Chapter 4

Non-standard inference

As we mentioned in Chapter 2 the the log-likelihood ratio statistic is useful in the context of statistical testing because typically it is "pivotal" (does not depend on any nuisance) under the null hypothesis. Typically, the log-likelihood ratio statistic follows a chi-square distribution under the null hypothesis. However, there are realistic situations where the this statistic does not follow a chi-square distribution and the purpose of this chapter is to consider some of these cases.

At the end of this chapter we consider what happens when the "regularity" conditions are not satisfied.

4.1 Detection of change points

This example is given in Davison (2004), pages 141, and will be considered in class. It is not related to the boundary problems discussed below but none the less is very interesting.

4.2 Estimation on the boundary of the parameter space

In this section we consider the distribution of parameters which are estimated on the boundary of the parameter space. We will use results from Chapter 2.

4.2.1 Estimating the mean on the boundary

There are situations where the parameter to be estimated lies on the boundary (or very, very close to it). In such cases the limiting distribution of the the parameter may not be normal (since when we maximise the likelihood we do so over the parameter space and not outside it). This will not impact Wald based tests (by much), but it will have an impact on the log-likelihood ratio test.

To understand the changes involved, we start with a simple example.

Suppose $X_i \sim \mathcal{N}(\mu, 1)$, where the mean μ is unknown. In addition it is known that the mean is non-negative hence the parameter space of the mean is $\Theta = [0, \infty)$. In this case $\bar{X}$ can no longer be the MLE because there will be some instances where $\bar{X} < 0$. Let us relook at the maximum likelihood on the restricted parameter space

$$\hat{\mu}_n = \arg\max_{\mu \in \Theta} \mathcal{L}_n(\mu) = \arg\max_{\mu \in \Theta} \frac{-1}{2} \sum_{i=1}^{n} (X_i - \mu)^2.$$

Since $\mathcal{L}_n(\mu)$ is concave over μ, we see that the MLE estimator is

$$\hat{\mu}_n = \begin{cases} \bar{X} & \bar{X} \geq 0 \\ 0 & \bar{X} < 0. \end{cases}$$

Hence in this restricted space it is not necessarily true that $\frac{\partial \mathcal{L}_n(\mu)}{\partial \mu} \rfloor_{\hat{\mu}_n} \neq 0$, and the usual Taylor expansion method cannot be used to derive normality. Indeed we will show that it is not normal.

We recall that $\sqrt{n}(\bar{X} - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\mu)^{-1})$ or equivalently $\frac{1}{\sqrt{n}} \frac{\partial \mathcal{L}_n(\mu)}{\partial \mu} \rfloor_{\bar{X}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(\mu))$. Hence if the true parameter $\mu = 0$, then approximately half the time $\bar{X}$ will be less than zero and the other half it will be greater than zero. This means that half the time $\hat{\mu}_n = 0$ and the other half it will be greater than zero. Therefore the distribution function of $\hat{\mu}_n$ is

$$\begin{aligned} & P(\sqrt{n}\widehat{\mu}_n \leq x) = P(\sqrt{n}\hat{\mu}_n = 0 \text{ or } 0 < \sqrt{n}\hat{\mu}_n \leq x) \\ \approx & \begin{cases} 0 & x \leq 0 \\ 1/2 & x = 0 \\ 1/2 + P(0 < \sqrt{n}\bar{X} \leq x) = \Phi(\sqrt{n}\bar{X} \leq x) & x > 0 \end{cases}, \end{aligned}$$

where Φ denotes the distribution function of the normal distribution. Observe the distribution of $\sqrt{n}\bar{X}$ is a mixture of a point mass and a density. However, this result does not

change our testing methodology based on the sample mean. For example, if we want to test $H_0 : \mu = 0$ vs $H_A : \mu > 0$, then we use the estimator $\widehat{\mu}_n$ and the p-value is

$$1 - \Phi\left(\sqrt{n}\widehat{\mu}_n\right)$$

which is the p-value for the one-sided test (using the normal distribution).

Now we consider using the log-likelihood ratio test to test the $H_0 : \mu = 0$ vs $H_A : \mu > 0$. In this set-up the test statistic is

$$W_n = 2\left\{\arg\max_{\mu\in[0,\infty)} \mathcal{L}_n(\mu) - \mathcal{L}_n(0)\right\} = 2\big\{\mathcal{L}_n(\widehat{\mu}_n) - \mathcal{L}_n(0)\big\}.$$

However, since the derivative of the likelihood at $\widehat{\mu}_n$ is not necessarily zero, means that W will not be a standard chi-square distribtion. To obtain the distribution we note that likelihoods under $\mu \in [0, \infty)$ and $\mu = 0$ can be written as

$$\mathcal{L}_n(\hat{\mu}_n) = -\frac{1}{2}\sum_{i=1}^{n}(X_i - \hat{\mu}_n)^2 \qquad \mathcal{L}_n(0) = -\frac{1}{2}\sum_{i=1}^{n} X_i^2.$$

Thus we observe that when $\bar{X} \leq 0$ then $\mathcal{L}_n(\hat{\mu}_n) = \mathcal{L}_n(0)$ and

$$2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \;=\; \left\{\begin{array}{ccc} 0 & \bar{X} \leq 0 & P(\bar{X} \leq 0) \approx 1/2 \\ n|\bar{X}|^2 & \bar{X} > 0 & P(\bar{X} > 0) \approx 1/2 \end{array}\right.$$

Hence we have that

$$\begin{aligned} & P(2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \leq x) \\ = \;& P(2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \leq x\big|\bar{X} \leq 0)P(\bar{X} \leq 0) + P(2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \leq x\big|\bar{X} > 0)P(\bar{X} > 0). \end{aligned}$$

Now using that

$$P(2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \leq x\big|\bar{X} \leq 0) = \left\{\begin{array}{cc} 0 & x < 0 \\ 1 & x \geq 0 \end{array}\right. ,$$

$$P(2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \leq x\big|\bar{X} > 0) = P(n\bar{X}^2 \leq x\big|\bar{X} > 0) \approx \left\{\begin{array}{cc} 0 & x < 0 \\ \chi^2_1 & x > 0 \end{array}\right.$$

and $P(\sqrt{n}\bar{X} < 0) = 1/2$, gives

$$P(2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \leq x) \;=\; \left\{\begin{array}{cc} 0 & x \leq 0 \\ 1/2 & x = 0 \\ 1/2 + \frac{1}{2}P(n|\bar{X}|^2 \leq x) & x > 0 \end{array}\right.$$

Therefore

$$P(2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \leq x) = \frac{1}{2} + \frac{1}{2}P(\chi^2 \leq x) = \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2,$$

we use the χ_0^2 notation to denote the point mass at zero. Therefore, suppose we want to test the hypothesis $H_0 : \mu = 0$ against the hypothesis $H_A : \mu > 0$ using log likelihood ratio test. We would evaluate $W_n = 2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\}$ and find the p such that

$$\frac{1}{2} + \frac{1}{2}P(W_n \leq \chi_1^2) = 1 - p.$$

This is the p-value, which we then use to make the decision on the test.

