, $\hat{\theta}_T$.

Exercise: Use (5.2) to construct a CI for θ_0 based on the score

The CI constructed above works well if θ is a scalar. But beyond dimension one, constructing a CI based on (5.1) (and the p-dimensional normal) is extremely difficult. More precisely, if θ_0 is a p-dimensional vector then the analogous version of (5.4) is

$$\sqrt{T}\Big\{\mathbb{E}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\Big)^2\Big\}^{1/2}\big(\hat{\theta}_T-\theta_0\big)\xrightarrow{\mathcal{D}}N(0,I_p),$$

using this it is difficult to obtain the CI of θ_0. One way to construct the CI is to 'square' $\big(\hat{\theta}_T-\theta_0\big)$ and use

$$\big(\hat{\theta}_T-\theta_0\big)'T\mathbb{E}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\Big)^2\big(\hat{\theta}_T-\theta_0\big)\xrightarrow{\mathcal{D}}\chi^2_p. \tag{5.5}$$

Based on above a 95% CI is

$$\Big\{\theta;\big(\hat{\theta}_T-\theta\big)'T\mathbb{E}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\Big)^2\big(\hat{\theta}_T-\theta\big)\leq\chi^2_p(0.95)\Big\}. \tag{5.6}$$

Note that as in the scalar case, this leads to the interval with the smallest length. A disadvantage of (5.6) is that we have to (a) estimate the information matrix and (b) try to find all θ such the above holds. This can be quite unwielding. An alternative method, which is asymptotically equivalent to the above but removes the need to estimate the information matrix and is to use (5.3). By using (5.3), a $100(1-\alpha)\%$ CI for θ_0 is

$$\Big\{\theta;2\big(\mathcal{L}_T(\hat{\theta}_T)-\mathcal{L}_T(\theta)\big)\leq\chi^2_p(100(1-\alpha))\Big\}. \tag{5.7}$$

The above is not easy to calculate, but it is feasible.

Example 5.1.1 *In the case that θ_0 is a scalar the 95% CI based on (5.7) is*

$$\left\{\theta; \mathcal{L}_T(\theta) \geq \mathcal{L}_T(\hat{\theta}_T) - \frac{1}{2}\chi^2_p(0.95)\right\}.$$

Both the 95% CIs in (5.6) and (5.7) will be very close for relatively large sample sizes. However one advantage of using (5.7) instead of (5.6) is that it is easier to evaluate - no need to obtain the second derivative of the likelihood etc.

Another feature which differentiates the CIs in (5.6) and (5.7) is that the CI based on (5.6) is symmetric about $\hat{\theta}_T$ (recall that $(\bar{X} - 1.96\sigma/\sqrt{T}, \bar{X} + 1.96\sigma/\sqrt{T})$) is symmetric about $\bar{X}$, whereas the symmetry condition may not hold for sample sizes when constructing a CI for θ_0 using (5.7). This is a positive advantage of using (5.7) instead of (5.6). A disadvantage of using (5.7) instead of (5.6) is that sometimes in the CI based on (5.7) may have more than one interval.

As you can see if the dimension of θ is large it is quite difficult to evaluate the CI (try it for the simple case that the dimension is two!). Indeed for dimensions greater than three it is extremely hard. However in most cases, we are only interested in constructing CIs for certain parameters of interest, the other unknown parameters are simply nuisance parameters and CIs for them are not of interest. For example, for the normal distribution we may only be interested in CIs for the mean but not the variance.

It is clear that directly using the log-likelihood ratio to construct CIs (and also test) will mean also constructing CIs for the nuisance parameters. Therefore below (in Section **??**) we construct a variant of the likelihood (called the Profile likelihood), which allows us to deal with nuisance parameters in a more efficient way.

5.1.2 Testing using the likelihood

Let us suppose we wish to test the hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_A : \theta \neq \theta_0$. We can use any of the results in (5.1), (5.2) and (5.3) to do the test - they will lead to slightly different p-values, but 'asympototically' they are all equivalent, because they are all based (essentially) on the same derivation.

We now list the three tests that one can use.

The Wald test

The Wald statistic is based on (5.1). We recall from (5.1) that if the null is true, then we have

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \left\{\mathbb{E}\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\rfloor_{\theta_0}\right)^2\right\}^{-1}\right).$$

Thus we can use as the test statistic

$$T_1 = \sqrt{T}\Big\{\mathbb{E}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\Big)^2\Big\}^{1/2}\big(\hat{\theta}_T - \theta_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

Let us now consider how the test statistics behaves under the alternative $H_A : \theta = \theta_1$. If the null is not true, then we have that

$$\begin{aligned}(\hat{\theta}_T - \theta_0) &= (\hat{\theta}_T - \theta_1) + (\theta_1 - \theta_0) \\ &\approx I(\theta_1)^{-1}\frac{1}{\sqrt{T}}\sum_t \frac{\partial \log f(X_t;\theta_1)}{\partial\theta_1}(\theta_1 - \theta_0)\end{aligned}$$

Thus the distribution of the test statistic T_1 becomes centered about $\sqrt{T}\Big\{\mathbb{E}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\Big)^2\Big\}^{1/2}\big(\theta - \theta_0\big)$. Thus for a larger sample size the more likely we are to reject the null.

Remark 5.1.1 (Types of alternatives) *In the case that the alternative is fixed, it is clear that the power in the test goes to* 100%. *Therefore, often to see the effectiveness of the test, one lets the alternative get closer to the the null as $T \to \infty$. For example*

- *Suppose that $\theta_1 = \theta_0 + \frac{1}{T}$, then the center of T_1 is $\frac{1}{\sqrt{T}}\Big\{\mathbb{E}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\Big)^2\Big\}^{1/2} \to 0$. Thus the alternative is too close to the null for us to discriminate between the the two.*
- *Suppose that $\theta_1 = \theta_0 + \frac{1}{\sqrt{T}}$, then the center of T_1 is $\Big\{\mathbb{E}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\Big)^2\Big\}^{1/2}$. Therefore, the test does have power, but it's not* 100%.

In the case that the dimension of θ is greater than one, we use the test statistic $\tilde{T}_1 = \big(\hat{\theta}_T - \theta_0\big)\sqrt{T}\mathbb{E}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\Big)^2\big(\hat{\theta}_T - \theta_0\big)$ instead of T_1. Noting that the distribution of T_1 is a chi-squared with p-degrees of freedom.

The Score test

The score test is based on the score. We recall from (**??**), that under the null the distribution of the score is

$$\frac{1}{\sqrt{T}}\frac{\partial \mathcal{L}_T}{\partial\theta}\rfloor_{\theta=\theta_0} \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \Big\{\mathbb{E}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\Big)^2\Big\}\Big).$$

Thus we use as the test statistic

$$T_2 = \frac{1}{\sqrt{T}}\Big\{\mathbb{E}\Big(\frac{\partial \log f(X;\theta)}{\partial\theta}\rfloor_{\theta_0}\Big)^2\Big\}^{-1/2}\frac{\partial \mathcal{L}_T}{\partial\theta}\rfloor_{\theta=\theta_0} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

An advantage of this test is that the maximum likelihood estimator (under either the null or alternative) does not have to be calculated.

Exercise: What does the test statistic look like under the alternative?

The log-likelihood ratio test

Probably one of the most popular test os the log-likelihood ratio tests. This test is based on (5.3), and the test statistic is

$$T_3 = 2\big(\mathcal{L}_T(\hat{\theta}_T) - \mathcal{L}_T(\theta_0)\big) \xrightarrow{\mathcal{D}} \chi^2_p.$$

An advantage of this test statistic is that it is pivotal, in the sense that the Fisher information etc. does not have to calculated, only the maximum likelihood estimator.