Remark 4.2.1 *Essentially what has been done is turned the log-likelihood test for the mean, which is a two-sided test, into a one-sided test.*

(i) It is clear that without a boundary testing $H_0 : \mu = 0$ against $H_A : \mu \neq 0$ the LLRT is simply

$$2\big\{\mathcal{L}_n(\bar{X}) - \mathcal{L}_n(0)\big\} = n|\bar{X}|^2 \xrightarrow{\mathcal{D}} \chi_1^2,$$

under the null.

Example, $n = 10$ and $\bar{x} = 0.65$ the p-value for the above hypothesis is

$$\begin{aligned} P\left(W_n > 10 \times (0.65)^2\right) &= P\left(\chi_1^2 > 10 \times (0.65)^2\right) \\ &= 1 - P(\chi_1^2 \leq 4.2) = 1 - 0.96 = 0.04. \end{aligned}$$

The p-value is 4%.

(ii) On the other hand, to test $H_0 : \mu = 0$ against the hypothesis $H_A : \mu > 0$ we use

$$2\big\{\mathcal{L}_n(\hat{\mu}_n) - \mathcal{L}_n(0)\big\} \xrightarrow{\mathcal{D}} \frac{1}{2} + \frac{1}{2}\chi_1^2.$$

Example: Using the same data, but the one-sided test we have

$$\begin{aligned} P\left(W_n > 10 \times (0.65)^2\right) &= 1 - P\left(W_n \leq 10 \times (0.65)^2\right) \\ &= 1 - \left(\frac{1}{2} + \frac{1}{2}P\left(\chi_1^2 \leq 10 \times (0.65)^2\right)\right) = \frac{1}{2}\left(1 - P(\chi_1^2 \leq 4.2)\right) = 0.02. \end{aligned}$$

The p-value is 2%. Thus, as we would expect, the result of the one-sided test simply gives half the p-value corresponding to the two-sided test.

Exercise 4.1 *The survivial time of disease A follow an exponential distribution, where the distribution function has the form* $f(x) = \lambda^{-1}\exp(-x/\lambda)$. *Suppose that it is known that at least one third of all people who have disease A survive for more than 2 years.*

(i) Based on the above information obtain the appropriate parameter space for λ. *Let* λ_B *denote the lower boundary of the parameter space and* Θ *the corresponding parameter space.*

(ii) What is the maximum likelihood estimator of $\widehat{\lambda}_n = \arg\max_{\lambda\in\Theta}\mathcal{L}_n(\lambda)$.

(iii) Derive the sampling properties of maximum likelihood estimator of λ, *for the cases* $\lambda = \lambda_B$ *and* $\lambda > \lambda_B$.

(iv) Suppose the true parameter is λ_B *derive the distribution of* $2[\max_{\theta\in\Theta}\mathcal{L}_n(\lambda) - \mathcal{L}_n(\lambda_B)]$.

4.2.2 General case with parameter on the boundary

It was straightfoward to derive the distributions in the above examples because a closed form expression exists for the estimator. However the same result holds for general maximum likelihood estimators; *so long as certain regularity conditions are satisfied.*

Suppose that the log-likelihood is $\mathcal{L}_n(\theta)$, the parameter space is $[0,\infty)$ and

$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta}\mathcal{L}_n(\theta).$$

We consider the case that the true parameter $\theta_0 = 0$. To derive the limiting distribution we extend the parameter space $\widetilde{\Theta}$ such that $\theta_0 = 0$ is an *interior point* of $\widetilde{\Theta}$. Let

$$\widetilde{\theta}_n \in \arg\max_{\theta\in\widetilde{\Theta}}\mathcal{L}_n(\theta),$$

this is the maximum likelihood estimator in the non-constrained parameter space. We assume that for this non-constrained estimator $\sqrt{n}(\widetilde{\theta}_n - 0) \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(0)^{-1})$ (this needs to be verified and may not always hold). This means that for sufficiently large n, the likelihood will have a maximum close to 0 and that in the neighbourhood of zero, the likelihood is concave (with only one maximum). We use this result to obtain the distribution of the restricted estimator. The log-likelihood ratio involving the restricted estimator is

$$\begin{aligned} W_n &= 2\left(\arg_{\theta\in[0,\infty)}\mathcal{L}_n(\theta) - \mathcal{L}_n(0)\right) \\ &= 2\left(\mathcal{L}_n(\widehat{\theta}_n) - \mathcal{L}_n(0)\right). \end{aligned}$$

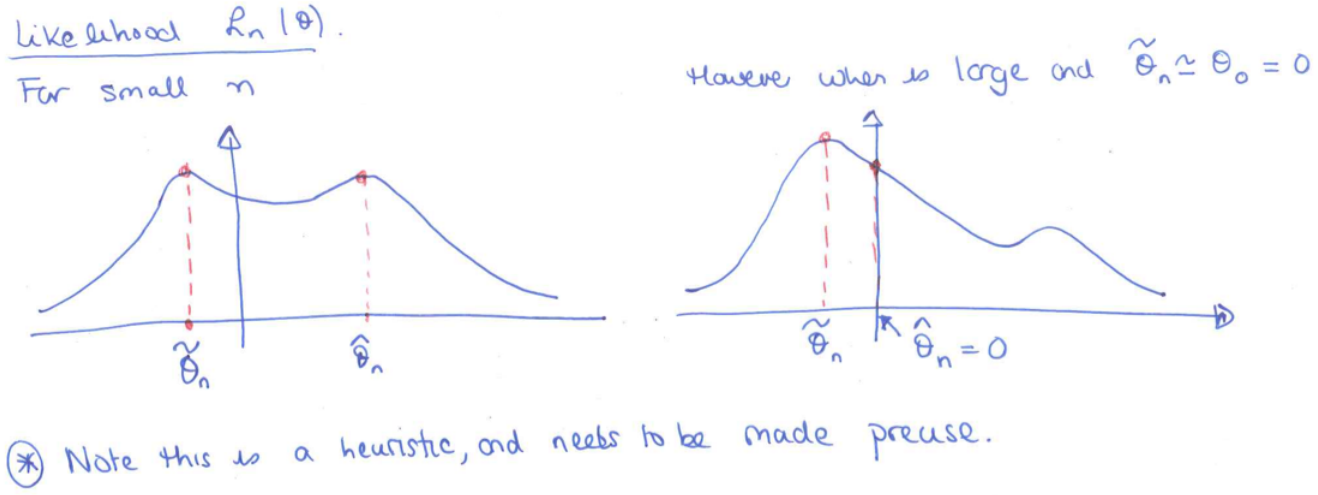

Figure 4.1: A plot of the likelihood for large and small n. For large n, the likelihood tends to be concave about the true parameter, which in this case is zero.

Roughly speaking $\widehat{\theta}_n$ can be considered as a "reflection" of $\widetilde{\theta}_n$ i.e. if $\widetilde{\theta}_n < 0$ then $\widehat{\theta}_n = 0$ else $\widehat{\theta}_n = \widetilde{\theta}_n$ (see Figure 4.1) (since for a sufficiently large sample size, if $\widetilde{\theta}_n < 0$, then the maximum within $[0, \infty)$ will lie at zero). We use this principle to obtain the distribution of W_n by conditioning on $\widetilde{\theta}_n$

$$P\left(W_n \leq x\right) = P(W_n \leq x | \widetilde{\theta}_n \leq 0) P(\widetilde{\theta}_n \leq 0) + P(W_n \leq x | \widetilde{\theta}_n > 0) P(\widetilde{\theta}_n > 0).$$

Now using that $\sqrt{n}\widetilde{\theta}_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, I(0)^{-1})$ and that $\mathcal{L}_n(\theta)$ is close to concave about its maximum thus for $\widetilde{\theta}_n \leq 0$ we have $W_n = 0$, and we have a result analogous to the mean case

$$P\left(W_n \leq x\right) \;\; = \;\; \frac{1}{2} P(W_n \leq x | \widetilde{\theta}_n \leq 0) + \frac{1}{2} P(W_n \leq x | \widetilde{\theta}_n > 0) = \frac{1}{2} + \frac{1}{2}\chi^2_1.$$

The precise argument for the above uses a result by Chernoff (1954), who shows that

$$W_n \stackrel{\mathcal{D}}{=} \max_{\theta \in [0,\infty)} \left[-(Z-\theta) I(0) (Z-\theta)\right] + Z I(0) Z + o_p(1), \tag{4.1}$$

where $Z \sim \mathcal{N}(0, I(0)^{-1})$ (and is the same for both quadratic forms). Observe that when $Z < 0$ the above is zero, whereas when $Z > 0$ $\max_{\theta \in [0,\infty)} \left[-(Z-\theta) I(0) (Z-\theta)\right] = 0$ and we have the usual chi-square statistic.

To understand the approximation in (4.1) we return to the log-likelihood ratio and add and subtract the maximum likelihood estimator based on the non-restricted parameter space $\widetilde{\Theta}$

$$2\left[\max_{\theta\in\Theta}\mathcal{L}_n(\theta)-\mathcal{L}_n(\theta_0)\right]=2\left[\max_{\theta\in\Theta}\mathcal{L}_n(\theta)-\max_{\theta\in\widetilde{\Theta}}\mathcal{L}_n(\theta)\right]+2\left[\max_{\theta\in\widetilde{\Theta}}\mathcal{L}_n(\theta)-\mathcal{L}_n(\theta_0)\right] \tag{4.2}$$

Now we do the usual Taylor expansion about $\widetilde{\theta}_n$ (which guarantees that the first derivative is zero) for both terms to give

$$\begin{aligned}
&2\left[\max_{\theta\in\Theta}\mathcal{L}_n(\theta)-\mathcal{L}_n(\theta_0)\right]\\
=\;&-n\left(\widetilde{\theta}_n-\widehat{\theta}_n\right)I(\theta_0)\left(\widetilde{\theta}_n-\widehat{\theta}_n\right)+n\left(\widetilde{\theta}_n-\theta_0\right)I(\theta_0)\left(\widetilde{\theta}_n-\theta_0\right)+o_p(1)\\
=\;&-n\left(\left[\widetilde{\theta}_n-\theta_0\right]-\left[\widehat{\theta}_n-\theta_0\right]\right)I(\theta_0)\left(\left[\widetilde{\theta}_n-\theta_0\right]-\left[\widehat{\theta}_n-\theta_0\right]\right)+n\left(\widetilde{\theta}_n-\theta_0\right)I(\theta_0)\left(\widetilde{\theta}_n-\theta_0\right).
\end{aligned}$$

We recall that asymptotically $\sqrt{n}\left(\widetilde{\theta}_n-\theta_0\right)\sim\mathcal{N}(0,I(\theta_0)^{-1})$. Therefore we define the random variable $\sqrt{n}\left(\widetilde{\theta}_n-\theta_0\right)\sim Z\sim\mathcal{N}(0,I(\theta_0)^{-1})$ and replace this in the above to give

$$\begin{aligned}
&2\left[\mathcal{L}_n(\widehat{\theta}_n)-\mathcal{L}_n(\theta_0)\right]\\
\stackrel{\mathcal{D}}{=}\;&-\left(Z-n^{1/2}\left[\widehat{\theta}_n-\theta_0\right]\right)I(\theta_0)\left(Z-n^{1/2}\left[\widehat{\theta}_n-\theta_0\right]\right)+ZI(\theta_0)Z.
\end{aligned}$$

Finally, it can be shown (see, for example, Self and Liang (1987), Theorem 2 or Andrews (1999), Section 4.1) that $\sqrt{n}(\widehat{\theta}_n-\theta_0)\in\Theta-\theta_0=\Lambda$, where Λ is a convex cone about θ_0 (this is the terminology that is often used); in the case that $\Theta=[0,\infty)$ and $\theta_0=0$ then $\sqrt{n}(\widehat{\theta}_n-\theta_0)\in\Lambda=[0,\infty)$ (the difference can never be negative). And that the maximum likelihood estimator is equivalent to the maximum of the quadratic form over Θ i.e.

$$\begin{aligned}
2\left[\mathcal{L}_n(\widehat{\theta}_n)-\mathcal{L}_n(\theta_0)\right]&\stackrel{\mathcal{D}}{=}\max_{\theta\in\Theta}-\left(Z-n^{1/2}\left[\widehat{\theta}_n-\theta_0\right]\right)I(\theta_0)\left(Z-n^{1/2}\left[\widehat{\theta}_n-\theta_0\right]\right)+ZI(\theta_0)Z\\
&=\max_{\theta\in\Theta-\theta_0=[0,\infty)=\Theta}-(Z-\theta)\,I(\theta_0)\,(Z-\theta)+ZI(\theta_0)Z,
\end{aligned}$$

which gives (4.1).