Exercise: What does the test statistic look like under the alternative?

5.1.3 Applications of the log-likeihood ratio to the multinomial distribution

Example 5.1.2 (The multinomial distribution) *This is a generalisation of the binomial distribution. In this case at any given trial there can arise m different events (in the Binomial case $m = 2$). Let Z_i denote the outcome of the ith trial and assume $P(Z_i = k) = \pi_i$ ($\pi_1 + \ldots + \pi_m = 1$). Suppose there were n trial conducted and let Y_1 denote the number of times event 1 arises, Y_2 denote the number of times event 2 arises and so on. Then it is straightforward to show that*

$$P(Y_1 = k_1, \ldots, Y_m = k_m) = \binom{n}{k_1, \ldots, k_m} \prod_{i=1}^{m} \pi_i^{k_i}.$$

If we do not impose any constraints on the probabilities $\{\pi_i\}$, given $\{Y_i\}_{i=1}^m$ is straightforward to derive the mle of $\{\pi_i\}$ (it is very intuitive too!).

Noting that $\pi_m = 1 - \sum_{i=1}^{m-1} \pi_i$, the log-likelihood of the multinomial is proportional to

$$\mathcal{L}_T(\underline{\pi}) \;=\; \sum_{i=1}^{m-1} y_i \log \pi_i + y_m \log(1 - \sum_{i=1}^{m-1} \pi_i).$$

Differentiating the above with respect to π_i and solving gives the mle estimator $\hat{\pi}_i = Y_i/n$, which is what we would have expected! We observe that though there are m probabilities to estimate due to the constraint $\pi_m = 1 - \sum_{i=1}^{m-1} \pi_i$, we only have to estimate $(m-1)$ probabilities. We mention, that the same estimators can also be obtained by using Lagrange multipliers, that is maximising $\mathcal{L}_T(\underline{\pi})$ subject to the parameter constraint that $\sum_{j=1}^{p} \pi_i = 1$. To enforce this constraint, we normally add an additional term to $\mathcal{L}_T(\underline{\pi})$ and include the dummy variable λ. That is we define the constrained likelihood

$$\tilde{\mathcal{L}}_T(\underline{\pi}, \lambda) \;=\; \sum_{i=1}^{m} y_i \log \pi_i + \lambda(\sum_{i=1}^{m} \pi_i - 1).$$

Now if we maximise $\tilde{\mathcal{L}}_T(\underline{\pi}, \lambda)$ with respect to $\{\pi_i\}_{i=1}^m$ and λ we will obtain the estimators $\hat{\pi}_i = Y_i/n$ (which is the same as the maximum of $\mathcal{L}_T(\underline{\pi})$).

To derive the limiting distribution we note that the second derivative is

$$-\frac{\partial^2 \mathcal{L}_T(\underline{\pi})}{\partial \pi_i \pi_j} = \begin{cases} \frac{y_i}{\pi_i^2} + \frac{y_m}{(1-\sum_{r=1}^{m-1}\pi_r)^2} & i = j \\ \frac{y_m}{(1-\sum_{r=1}^{m-1}\pi_r)^2} & i \neq j \end{cases}$$

Hence taking expectations of the above the information matrix is the $(k-1) \times (k-1)$ matrix

$$I(\pi) = n \begin{pmatrix} \frac{1}{\pi_1} + \frac{1}{\pi_m} & \frac{1}{\pi_m} & \cdots & \frac{1}{\pi_m} \\ \frac{1}{\pi_m} & \frac{1}{\pi_2} + \frac{1}{\pi_m} & \cdots & \frac{1}{\pi_m} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1}{\pi_{m-1}} & \cdots & \frac{1}{\pi_{m-1}} + \frac{1}{\pi_m} & \end{pmatrix}.$$

Provided no of π_i is equal to either 0 or 1 (which would drop the dimension of m and make $I(\pi)$) singular, then the asymptotic distribution of the mle the normal with variance $I(\pi)^{-1}$.

Sometimes the probabilities $\{\pi_i\}$ will not be 'free' and will be determined by a parameter θ (where θ is an r-dimensional vector where $r < m$), ie. $\pi_i = \pi_i(\theta)$, in this case the likelihood of the multinomial is

$$\mathcal{L}_T(\underline{\pi}) = \sum_{i=1}^{m-1} y_i \log \pi_i + y_m \log(1 - \sum_{i=1}^{m-1} \pi_i(\theta)).$$

By differentiating the above with respect to θ and solving will give the mle.

Pearson's goodness of Fit test

We now derive Pearson's goodness of Fit test using the log-likelihood ratio, though Pearson did not use this method to derive his test.

Suppose the null is $H_0 : \pi_1 = \tilde{\pi}_1, \ldots, \pi_m = \tilde{\pi}_m$ (where $\{\tilde{\pi}_i\}$ are some pre-set probabilities) and H_A : the probabilities are not the given probabilities. Hence we are testing restricted model (where we do not have to estimate anything) against the full model where we estimate the probabilities using $\pi_i = Y_i/n$.

The log-likelihood ratio in this case is

$$W = 2\big\{ \arg\max_{\pi} \mathcal{L}_T(\pi) - \mathcal{L}_T(\tilde{\pi}) \big\}.$$

Under the null we know that $W = 2\big\{ \arg\max_{\pi} \mathcal{L}_T(\pi) - \mathcal{L}_T(\tilde{\pi}) \big\} \xrightarrow{\mathcal{P}} \chi^2_{m-1}$ (because we have to estimate $(m-1)$ parameters). We now derive an expression for W and show that the Pearson-

statistic is an approximation of this.

$$\begin{aligned}\frac{1}{2}W &= \sum_{i=1}^{m-1} Y_i \log\big(\frac{Y_i}{n}\big) + Y_m \log\frac{Y_m}{n} - \sum_{i=1}^{m-1} Y_i \log\tilde{\pi}_i - Y_m \log\tilde{\pi}_m \\ &= \sum_{i=1}^{m} Y_i \log\big(\frac{Y_i}{n\tilde{\pi}_i}\big).\end{aligned}$$

Recall that Y_i is often called the observed $Y_i = O_i$ and $n\tilde{\pi}_i$ the expected under the null $E_i = n\tilde{\pi}_i$. Then $W = 2\sum_{i=1}^{m} O_i \log\big(\frac{O_i}{E_i}\big) \xrightarrow{\mathcal{P}} \chi^2_{m-1}$. By using that for a close to x and making a Taylor expansion of $x\log(xa^{-1})$ about $x = a$ we have $x\log(xa^{-1}) \approx a\log(aa^{-1}) + (x-a) + \frac{1}{2}(x-a)^2/a$. We let $O = x$ and $E = a$, then assuming the null is true and $E_i \approx O_i$ we have

$$W = 2\sum_{i=1}^{m} Y_i \log\big(\frac{Y_i}{n\tilde{\pi}_i}\big) \approx 2\sum_{i=1}^{m} \big((O_i - E_i) + \frac{1}{2}\frac{(O_i - E_i)^2}{E_i}\big).$$

Now we note that $\sum_{i=1}^{m} E_i = \sum_{i=1}^{m} O_i = n$ hence the above reduces to

$$W \approx \frac{(O_i - E_i)^2}{E_i}\big) \xrightarrow{\mathcal{D}} \chi^2_{m-1}.$$

We recall that the above is the Pearson test statistic. Hence this is one methods for deriving the Pearson chi-squared test for goodness of fit.