Example 4.2.1 (Example 4.39 (page 140) in Davison (2002)) *In this example Davison reparameterises the t-distribution. It is well known that if the number of degrees of freedom of a t-distribution is one, it is the Cauchy distribution, which has extremely thick tails (such that the mean does not exist). At the other extreme, if we let the number of*

degrees of freedom tend to ∞, then the limit is a normal distribution (where all moments exist). In this example, the t-distribution is reparameterised as

$$f(y;\mu,\sigma^2,\psi)=\frac{\Gamma\left[\frac{(1+\psi^{-1})}{2}\right]\psi^{1/2}}{(\sigma^2\pi)^{1/2}\Gamma(\frac{1}{2\pi})}\left(1+\frac{\psi(y-\mu)^2}{\sigma^2}\right)^{-(\psi^{-1}+1)/2}$$

It can be shown that $\lim_{\psi\to 1} f(y;\mu,\sigma^2,\psi)$ is a t-distribution with one-degree of freedom and at the other end of the spectrum $\lim_{\psi\to 0} f(y;\mu,\sigma^2,\psi)$ is a normal distribution. Thus $0<\psi\leq 1$, and the above generalisation allows for fractional orders of the t-distribution.

In this example it is assumed that the random variables $\{X_i\}$ have the density $f(y;\mu,\sigma^2,\psi)$, and our objective is to estimate ψ, when $\psi\to 0$, this the true parameter is on the boundary of the parameter space $(0,1]$ (it is just outside it!). Using similar, arguments to those given above, Davison shows that the limiting distribution of the MLE estimator is close to a mixture of distributions (as in the above example).

Testing on the boundary in the presence of independent nuisance parameters

Suppose that the iid random variables come from the distribution $f(x;\theta,\psi)$, where (θ,ψ) are unknown. We will suppose that θ is a univariate random variable and ψ can be multivariate. Suppose we want to test $H_0:\theta=0$ vs $H_A:\theta>0$. In this example we are testing on the boundary in the presence of nuisance parameters ψ.

Example 4.2.2 *Examples include the random coefficient regression model*

$$Y_i=(\alpha+\eta_i)X_i+\varepsilon_i, \tag{4.3}$$

where $\{(Y_i,X_i)\}_{i=1}^n$ are observed variables. $\{(\eta_i,\varepsilon_i)\}_{i=1}^n$ are independent zero mean random vector, where $var((\eta_i,\varepsilon_i))=\mathrm{diag}(\sigma_\eta^2,\sigma_\varepsilon^2)$. We may want to test whether the underlying model is a classical regression model of the type

$$Y_i=\alpha X_i+\varepsilon_i,$$

vs the random regression model in (4.3). This reduces to testing $H_0:\sigma_\eta^2=0$ vs $H_A:\sigma_\eta^2>0$.

In this section we will assume that the Fisher information matrix associated for the mle of (θ,ψ) is block diagonal i.e. $\mathrm{diag}(I(\theta),I(\psi))$. In other words, if we did not constrain

the parameter space in the maximum likelihood estimation then

$$\sqrt{n}\begin{pmatrix} \widetilde{\theta}_n - \theta \\ \widetilde{\psi}_n - \psi \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \text{diag}(I(\theta), I(\psi))\right).$$

The log-likelihood ratio statistic for testing the hypothesis is

$$W_n = 2\left[\max_{\theta\in[0,\infty),\psi} \mathcal{L}_n(\theta,\psi) - \max_{\psi} \mathcal{L}_n(0,\psi)\right]$$

Now using the heuristics presented in the previous section we have

$$P\left(W_n \leq x\right) = P(W_n \leq x|\widetilde{\theta}_n \leq 0)P(\widetilde{\theta}_n \leq 0) + P(W_n \leq x|\widetilde{\theta}_n > 0)P(\widetilde{\theta}_n > 0).$$

The important observation is that because $(\widetilde{\theta}_n, \widetilde{\psi}_n)$ are asymptotically independent of each other, the estimator of $\widetilde{\theta}_n$ has no influence on the estimate of $\widetilde{\psi}_n$. Thus setting $\widehat{\theta}_n = 0$ will not change the estimator of ψ and $\widetilde{\psi}_n = \widehat{\psi}_n$

$$\begin{aligned} & 2\left[\max_{\theta\in[0,\infty),\psi} \mathcal{L}_n(\theta,\psi) - \max_{\psi} \mathcal{L}_n(0,\psi)\right] \big|\widetilde{\theta}_n < 0 \\ = \; & 2\left[\max_{\theta\in[0,\infty),\psi} \mathcal{L}_n(\widehat{\theta},\widehat{\psi}) - \max_{\psi} \mathcal{L}_n(0,\psi)\right] \big|\widetilde{\theta}_n < 0 \\ = \; & 2\left[\mathcal{L}_n(0,\widetilde{\psi}) - \max_{\psi} \mathcal{L}_n(0,\psi)\right] = 0. \end{aligned}$$

This gives the result

$$P\left(W_n \leq x\right) = \frac{1}{2}P(W_n \leq x|\widetilde{\theta}_n \leq 0) + \frac{1}{2}P(W_n \leq x|\widetilde{\theta}_n > 0) = \frac{1}{2} + \frac{1}{2}\chi^2_1.$$

However, it relies on the asymptotic independence of $\widetilde{\theta}_n$ and $\widetilde{\psi}_n$.

4.2.3 Estimation on the boundary with several parameters when the Fisher information is block diagonal

In the following section we summarize some of the results in Self and Liang (1987).

One parameter lies on the boundary and the rest do not

We now generalize the above to estimating the parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_{p+1})$. We start by using an analogous argument to that used in the mean case and then state the precise result from which it comes from.

Suppose the true parameter θ_1 lies on the boundary, say zero, however the other parameters $\theta_2, \ldots, \theta_{p+1}$ lie within the interior of the parameter space and the parameter space is denoted as Θ. Examples include mixture models where θ_1 is the variance (and cannot be negative!). We denote the true parameters as $\theta_0 = (\theta_{10} = 0, \theta_{20}, \ldots, \theta_{p+1,0})$. Let $\mathcal{L}_n(\theta)$ denote the log-likelihood. We make the informal assumption that if we were to extend the parameter space such that $\theta_0 = 0$ were in the interior of this new parameter space $\widetilde{\Theta}$ i.e. $(\theta_{10} = 0, \theta_{20}, \ldots, \theta_{p+1,0}) = (\theta_{10}, \underline{\theta}_p) \in int(\widetilde{\Theta})$, and $\widetilde{\theta} = (\widetilde{\theta}_1, \widetilde{\underline{\theta}}_p) = \arg\max_{\theta \in \widetilde{\Theta}} \mathcal{L}_n(\theta)$ then

$$\sqrt{n}\begin{pmatrix} \widetilde{\theta}_{1n} - \theta_0 \\ \widetilde{\underline{\theta}}_{pn} - \underline{\theta}_{p0} \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \begin{pmatrix} I_{11}(\theta_0) & 0 \\ 0 & I_{pp}(\theta_0) \end{pmatrix}^{-1}\right).$$

It is worth noting that the block diagonal nature of the information matrix assumes that the two sets of parameters are asymptotically independent. The asymptotic normality results needs to be checked; it does not always hold.[1]. Let $\widehat{\theta}_n = \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta)$ denote the maximum likelihood estimator in the restricted parameter space (with the cut off at zero). Our aim is to derive the distribution of

$$W_n = 2\left(\max_{\theta \in \Theta} \mathcal{L}_n(\theta) - \mathcal{L}_n(\theta_0)\right).$$

We use heuristics to obtain the distribution, by conditioning on the unrestricted estimator $\widetilde{\theta}_n$ (we make this a little more precisely later on). Conditioning on $\widetilde{\theta}_{1n}$ we have

$$P\left(W_n \leq x\right) = P(W_n \leq x | \widetilde{\theta}_{1n} \leq 0) P(\widetilde{\theta}_{1n} \leq 0) + P(W_n \leq x | \widetilde{\theta}_{1n} > 0) P(\widetilde{\theta}_{1n} > 0).$$

Again assuming that for large n, $\mathcal{L}_n(\theta)$ is concave about $\widetilde{\theta}_n$ such that when $\widetilde{\theta}_n < 0$, $\widehat{\theta}_n = 0$. However, asymptotic independence between $\widetilde{\theta}_{n1}$ and $\widetilde{\theta}_{np}$ (since the Fisher information matrix is block diagonal) means that setting $\widehat{\theta}_{n1} = 0$ does not change the estimator of θ_p $\widetilde{\theta}_{np}$ i.e. roughly speaking

$$2[\mathcal{L}_n(\widehat{\theta}_{1n}, \widehat{\theta}_{pn}) - \mathcal{L}_n(0, \theta_p)] | \widetilde{\theta}_{n2} < 0 = \underbrace{2[\mathcal{L}_n(0, \widetilde{\theta}_{pn}) - \mathcal{L}_n(0, \theta_p)]}_{\chi^2_p}$$

[1]Sometimes we cannot estimate on the boundary (consider some of the example considered in Chapter 2.9 with regards to the exponential family), sometimes the $\sqrt{n}$-rates and/or the normality result is completely different for parameters which are defined at the boundary (the Dickey-Fuller test is a notable example)

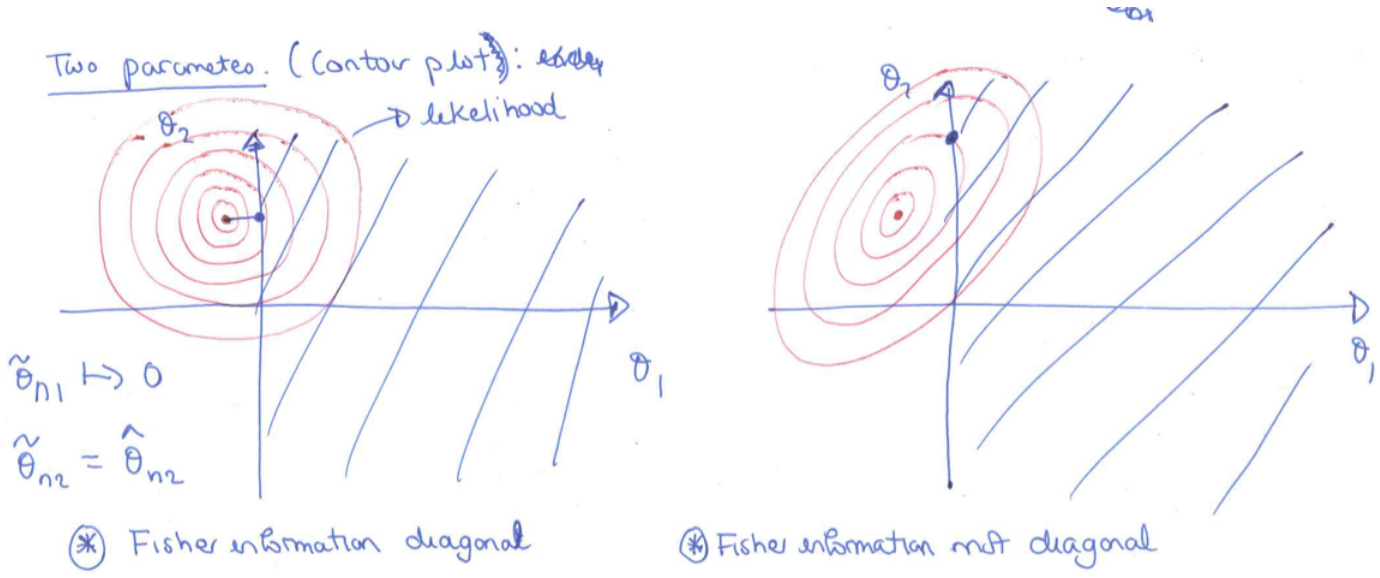

Figure 4.2: The likelihood for diagonal and nondiagonal Fisher information matrices.