By using a similar argument, we can also obtain the test statistic of the chi-squared test for independent (and an explanation for the rather strange number of degrees of freedom!).

Chapter 6

The Profile Likelihood

6.1 The Profile Likelihood

See also Section 4.5.2, Davison (2002).

6.1.1 The method of profiling

Let us suppose that the unknown parameters θ can be partitioned as $\theta' = (\psi', \lambda')$, where ψ are the p-dimensional parameters of interest (eg. mean) and λ are the q-dimensional nuisance parameters (eg. variance). We will need to estimate both ψ and λ, but our interest is in testing only the parameter ψ (without any information on λ) and construction confidence intervals for ψ (without constructing unnecessary confidence intervals for λ - confidence intervals for a large number of parameters are wider than those for a few parameters). To achieve this one often uses the profile likelihood. To motivate the profile likelihood, we first describe a method to estimate the parameters (ψ, ψ) in two stages and consider some examples.

Let us suppse that $\{X_t\}$ are iid random variables, with density $f(x; \psi, \lambda)$ where our objective is to estimate ψ and λ. In this case the log-likelihood is

$$\mathcal{L}_T(\psi, \lambda) = \sum_{t=1}^{T} \log f(X_t; \psi, \lambda).$$

To estimate ψ and λ one can use $(\hat{\lambda}_T, \hat{\psi}_T) = \arg\max_{\lambda,\psi} \mathcal{L}_T(\psi, \lambda)$. However, this can quite difficult, and lead to expression which are hard to maximise. Instead let us consider a different method, which may, sometimes, be easier to evaluate. Suppose, for now, ψ is known, then we rewrite the likelihood as $\mathcal{L}_T(\psi, \lambda) = \mathcal{L}_\psi(\lambda)$ (to show that ψ is fixed but λ varies). To estimate λ we maximise $\mathcal{L}_\psi(\lambda)$ with respect to λ, ie.

$$\hat{\lambda}_\psi = \arg\max_{\lambda} \mathcal{L}_\psi(\lambda).$$

In reality ψ this unknown, hence for each ψ we can evaluate $\hat{\lambda}_\psi$. Note that for each ψ, we have a new curve $\mathcal{L}_\psi(\lambda)$ over λ. Now to estimate ψ, we evaluate the maximum $\mathcal{L}_\psi(\lambda)$, over λ, and choose the ψ, which is the maximum over all these curves. In other words, we evaluate

$$\hat{\psi}_T = \arg\max_\psi \mathcal{L}_\psi(\hat{\lambda}_\psi) = \arg\max_\psi \mathcal{L}_T(\psi, \hat{\lambda}_\psi).$$

A bit of logical deduction shows that $\hat{\psi}_T$ and $\lambda_{\hat{\psi}_T}$ are the maximum likelihood estimators $(\hat{\lambda}_T, \hat{\psi}_T) = \arg\max_{\psi,\lambda} \mathcal{L}_T(\psi, \lambda)$.

We note that we have *profiled* out nuisance parameter λ, and the likelihood $\mathcal{L}_\psi(\hat{\lambda}_\psi) = \mathcal{L}_T(\psi, \hat{\lambda}_\psi)$ is completely in terms of the parameter of interest ψ.

The advantage of this best illustrated through some examples.

Example 6.1.1 *Let us suppose that $\{X_t\}$ are iid random variables from a Weibull distribution with density $f(x;\alpha,\theta) = \frac{\alpha y^{\alpha-1}}{\theta^\alpha}\exp(-(y/\theta)^\alpha)$. We know from Example 4.1.2, that if α, were known an explicit expression for the MLE can be derived, it is*

$$\begin{aligned}\hat{\theta}_\alpha &= \arg\max_\theta \mathcal{L}_\alpha(\theta) \\ &= \arg\max_\theta \sum_{t=1}^T \Big(\log\alpha + (\alpha-1)\log Y_t - \alpha\log\theta - \big(\frac{Y_t}{\theta}\big)^\alpha \Big) \\ &= \arg\max_\theta \sum_{t=1}^T \Big(-\alpha\log\theta - \big(\frac{Y_t}{\theta}\big)^\alpha \Big) = \big(\frac{1}{T}\sum_{t=1}^T Y_t^\alpha\big)^{1/\alpha},\end{aligned}$$

where $\mathcal{L}_\alpha(\underline{X};\theta) = \sum_{t=1}^T \Big(\log\alpha + (\alpha-1)\log Y_t - \alpha\log\theta - \big(\frac{Y_t}{\theta}\big)^\alpha \Big)$. Thus for a given α, the maximum likelihood estimator of θ can be derived. The maximum likelihood estimator of α is

$$\hat{\alpha}_T = \arg\max_\alpha \sum_{t=1}^T \Big(\log\alpha + (\alpha-1)\log Y_t - \alpha\log\big(\frac{1}{T}\sum_{t=1}^T Y_t^\alpha\big)^{1/\alpha} - \big(\frac{Y_t}{(\frac{1}{T}\sum_{t=1}^T Y_t^\alpha)^{1/\alpha}}\big)^\alpha \Big).$$

Therefore, the maximum likelihood estimator of θ is $(\frac{1}{T}\sum_{t=1}^T Y_t^{\hat{\alpha}_T})^{1/\hat{\alpha}_T}$. We observe that evaluating $\hat{\alpha}_T$ can be tricky but no worse than maximising the likelihood $\mathcal{L}_T(\alpha,\theta)$ over α and θ.

As we mentioned above, we often do not have any interest in the nuisance parameters λ and are only interesting in testing and constructing CIs for α. In this case, we are interested in the limiting distribution of the MLE $\hat{\alpha}_T$. This can easily be derived by observing that

$$\sqrt{T}\begin{pmatrix} \hat{\psi}_T - \psi \\ \hat{\lambda}_T - \lambda \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \begin{pmatrix} I_{\psi\psi} & I_{\psi\lambda} \\ I_{\lambda\psi} & I_{\lambda\lambda} \end{pmatrix}^{-1}\Big).$$

where

$$\begin{pmatrix} I_{\psi\psi} & I_{\psi\lambda} \\ I_{\lambda\psi} & I_{\lambda\lambda} \end{pmatrix} = \begin{pmatrix} \mathbb{E}\big(-\frac{\partial^2 \log f(X_t;\psi,\lambda)}{\partial\psi^2}\big) & \mathbb{E}\big(-\frac{\partial^2 \log f(X_t;\psi,\lambda)}{\partial\psi\partial\lambda}\big) \\ \mathbb{E}\big(-\frac{\partial^2 \log f(X_t;\psi,\lambda)}{\partial\psi\partial\lambda}\big)' & \mathbb{E}\big(-\frac{\partial^2 \log f(X_t;\psi,\lambda)}{\partial\psi^2}\big) \end{pmatrix}. \tag{6.1}$$

To derive an exact expression for the limiting variance of $\sqrt{T}(\hat{\psi}_T - \psi)$, we note that the inverse of a block matrix is

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -D^{-1}CB(A - BD^{-1}C)^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}.$$

Thus the above implies that

$$\sqrt{T}(\hat{\psi}_T - \psi) \xrightarrow{\mathcal{D}} \mathcal{N}(0, (I_{\psi,\psi} - I_{\psi,\lambda}I_{\lambda\lambda}^{-1}I_{\lambda,\psi})^{-1}).$$

Thus if ψ is a scalar we can easily use the above to construct confidence intervals for ψ.