If $\widetilde{\theta}_{n1}$ and $\widetilde{\theta}_{np}$ were dependent then the above equality does not hold and it is not a χ^2_p (see Figure 4.2). The above gives

$$\begin{aligned} P\left(W_n \leq x\right) &= P(\underbrace{W_n \leq x}_{\chi^2_p} | \widetilde{\theta}_{1n} \leq 0) P(\widetilde{\theta}_{1n} \leq 0) + P(\underbrace{W_n \leq x}_{\chi^2_{p+1}} | \widetilde{\theta}_{1n} > 0) P(\widetilde{\theta}_{1n} > 0) \\ &= \frac{1}{2}\chi^2_p + \frac{1}{2}\chi^2_{p+1}. \end{aligned} \tag{4.4}$$

See Figure 4.3 for a plot of the parameter space and associated probabilities.

The above is a heuristic argument. If one wanted to do it precisely one needs to use the asymptotic equivalent (based on the same derivations given in(4.2)) where (under certain regularity conditions) we have

$$\begin{aligned} W_n &\stackrel{\mathcal{D}}{=} \max_{\theta \in \Theta} [-(Z-\theta) I(0) (Z-\theta)] + Z I(\theta_0) Z + o_p(1) \\ &= \max_{\theta_1 \in [0,\infty)} [-(Z-\theta_1) I_{11}(\theta_0)(Z-\theta_1)] + Z I_{11}(\theta_0) Z \\ &\quad + \underbrace{\max_{\underline{\theta}_p \in \Theta_p} \left[-(\underline{Z}_p - \underline{\theta}_p) I_{11}(\theta_0)(\underline{Z}_p - \underline{\theta}_p)\right]}_{=0} + \underline{Z}_p I_{pp}(\theta_0) \underline{Z}_p \\ &= \max_{\theta_1 \in [0,\infty)} [-(Z-\theta_1) I_{11}(\theta_0)(Z-\theta_1)] + Z I_{11}(\theta_0) Z \\ &\quad + \underline{Z}_p I_{pp}(\theta_0) \underline{Z}_p \end{aligned}$$

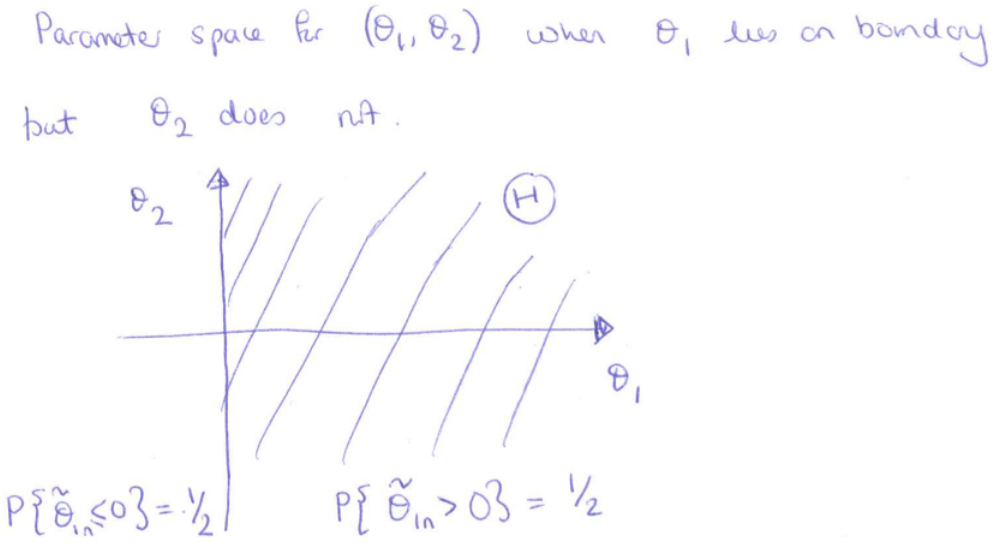

Figure 4.3: Two parameters: one on boundary and one in interior.

where $Z \sim N(0, I_{11}(\theta_0)^{-1})$ and $\underline{Z}_p \sim N(0, I_{pp}(\theta_0)^{-1})$ (Z and $\underline{Z}_p$ are independent). Using the above we can obtain the same distribution as that given in (4.4)

More than one parameter lies on the boundary

Suppose that the parameter space is $\Theta = [0, \infty) \times [0, \infty)$ and the true parameter $\theta_0 = (\theta_{10}, \theta_{20}) = (0, 0)$ (thus is on the boundary). As before we make the informal assumption that we can extend the parameter space such that θ_0 lies within its interior of $\widetilde{\Theta}$. In this extended parameter space we have

$$\sqrt{n}\begin{pmatrix} \widetilde{\theta}_1 - \theta_{10} \\ \widetilde{\theta}_2 - \theta_{20} \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \begin{pmatrix} I_{11}(\theta_0) & 0 \\ 0 & I_{22}(\theta_0) \end{pmatrix}^{-1}\right).$$

In order to derive the limiting distribution of the log-likelihood ratio statistic

$$W_n = 2\left(\max_{\theta \in \Theta} \mathcal{L}_n(\theta) - \mathcal{L}_n(\theta_0)\right)$$

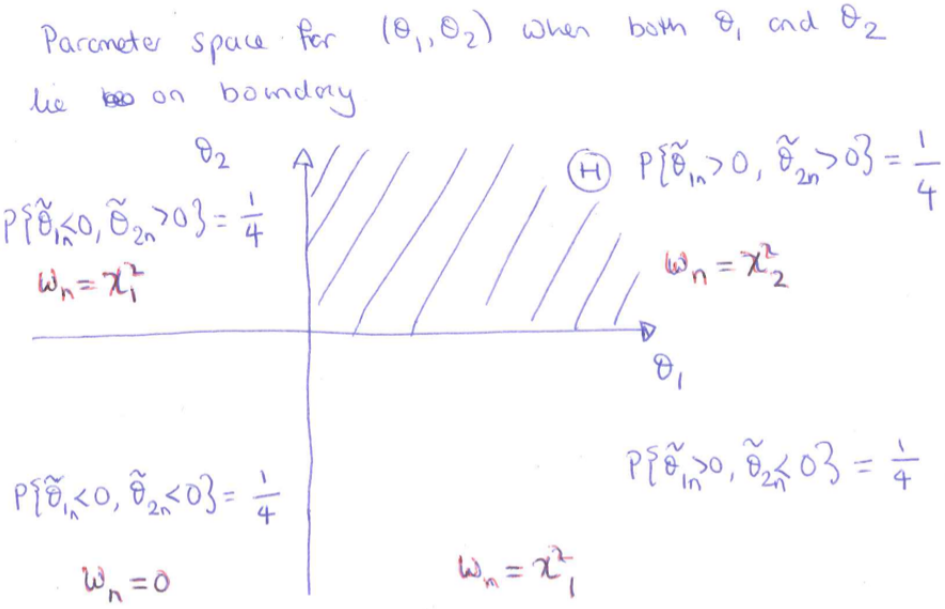

Figure 4.4: Two parameters: both parameters on boundary.

we condition on $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$. This gives

$$\begin{aligned} & P\left(W_n \leq x\right) \\ = \; & P\left(W_n \leq x | \widetilde{\theta}_1 \leq 0, \widetilde{\theta}_2 \leq 0\right) P\left(\widetilde{\theta}_1 \leq 0, \widetilde{\theta}_2 \leq 0\right) + P\left(W_n \leq x | \widetilde{\theta}_1 \leq 0, \widetilde{\theta}_2 > 0\right) P\left(\widetilde{\theta}_1 \leq 0, \widetilde{\theta}_2 > 0\right) + \\ & P\left(W_n \leq x | \widetilde{\theta}_1 > 0, \widetilde{\theta}_2 \leq 0\right) P\left(\widetilde{\theta}_1 > 0, \widetilde{\theta}_2 \leq 0\right) + P\left(W_n \leq x | \widetilde{\theta}_1 > 0, \widetilde{\theta}_2 > 0\right) P\left(\widetilde{\theta}_1 > 0, \widetilde{\theta}_2 > 0\right). \end{aligned}$$

Now by using the asymptotic independence of $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$ and for $\widetilde{\theta}_1 > 0, \widetilde{\theta}_2 > 0$ $W_n = 0$ the above is

$$P(W_n \leq x) = \frac{1}{4} + \frac{1}{2}\chi^2_1 + \frac{1}{4}\chi^2_2.$$

This is easiest seen in Figure 4.4.

Again the above argument can be made precise by using that the distribution of W_n can be approximated with the quadratic form

$$\begin{aligned} W_n & \stackrel{\mathcal{D}}{=} \max_{\theta_1 \in [0,\infty)} \left[-(Z_1 - \theta_1) I_{11}(0)(Z_1 - \theta_1)\right] + Z_1 I_{11}(0) Z_1 \\ & = + \max_{\theta_2 \in [0,\infty)} \left[-(Z_2 - \theta_2) I_{22}(0)(Z_2 - \theta_2)\right] + Z_2 I_{22}(0) Z_2 \end{aligned}$$

where $Z_1 \sim N(0, I_{11}(\theta_0)^{-1})$ and $Z_2 \sim N(0, I_{22}(\theta_0)^{-1})$. This approximation gives the same result.

4.2.4 Estimation on the boundary when the Fisher information is not block diagonal

In the case that the Fisher information matrix is not block diagonal the same procedure can be use, but the results are no longer so clean. In particular, the limiting distribution may no longer be a mixture of chi-square distributions and/or the weighting probabilities will depend on the parameter θ (thus the log-likelihood ratio will not be pivotal).

Let us consider the example where one parameter lies on the boundary and the other does not. i.e the parameter space is $[0, \infty) \times (-\infty, \infty)$. The true parameter $\theta_0 = (0, \theta_{20})$ however, unlike the examples considered above the Fisher information matrix is not diagonal. Let $\widehat{\theta}_n = (\widehat{\theta}_1, \widehat{\theta}_2) = \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta)$. We can use the conditioning arguments given above however they become ackward because of the dependence between the estimators of $\widehat{\theta}_1$ and $\widehat{\theta}_2$. Instead we use the quadratic form approximation

$$\begin{aligned} W_n &= 2\left(\max_{\theta \in \Theta} \mathcal{L}_n(\theta) - \mathcal{L}_n(\theta_0)\right) \\ &\stackrel{\mathcal{D}}{=} \max_{\theta \in \Theta} [-(Z-\theta) I(\theta_0)(Z-\theta)] + Z' I(\theta_0) Z + o_p(1) \end{aligned}$$

where $Z \sim \mathcal{N}(0, I(\theta_0)^{-1})$. To simplify the derivation we let $\overline{Z} \sim \mathcal{N}(0, I_2)$. Then the above can be written as

$$\begin{aligned} W_n &= 2\left(\max_{\theta \in \Theta} \mathcal{L}_n(\theta) - \mathcal{L}_n(\theta_0)\right) \\ &\stackrel{\mathcal{D}}{=} \max_{\theta \in \Theta} \left[-\left\{ I(\theta_0)^{-1/2}\overline{Z} - I(\theta_0)^{-1/2} I(\theta_0)^{1/2}\theta \right\}' I(\theta_0) \left\{ I(\theta_0)^{-1/2}\overline{Z} - I(\theta_0)^{-1/2} I(\theta_0)^{1/2}\theta \right\} \right] \\ &\quad + \left\{ I(\theta_0)^{-1/2}\overline{Z} \right\}' I(\theta_0) \left\{ I(\theta_0)^{-1/2}\overline{Z} \right\} + o_p(1) \\ &= \max_{\overline{\theta} \in \overline{\Theta}} \left[-(\overline{Z} - \overline{\theta})'(\overline{Z} - \overline{\theta}) \right] + \overline{Z}'\overline{Z} + o_p(1) \end{aligned}$$

where $\overline{\Theta} = \{\overline{\theta} = I(\theta_0)^{1/2}\theta; \theta \in \Theta\}$. This orthogonalisation simplifies the calculations. Using the spectral decomposition of $I(\theta) = P\Lambda P'$ where $P = (\underline{p}_1, \underline{p}_2)$ (thus $I(\theta)^{1/2} = P\Lambda^{1/2}P'$) we see that the half plane (which defines Θ) turns into the rotated half plane $\overline{\Theta}$ which is determined by the eigenvectors $\underline{p}_1$ and $\underline{p}_2$ (which rotates the line $\alpha(0,1)$ into

$$L = \alpha[\lambda_1^{1/2}\langle \underline{p}_1, (0,1)\rangle \underline{p}_1 + \lambda_2^{1/2}\langle \underline{p}_2, (0,1)\rangle \underline{p}_2] = \alpha[\lambda_1^{1/2}\langle \underline{p}_1, \underline{1}\rangle \underline{p}_1 + \lambda_2^{1/2}\langle \underline{p}_2, \underline{1}\rangle \underline{p}_2]$$