Exercise: How to estimate $I_{\psi,\psi} - I_{\psi,\lambda}I_{\lambda\lambda}^{-1}I_{\lambda,\psi}$?

6.1.2 The score and the log-likelihood ratio for the profile likelihood

To ease notation, let us suppose that ψ_0 and λ_0 are the true parameters in the distribution. The above gives us the limiting distribution of $(\hat{\psi}_T - \psi_0)$, this allows us to test ψ, however the test ignores any dependency that may exist with the nusiance estimator parameter $\hat{\lambda}_T$. An alternative test, which circumvents this issue is to do a log-likelihood ratio test of the type

$$2\Big\{ \max_{\psi,\lambda} \mathcal{L}_T(\psi,\lambda) - \max_{\lambda} \mathcal{L}_T(\psi_0,\lambda) \Big\}. \tag{6.2}$$

However, to derive the limiting distribution in this case for this statistic is a little more complicated than the log-likelihood ratio test that does not involve nusiance parameters. This is because a direct Taylor expansion does not work. However we observe that

$$2\Big\{ \max_{\psi,\lambda} \mathcal{L}_T(\psi,\lambda) - \max_{\lambda} \mathcal{L}_T(\psi_0,\lambda) \Big\} = 2\Big\{ \max_{\psi,\lambda} \mathcal{L}_T(\psi,\lambda) - \mathcal{L}_T(\psi_0,\lambda_0) \Big\} - 2\Big\{ \max_{\lambda} \mathcal{L}_T(\psi_0,\lambda) - \max_{\lambda} \mathcal{L}_T(\psi_0,\lambda_0) \Big\},$$

now we will show below that by using a few Taylor expansions we can derive the limiting distribution of (6.2).

In the theorem below we will derive the distribution of the score and the nested- loglikelihood. Please note you do not have to learn this proof.

Theorem 6.1.1 *Suppose Assumption 4.1.1 holds. Suppose that (ψ_0, λ_0) are the true parameters. Then we have*

$$\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial\psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0} \approx \frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial\psi}\rfloor_{\lambda_0,\psi_0} - I_{\psi_0\lambda_0}I_{\lambda_0\lambda_0}^{-1}\frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial\lambda}\rfloor_{\psi_0,\lambda_0} \tag{6.3}$$

$$\frac{1}{\sqrt{T}}\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0} \xrightarrow{\mathcal{D}} \mathcal{N}(0,(I_{\psi_0\psi_0}-I_{\psi_0\lambda_0}I^{-1}_{\lambda_0\lambda_0}I_{\lambda,\psi})) \tag{6.4}$$

and

$$2\Big\{\mathcal{L}_T(\hat{\psi}_T,\hat{\lambda}_T)-\mathcal{L}_T(\psi_0,\hat{\lambda}_{\psi_0})\Big\}\xrightarrow{\mathcal{D}}\chi^2_p \tag{6.5}$$

where I is defined as in (6.1).

PROOF. We first prove (6.3) which is the basis of the proofs of (6.4) and (6.5) - in the remark below we try to interprete (6.3). To avoid, notationaly difficulties by considering the elements of the vector $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$ and $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \lambda}\rfloor_{\lambda=\lambda_0,\psi_0}$ (as discussed in Section 4.1.3) we will suppose that these are univariate random variables.

Our objective is to find an expression for $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$ in terms of $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \lambda}\rfloor_{\lambda=\lambda_0,\psi_0}$ and $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\lambda=\lambda_0,\psi_0}$ which will allow us to obtain its variance and asymptotic distribution easily.

Now making a Taylor expansion of $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$ about $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\lambda_0,\psi_0}$ gives

$$\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}\approx\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\lambda_0,\psi_0}+\frac{\partial^2 \mathcal{L}_T(\psi,\lambda)}{\partial \lambda\partial \psi}\rfloor_{\lambda_0,\psi_0}(\hat{\lambda}_{\psi_0}-\lambda_0).$$

Notice that we have used $\approx$ instead of $=$ because we replace the second derivative with its true parameters. Now if the sample size is large enough then we can say that $\frac{\partial^2 \mathcal{L}_T(\psi,\lambda)}{\partial \lambda\partial \psi}\rfloor_{\lambda_0,\psi_0}\approx \mathbb{E}\big(\frac{\partial^2 \mathcal{L}_T(\psi,\lambda)}{\partial \lambda\partial \psi}\rfloor_{\lambda_0,\psi_0}\big)$. To see why this is true consider the case that of iid random variables then

$$\begin{aligned}\frac{1}{T}\frac{\partial^2 \mathcal{L}_T(\psi,\lambda)}{\partial \lambda\partial \psi}\rfloor_{\lambda_0,\psi_0} &= \frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 \log f(X_t;\psi,\lambda)}{\partial \lambda\partial \psi}\rfloor_{\lambda_0,\psi_0}\\ &\approx \mathbb{E}\bigg(\frac{\partial^2 \log f(X_t;\psi,\lambda)}{\partial \lambda\partial \psi}\rfloor_{\lambda_0,\psi_0}\bigg).\end{aligned}$$

Therefore we have that

$$\frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0}}\approx\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\lambda_0,\psi_0}+T\cdot I_{\lambda\psi}(\hat{\lambda}_{\psi_0}-\lambda_0) \tag{6.6}$$

Hence we have the first part of the decomposition of $\frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0}}$ into the distribution which is known, now we need find a decomposition of $(\hat{\lambda}_{\psi_0}-\lambda_0)$ into known distributions. We first recall that since $\mathcal{L}_T(\psi_0,\hat{\lambda}_{\psi_0})=\arg\max_\lambda \mathcal{L}_T(\psi_0,\lambda)$ then

$$\frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\hat{\lambda}_{\psi_0}}=0$$

(as long as the parameter space is large enough and the maximum is not on the boundary). Therefore making a Taylor expansion of $\frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\hat{\lambda}_{\psi_0}}$ about $\frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\lambda_0,\psi=\psi_0}$ gives

$$\frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\hat{\lambda}_{\psi_0}}\approx\frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\lambda_0,\psi_0}+\frac{\partial^2 \mathcal{L}_T(\psi_0,\lambda)}{\partial \lambda^2}\rfloor_{\lambda_0,\psi_0}(\hat{\lambda}_{\psi_0}-\lambda_0).$$

Again using the same trick as in (6.6) we have

$$\frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\hat{\lambda}_{\psi_0}} \approx \frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\lambda_0,\psi_0} + T\cdot I_{\lambda\lambda}(\hat{\lambda}_{\psi_0}-\lambda_0) = 0.$$

Therefore

$$(\hat{\lambda}_{\psi_0}-\lambda_0) = -\frac{I_{\lambda\lambda}^{-1}}{T}\frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\lambda_0,\psi_0}. \tag{6.7}$$

Therefore substituting (6.6) into (6.7) gives

$$\frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0}} \approx \frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\lambda_0,\psi_0} - I_{\psi\lambda}I_{\lambda\lambda}^{-1}\frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\psi_0,\lambda_0}$$

and (6.3).

To prove (6.4) (ie. obtain the asympototic distribution and limiting variance of $\frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0}}$), we recall that the regular score function satisfies

$$\frac{1}{\sqrt{T}}\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\lambda_0,\psi_0} = \frac{1}{\sqrt{T}}\begin{pmatrix} \frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\lambda_0,\psi_0} \\ \frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\psi_0,\lambda_0} \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\theta_0)).$$

Now by substituting the above into (6.4) we immediately obtain (6.4).

Finally to prove (6.5) we the following decomposition, Taylor expansions and the trick in (6.6) to obtain

$$\begin{aligned} 2\Big\{\mathcal{L}_T(\hat{\psi}_T,\hat{\lambda}_T) - \mathcal{L}_T(\psi_0,\hat{\lambda}_{\psi_0})\Big\} &= 2\Big\{\mathcal{L}_T(\hat{\psi}_T,\hat{\lambda}_T) - \mathcal{L}_T(\psi_0,\lambda_0)\Big\} - 2\Big\{\mathcal{L}_T(\psi_0,\hat{\lambda}_{\psi_0}) - \mathcal{L}_T(\psi_0,\lambda_0)\Big\} \\ &\approx (\hat{\theta}_T-\theta_0)'I(\theta)(\hat{\theta}_T-\theta_0) - (\hat{\lambda}_{\psi_0}-\lambda_0)'I_{\lambda\lambda}(\hat{\lambda}_{\psi_0}-\lambda_0), \end{aligned} \tag{6.8}$$

where $\hat{\theta}_T' = (\hat{\psi},\hat{\lambda})$ (the mle). Now we want to rewrite $(\hat{\lambda}_{\psi_0}-\lambda_0)'$ in terms of $(\hat{\theta}_T-\theta_0)$. We start by recalling that from (6.6) we have

$$(\hat{\lambda}_{\psi_0}-\lambda_0) = -\frac{I_{\lambda\lambda}^{-1}}{T}\frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\lambda_0,\psi_0}.$$

Now we will rewrite $\frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\lambda_0,\psi_0}$ in terms of $(\hat{\theta}_T-\theta_0)$ by using

$$\begin{aligned} &\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta}\rfloor_{\hat{\theta}_T} \approx \frac{\partial \mathcal{L}_T(\theta)}{\partial \theta}\rfloor_{\theta_0} + T\cdot I(\theta)(\hat{\theta}_T-\theta_0) \\ \Rightarrow\quad &\frac{\partial \mathcal{L}_T(\theta)}{\partial \theta}\rfloor_{\theta_0} \approx -I(\theta)(\hat{\theta}_T-\theta_0). \end{aligned}$$

Therefore concentrating on the subvector $\frac{\partial \mathcal{L}_T(\theta)}{\partial \lambda}\rfloor_{\psi_0,\lambda_0}$ we see that

$$\frac{\partial \mathcal{L}_T(\theta)}{\partial \lambda}\rfloor_{\psi_0,\lambda_0} \approx I_{\lambda\psi}(\hat{\psi}-\psi_0) + I_{\lambda\lambda}(\hat{\lambda}-\lambda_0). \tag{6.9}$$

Substituting (6.9) into (6.7) gives

$$(\hat{\lambda}_{\psi_0} - \lambda_0) \approx -I_{\lambda\lambda}^{-1} I_{\lambda\psi}(\hat{\psi} - \psi_0) + (\hat{\lambda} - \lambda_0).$$

Finally substituting the above into (6.8) and making lots of cancellations we have

$$2\Big\{\mathcal{L}_T(\hat{\psi}_T, \hat{\lambda}_T) - \mathcal{L}_T(\psi_0, \hat{\lambda}_{\psi_0})\Big\} \approx T(\hat{\psi} - \psi_0)'(I_{\psi\psi} - I_{\psi\lambda} I_{\lambda,\lambda}^{-1} I_{\lambda,\psi})(\hat{\psi} - \psi_0).$$

Finally, since

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\theta)^{-1}),$$

by using inversion formulas for block matrices we have that $\sqrt{T}(\hat{\psi} - \psi_0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, (I_{\psi\psi} - I_{\psi\lambda} I_{\lambda,\lambda}^{-1} I_{\lambda,\psi})^{-1})$, which gives the desired result. □

Remark 6.1.1 *(i) We first make the rather interesting observation. The limiting variance of* $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\psi_0,\lambda_0}$ *is* $I_{\psi\psi}$*, whereas the the limiting variance of* $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$ *is* $(I_{\psi\psi} - I_{\psi\lambda} I_{\lambda,\lambda}^{-1} I_{\lambda,\psi})$ *and the limiting variance of* $\sqrt{T}(\hat{\psi} - \psi_0)$ *is* $(I_{\psi\psi} - I_{\psi\lambda} I_{\lambda,\lambda}^{-1} I_{\lambda,\psi})^{-1}$*.*

(ii) Look again at the expression

$$\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0} \approx \frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\lambda_0,\psi_0} - I_{\psi\lambda} I_{\lambda\lambda}^{-1} \frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \lambda}\rfloor_{\lambda_0,\psi_0} \tag{6.10}$$

It is useful to understand where it came from. Consider the problem of linear regression. Suppose X *and* Y *are random variables and we want to construct the best linear predictor of* Y *given* X*. We know that the best linear predictor is* $\hat{Y}(X) = \mathbb{E}(XY)/\mathbb{E}(Y^2)X$ *and the residual and mean squared error is*

$$Y - \hat{Y}(X) = Y - \frac{\mathbb{E}(XY)}{\mathbb{E}(Y^2)}X \text{ and } \mathbb{E}\left(Y - \frac{\mathbb{E}(XY)}{\mathbb{E}(Y^2)}X\right)^2 = \mathbb{E}(Y^2) - \mathbb{E}(XY)\mathbb{E}(Y^2)^{-1}\mathbb{E}(XY).$$

Compare this expression with (6.10). We see that in some sense $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$ *can be treated as the residual (error) of the projection of* $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\lambda_0,\psi_0}$ *onto* $\frac{\partial \mathcal{L}_T(\psi_0,\lambda)}{\partial \lambda}\rfloor_{lambda_0,\psi_0}$*. This is quite surprising!*

We now aim to use the above result. It is immediately clear that (6.5) can be used for both constructing likelihoods and testing. For example, to construct a 95% CI for ψ we can use the mle $\hat{\theta}_T = (\hat{\psi}_T, \hat{\lambda}_T)$ and the profile likelihood and use the 95% CI

$$\left\{\psi; 2\Big\{\mathcal{L}_T(\hat{\psi}_T, \hat{\lambda}_T) - \mathcal{L}_T(\psi, \hat{\lambda}_\psi)\Big\} \leq \chi_p^2(0.95)\right\}.$$

As you can see by profiling out the parameter λ, we have avoided the need to also construct a CI for λ too. This has many advantages, from a practical perspective it reduced the dimension of the parameters.

The log-likelihood ratio test in the presence of nuisance parameters

An application of Theorem 6.1.1 is for nested hypothesis testing, as stated at the beginning of this section. (6.5) can be used to test $H_0 : \psi = \psi_0$ against $H_A : \psi \neq \psi_0$ since

$$2\Big\{ \max_{\psi,\lambda} \mathcal{L}_T(\psi, \lambda) - \max_{\lambda} \mathcal{L}_T(\psi_0, \lambda) \Big\} \xrightarrow{\mathcal{D}} \chi^2_p.$$

Example 6.1.2 (χ^2-test for independence) *Now it is worth noting that using the Profile likelihood one can derive the chi-squared test for independence (in much the same way that the Pearson goodness of fit test was derived using the log-likelihood ratio test).*

Do this as an exercise (see Davison, Example 4.37, page 135).

The score test in the presence of nuisance parameters

We recall that we used Theorem 6.1.1 to obtain the distribution of $2\big\{ \max_{\psi,\lambda} \mathcal{L}_T(\psi, \lambda) - \max_{\lambda} \mathcal{L}_T(\psi_0, \lambda)\big\}$ under the null, we now motivate an alternative test to test the same hypothesis (which uses the same Theorem). We recall that under the null $H_0 : \psi = \psi_0$ the derivative $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \lambda}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0} = 0$, but the same is not true of $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$. However, if the null is true we would expect if $\hat{\lambda}_{\psi_0}$ to be close to the true λ_0 and for $\frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0},\psi_0}$ to be close to zero. Indeed this is what we showed in (6.4), where we showed that under the null

$$\frac{\partial \frac{1}{\sqrt{T}} \mathcal{L}_T(\psi, \lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, I_{\psi\psi} - I_{\psi\lambda} I^{-1}_{\lambda,\lambda} I_{\lambda,\psi}), \tag{6.11}$$

where $\lambda_{\psi_0} = \arg\max_{\lambda} \mathcal{L}_T(\psi_0, \lambda)$.

Therefore (6.11) suggests an alternative test for $H_0 : \psi = \psi_0$ against $H_A : \psi \neq \psi_0$. We can use $\frac{1}{\sqrt{T}} \frac{\partial \mathcal{L}_T(\psi,\lambda)}{\partial \psi}\rfloor_{\hat{\lambda}_{\psi_0}}$ as the test statistic. This is called the score or LM test.

The log-likelihood ratio test and the score test are asymptotically equivalent. There are advantages and disadvantages of both.

(i) An advantage of the log-likelihood ratio test is that we do not need to calculate the information matrix.

(ii) An advantage of the score test is that we do not have to evaluate the the maximum likelihood estimates under the alternative model.

6.1.3 Examples

Example: An application of profiling to frequency estimation

Question

Suppose that the observations $\{X_t; t = 1, \ldots, T\}$ satisfy the following nonlinear regression model

$$X_t = A\cos(\omega t) + B\sin(\omega t) + \varepsilon_t$$

where $\{\varepsilon_t\}$ are iid standard normal random variables and $0 < \omega < \pi$. The parameters A, B, and ω are real and unknown.

Some useful identities are given at the end of the question.

(i) Ignoring constants, obtain the log-likelihood of $\{X_t\}$. Denote this likelihood as $\mathcal{L}_T(A, B, \omega)$.

(ii) Let

$$\mathcal{S}_T(A, B, \omega) = \Big(\sum_{t=1}^{T} X_t^2 - 2\sum_{t=1}^{T} X_t\big(A\cos(\omega t) + B\sin(\omega t)\big) + \frac{1}{2}T(A^2 + B^2) \Big).$$

Show that

$$2\mathcal{L}_T(A, B, \omega) + \mathcal{S}_T(A, B, \omega) = \frac{(A^2 - B^2)}{2}\sum_{t=1}^{T}\cos(2\omega) + AB\sum_{t=1}^{T}\sin(2\omega).$$

Thus show that $|\mathcal{L}_T(A, B, \omega) + \frac{1}{2}\mathcal{S}_T(A, B, \omega)| = O(1)$ (ie. the difference does not grow with T).

Since $\mathcal{L}_T(A, B, \omega)$ and $-\frac{1}{2}\mathcal{S}_T(A, B, \omega)$ are asymptotically equivalent, for the rest of this question, *use* $\frac{-1}{2}\mathcal{S}_T(A, B, \omega)$ instead of the likelihood $\mathcal{L}_T(A, B, \omega)$.

(iii) Obtain the profile likelihood of ω.

(hint: Profile out the parameters A and B, to show that $\hat{\omega}_T = \arg\max_\omega |\sum_{t=1}^{T} X_t \exp(it\omega)|^2$).

Suggest, a graphical method for evaluating $\hat{\omega}_T$?

(iv) By using the identity

$$\sum_{t=1}^{T}\exp(i\Omega t) = \begin{cases} \frac{\exp(\frac{1}{2}i(T+1)\Omega)\sin(\frac{1}{2}T\Omega)}{\sin(\frac{1}{2}\Omega)} & 0 < \Omega < 2\pi \\ T & \Omega = 0 \text{ or } 2\pi. \end{cases} \tag{6.12}$$

show that for $0 < \Omega < 2\pi$ we have

$$\sum_{t=1}^{T} t\cos(\Omega t) = O(T) \quad \sum_{t=1}^{T} t\sin(\Omega t) = O(T)$$
$$\sum_{t=1}^{T} t^2\cos(\Omega t) = O(T^2) \quad \sum_{t=1}^{T} t^2\sin(\Omega t) = O(T^2).$$

(v) By using the results in part (iv) show that the Fisher Information of $\mathcal{L}_T(A, B, \omega)$ (denoted as $I(A, B, \omega)$) is asymptotically equivalent to

$$2I(A, B, \omega) = E\big(\frac{\partial^2 \mathcal{S}_T}{\partial \omega^2}\big) = \begin{pmatrix} \frac{T}{2} & 0 & \frac{T^2}{2}B + O(T) \\ 0 & \frac{T}{2} & -\frac{T^2}{2}A + O(T) \\ \frac{T^2}{2}B + O(T) & -\frac{T^2}{2}A + O(T) & \frac{T^3}{3}(A^2 + B^2) + O(T^2) \end{pmatrix}.$$

(vi) Derive the asymptotic variance of maximum likelihood estimator, $\hat{\omega}_T$, derived in part (iv).

Comment on the rate of convergence of $\hat{\omega}_T$.

Useful information: In this question the following quantities may be useful:

$$\sum_{t=1}^{T} \exp(i\Omega t) = \begin{cases} \frac{\exp(\frac{1}{2}i(T+1)\Omega)\sin(\frac{1}{2}T\Omega)}{\sin(\frac{1}{2}\Omega)} & 0 < \Omega < 2\pi \\ T & \Omega = 0 \text{ or } 2\pi. \end{cases} \tag{6.13}$$

the trignometric identities: $\sin(2\Omega) = 2\sin\Omega\cos\Omega$, $\cos(2\Omega) = 2\cos^2(\Omega) - 1 = 1 - 2\sin^2\Omega$, $\exp(i\Omega) = \cos(\Omega) + i\sin(\Omega)$ and

$$\sum_{t=1}^{T} t = \frac{T(T+1)}{2} \qquad \sum_{t=1}^{T} t^2 = \frac{T(T+1)(2T+1)}{6}.$$

Solution

(i) Since $\{\varepsilon_t\}$ are standard normal iid random variables the likelihood is

$$\mathcal{L}_T(A, B, \omega) = -\frac{1}{2}\sum_{t=1}^{T}(X_t - A\cos(\omega t) - B\sin(\omega t))^2.$$

(ii) It is straightforward to show that

$$\begin{aligned}
& -2\mathcal{L}_T(A,B,\omega) \\
= & \sum_{t=1}^{T} X_t^2 - 2\sum_{t=1}^{T} X_t\big(A\cos(\omega t) + B\sin(\omega t)\big) \\
& + A^2\sum_{t=1}^{T}\cos^2(\omega t) + B^2\sum_{t=1}^{T}\sin^2(\omega t) + 2AB\sum_{t=1}^{T}\sin(\omega t)\cos(\omega t) \\
= & \sum_{t=1}^{T} X_t^2 - 2\sum_{t=1}^{T} X_t\big(A\cos(\omega t) + B\sin(\omega t)\big) + \\
& \frac{A^2}{2}\sum_{t=1}^{T}(1+\cos(2\omega)) + \frac{B^2}{2}\sum_{t=1}^{T}(1-\cos(2\omega)) + AB\sum_{t=1}^{T}\sin(2\omega) \\
= & \sum_{t=1}^{T} X_t^2 - 2\sum_{t=1}^{T} X_t\big(A\cos(\omega t) + B\sin(\omega t)\big) + \frac{T}{2}(A^2+B^2) + \\
& \frac{(A^2-B^2)}{2}\sum_{t=1}^{T}\cos(2\omega) + AB\sum_{t=1}^{T}\sin(2\omega) \\
= & \mathcal{S}_T(A,B,\omega) + \frac{(A^2-B^2)}{2}\sum_{t=1}^{T}\cos(2\omega) + AB\sum_{t=1}^{T}\sin(2\omega)
\end{aligned}$$

Now by using (6.13) we have

$$-2\mathcal{L}_T(A,B,\omega) \;=\; \mathcal{S}_T(A,B,\omega) + O(1),$$

as required.

(iii) To obtain the profile likelihood, let us suppose that ω is known, Then the mle of A and B (using $\frac{-1}{2}\mathcal{S}_T$) is

$$\hat{A}_T(\omega) = \frac{2}{T}\sum_{t=1}^{T} X_t\cos(\omega t) \quad \hat{B}_T(\omega) = \frac{2}{T}\sum_{t=1}^{T} X_t\sin(\omega t).$$

Thus the profile likelihood (using the approximation $\mathcal{S}_T$) is

$$\begin{aligned}
-\frac{1}{2}\mathcal{S}_p(\omega) &= \frac{-1}{2}\bigg(\sum_{t=1}^{T} X_t^2 - 2\sum_{t=1}^{T} X_t\big(\hat{A}_T(\omega)\cos(\omega t) + \hat{B}(\omega)\sin(\omega t)\big) + \frac{T}{2}(\hat{A}_T(\omega)^2 + \hat{B}(\omega)^2)\bigg) \\
&= \frac{-1}{2}\bigg(\sum_{t=1}^{T} X_t^2 - \frac{T}{2}\Big[\hat{A}_T(\omega)^2 + \hat{B}_T(\omega)^2\Big]\bigg).
\end{aligned}$$

Thus the ω which maximises $-\frac{1}{2}\mathcal{S}_p(\omega)$ is the parameter that maximises $\hat{A}_T(\omega)^2 + \hat{B}_T(\omega)^2$. Since $\hat{A}_T(\omega)^2 + \hat{B}_T(\omega)^2 = \frac{1}{2T}|\sum_{t=1}^T X_t \exp(it\omega)|$, we have

$$\begin{aligned}\hat{\omega}_T &= \arg\max_{\omega}(-1/2)\mathcal{S}_p(\omega) = \arg\max_{\omega}\big(\hat{A}_T(\omega)^2 + \hat{B}_T(\omega)^2\big) \\ &= \arg\max_{\omega}\big|\sum_{t=1}^T X_t \exp(it\omega)\big|^2,\end{aligned}$$

as required.

(iv) Differentiating both sides of (6.12) with respect to Ω and considering the real and imaginary terms gives $\sum_{t=1}^T t\cos(\Omega t) = O(T) \quad \sum_{t=1}^T t\sin(\Omega t) = O(T)$. Differentiating both sides of (6.12) twice wrt to Ω gives the second term.

(v) Differentiating $\mathcal{S}_T(A,B,\omega) = \big(\sum_{t=1}^T X_t^2 - 2\sum_{t=1}^T X_t\big(A\cos(\omega t)+B\sin(\omega t)\big) + \frac{1}{2}T(A^2+B^2)\big)$ twice wrt to A, B and ω gives

$$\begin{aligned}\frac{\partial \mathcal{S}_T}{\partial A} &= -2\sum_{t=1}^T X_t\cos(\omega t) + AT \\ \frac{\partial \mathcal{S}_T}{\partial B} &= -2\sum_{t=1}^T X_t\sin(\omega t) + BT \\ \frac{\partial \mathcal{S}_T}{\partial \omega} &= 2\sum_{t=1}^T AX_t t\sin(\omega t) - 2\sum_{t=1}^T BX_t t\cos(\omega t).\end{aligned}$$

and $\frac{\partial^2 \mathcal{S}_T}{\partial A^2} = T$, $\frac{\partial^2 \mathcal{S}_T}{\partial B^2} = T$, $\frac{\partial^2 \mathcal{S}_T}{\partial A \partial B} = 0$,

$$\begin{aligned}\frac{\partial^2 \mathcal{S}_T}{\partial \omega \partial A} &= 2\sum_{t=1}^T X_t t\sin(\omega t) \\ \frac{\partial^2 \mathcal{S}_T}{\partial \omega \partial B} &= -2\sum_{t=1}^T X_t t\cos(\omega t) \\ \frac{\partial^2 \mathcal{S}_T}{\partial \omega^2} &= 2\sum_{t=1}^T t^2 X_t\big(A\cos(\omega t) + B\sin(\omega t)\big).\end{aligned}$$

Now taking expectations of the above and using (v) we have

$$\begin{aligned}E\Big(\frac{\partial^2 \mathcal{S}_T}{\partial \omega \partial A}\Big) &= 2\sum_{t=1}^T t\sin(\omega t)\big(A\cos(\omega t) + B\sin(\omega t)\big) \\ &= 2B\sum_{t=1}^T t\sin^2(\omega t) + 2\sum_{t=1}^T At\sin(\omega t)\cos(\omega t) \\ &= B\sum_{t=1}^T t(1-\cos(2\omega t)) + A\sum_{t=1}^T t\sin(2\omega t) = \frac{T(T+1)}{2}B + O(T) = B\frac{T^2}{2} + O(T).\end{aligned}$$

Using a similar argument we can show that $E(\frac{\partial^2 \mathcal{S}_T}{\partial\omega\partial B}) = -A\frac{T^2}{2} + O(T)$ and

$$\begin{aligned} E(\frac{\partial^2 \mathcal{S}_T}{\partial\omega^2}) &= 2\sum_{t=1}^{T} t^2\Big(A\cos(\omega t) + B\sin(\omega t)\Big)^2 \\ &= (A^2+B^2)\frac{T(T+1)(2T+1)}{6} + O(T^2) = (A^2+B^2)T^3/3 + O(T^2). \end{aligned}$$

Since $E(-\nabla^2\mathcal{L}_T) \approx \frac{1}{2}E(\nabla^2\mathcal{S}_T)$, this gives the required result.

(vi) Noting that the asymptotic variance for the profile likelihood estimator $\hat{\omega}_T$

$$\Big(I_{\omega,\omega} - I_{\omega,(AB)}I_{A,B}^{-1}I_{(BA),\omega}\Big)^{-1},$$

by subsituting (vi) into the above we have

$$2\Big(\frac{A^2+B^2}{6}T^3 + O(T^2)\Big)^{-1} \approx \frac{12}{(A^2+B^2)T^3}$$

Thus we observe that the asymptotic variance of $\hat{\omega}_T$ is $O(T^{-3})$.

Typically estimators have a variance of order $O(T^{-1})$, so we see that the estimator $\hat{\omega}_T$ variance which converges to zero, much faster. Thus the estimator is extremely good compared with the majority of parameter estimators.

Example: An application of profiling in survival analysis

Question (This question also uses some methods from Survival Analysis which is covered later in this course - see Sections 13.1 and 19.1).

Let T_i denote the survival time of an electrical component. It is known that the regressors x_i influence the survival time T_i. To model the influence the regressors have on the survival time the Cox-proportional hazard model is used with the exponential distribution as the baseline distribution and $\psi(x_i;\beta) = \exp(\beta x_i)$ as the link function. More precisely the survival function of T_i is

$$\mathcal{F}_i(t) = \mathcal{F}_0(t)^{\psi(x_i;\beta)},$$

where $\mathcal{F}_0(t) = \exp(-t/\theta)$. Not all the survival times of the electrical components are observed, and there can arise censoring. Hence we observe $Y_i = \min(T_i, c_i)$, where c_i is the censoring time and δ_i, where δ_i is the indicator variable, where $\delta_i = 0$ denotes censoring of the ith component and $\delta_i = 1$ denotes that it is not censored. The parameters β and θ are unknown.

(i) Derive the log-likelihood of $\{(Y_i, \delta_i)\}$.

(ii) Compute the profile likelihood of the regression parameters β, profiling out the baseline parameter θ.

Solution

(i) The survivial function and the density are

$$f_i(t) = \psi(x_i;\beta)\big\{\mathcal{F}_0(t)\big\}^{[\psi(x_i;\beta)-1]} f_0(t) \quad \text{and} \quad \mathcal{F}_i(t) = \mathcal{F}_0(t)^{\psi(x_i;\beta)}.$$

Hence for this example we have

$$\begin{aligned} \log f_i(t) &= \log\psi(x_i;\beta) - \big[\psi(x_i;\beta)-1\big]\frac{t}{\theta} - \log\theta - \frac{t}{\theta} \\ \log \mathcal{F}_i(t) &= -\psi(x_i;\beta)\frac{t}{\theta}. \end{aligned}$$

Therefore, the likelihood is

$$\begin{aligned} \mathcal{L}_n(\beta,\theta) &= \sum_{i=1}^n \delta_i\big\{\log\psi(x_i;\beta) + \log f_0(T_i) + (\psi(x_i;\beta)-1)\log\mathcal{F}_0(t)\big\} + \\ &\quad \sum_{i=1}^n (1-\delta_i)\big\{\psi(x_i;\beta)\log\mathcal{F}_0(t)\big\} \\ &= \sum_{i=1}^n \delta_i\big\{\log\psi(x_i;\beta) - \log\theta\big\} - \sum_{i=1}^n \psi(x_i;\beta)\frac{T_i}{\theta} \end{aligned}$$

(ii) Keeping β fixed and differentiating the above with respect to θ and equating to zero gives

$$\frac{\partial\mathcal{L}_n}{\partial\theta} = \sum_{i=1}^n \delta_i\big\{-\frac{1}{\theta}\big\} + \sum_{i=1} \psi(x_i;\beta)\frac{T_i}{\theta^2}$$

and

$$\hat{\theta}(\beta) = \frac{\sum_{i=1}^n \psi(x_i;\beta)T_i}{\sum_{i=1}^n \delta_i}.$$

Hence the profile likelihood is

$$\ell_P(\beta) = \sum_{i=1}^n \delta_i\big\{\log\psi(x_i;\beta) - \log\hat{\theta}(\beta)\big\} - \sum_{i=1}^n \psi(x_i;\beta)\frac{T_i}{\hat{\theta}(\beta)}.$$

Hence to obtain an estimator of β we maximise the above with respect to β.

An application of profiling in semi-parametric regression

We now consider how the profile 'likelihood' (we use inverted commas here because we do not use the likelihood, but least squares instead) can be used in semi-parametric regression. Recently this type of method has been used widely in various semi-parametric models. This section needs a little knowledge of nonparametric regression, which is considered later in this course. Suppose we observe (Y_t, U_t, X_t) where

$$Y_t = \beta X_t + \phi(U_t) + \varepsilon_t,$$

(Y_t, X_t, U_t) are iid random variables and ϕ is an unknown function. To estimate β, we first profile out $\phi(\cdot)$, which we estimate as if β were known. In other other words, we suppose that β is known and let $Y_t(\beta) = Y_t - \beta X_t$. We then estimate $\phi(\cdot)$ using the classic local least estimator, in other words the $\phi(\cdot)$ which minimises the criterion

$$\begin{aligned}
\hat{\phi}_\beta(u) &= \arg\min_a \sum_t W_b(u-U_t)(Y_t(\beta)-a)^2 = \frac{\sum_t W_b(u-U_t)Y_t(\beta)}{\sum_t W_b(u-U_t)} \\
&= \frac{\sum_t W_b(u-U_t)Y_t}{\sum_t W_b(u-U_t)} - \beta\frac{\sum_t W_b(u-U_t)X_t}{\sum_t W_b(u-U_t)} \\
&:= G_b(u) - \beta H_b(u), \qquad (6.14)
\end{aligned}$$

where

$$G_b(u) = \frac{\sum_t W_b(u-U_t)Y_t}{\sum_t W_b(u-U_t)} \quad \text{and} \quad H_b(u) = \frac{\sum_t W_b(u-U_t)X_t}{\sum_t W_b(u-U_t)}.$$

Thus, given β the estimator of ϕ and the residuals ε_t are $\hat{\phi}_\beta(u) = G_b(u) - \beta H_b(u)$ and $Y_t - \beta X_t - \hat{\phi}_\beta(U_t)$. Given the estimated residuals $Y_t - \beta X_t - \hat{\phi}_\beta(U_t)$ we can now use least squares to estimate coefficient β, where

$$\begin{aligned}
\mathcal{L}_T(\beta) &= \sum_t \big(Y_t - \beta X_t - \hat{\phi}_\beta(U_t)\big)^2 \\
&= \sum_t \big(Y_t - \beta X_t - G_b(U_t) + \beta H_b(U_t)\big)^2 \\
&= \sum_t \big(Y_t - G_b(U_t) - \beta[X_t - H_b(U_t)]\big)^2.
\end{aligned}$$

Therefore, the least squares estimator of β is

$$\hat{\beta}_{b,T} = \frac{\sum_t [Y_t - G_b(U_t)][X_t - H_b(U_t)]}{\sum_t [X_t - H_b(U_t)]^2}.$$

Using $\beta_{b,T}$ we can then estimate (6.15). We observe how we have the used the principle of profiling to estimate the unknown parameters. There is a large literature on this, including

Wahba, Speckman, Carroll, Fan etc. In particular it has been shown that under some conditions on b (as $T \to \infty$), the estimator $\hat{\beta}_{b,T}$ has the usual $\sqrt{T}$ rate of convergence.

It should be mentioned that using random regressors U_t are not necessary. It could be that $U_t = \frac{t}{T}$ (on a grid). In this case

$$\begin{aligned} \hat{\phi}_\beta(u) &= \arg\min_a \sum_t W_b(u - \frac{t}{T})(Y_t(\beta) - a)^2 = \frac{\sum_t W_b(u - \frac{t}{T})Y_t(\beta)}{\sum_t W_b(u - \frac{t}{T})} \\ &= \sum_t W_b(u - \frac{t}{T})Y_t - \beta \sum_t W_b(u - U_t)X_t \\ &:= G_b(u) - \beta H_b(u), \end{aligned} \tag{6.15}$$

where

$$G_b(u) = \sum_t W_b(u - \frac{t}{T})Y_t \quad \text{and} \quad H_b(u) = \sum_t W_b(u - \frac{t}{T})X_t.$$

Using the above estimator of $\phi(\cdot)$ we continue as before.