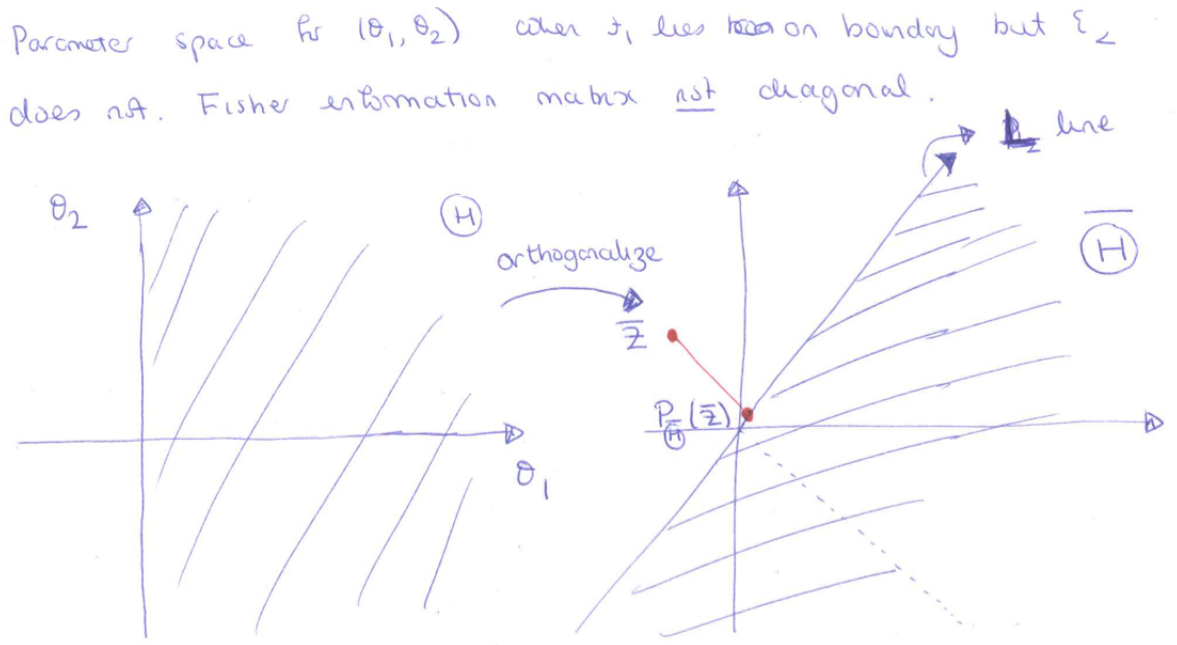

Figure 4.5: Two parameters: one parameter on boundary and the other in interior.

where $\underline{1} = (0, 1)$. We observe that

$$W_n = \begin{cases} \underbrace{\overline{Z}'\overline{Z}}_{\chi^2_2} & \overline{Z} \in \overline{\Theta} \\ -[\overline{Z} - P_{\overline{\Theta}}(\overline{Z})]'[\overline{Z} - P_{\overline{\Theta}}(\overline{Z})] + \overline{Z}'\overline{Z} & \overline{Z} \in \overline{\Theta}^c. \end{cases}$$

We note that $P_{\overline{\Theta}}(\overline{Z})$ is the nearest closest point on the line L, thus with some effort one can calculate the distribution of $-[\overline{Z}-P_{\overline{\Theta}}(\overline{Z})]'[\overline{Z}-P_{\overline{\Theta}}(\overline{Z})]+\overline{Z}'\overline{Z}$ (it will be some weighted chi-square), noting that $P(\overline{Z} \in \overline{\Theta}) = 1/2$ and $P(\overline{Z} \in \overline{\Theta}^c) = 1/2$ (since they are both in half a plane). Thus we observe that the above is a mixture of distributions, but they are not as simple (or useful) as when the information matrix has a block diagonal structure.

The precise details can be found in Chernoff (1954), Moran (1971), Chant (1974), Self and Liang (1987) and Andrews (1999). For the Bayesian case see, for example, Botchkina and Green (2014).

Exercise 4.2 *The parameter space of θ is $[0, \infty) \times [0, \infty)$. The Fisher information matrix corresponding to the distribution is*

$$I(\theta) = \begin{pmatrix} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) \end{pmatrix}.$$

Suppose that the true parameter is $\theta = (0, 0)$ obtain (to the best you can) the limiting distribution of the log-likelihood ratio statistic $2(\arg\max_{\theta\in\Theta} \mathcal{L}_n(\theta) - \mathcal{L}_n(0, 0))$.

4.3 Regularity conditions which are not satisfied

In this section we consider another aspect of nonstandard inference. Namely, deriving the asymptotic sampling properties of estimators (mainly MLEs) when the usual regularity conditions are not satisfied, thus the results in Chapter 2 do not hold. Some of this material was covered or touched on previously. Here, for completeness, we have collected the results together.

The uniform distribution

The standard example where the regularity conditions (mainly Assumption 1.3.1(ii)) are not satisfied is the uniform distribution

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

We can see that the likelihood in this case is

$$L_n(\underline{X};\theta) = \prod_{i=1}^{n} \theta^{-1} I(0 < X_i < \theta).$$

In this case the the derivative of $L_n(\underline{X};\theta)$ is not well defined, hence we cannot solve for the derivative. Instead, to obtain the mle we try to reason what the maximum is. We should plot $L_n(\underline{X};\theta)$ against θ and place X_i on the θ axis. We can see that if $\theta < X_i$, then L_n is zero. Let $X_{(i)}$ denote the ordered data $X_{(1)} \leq X_{(2)}, \ldots \leq X_{(T)}$. We see that for $\theta = X_{(T)}$, we have $L_n(\underline{X};\theta) = (X_{(T)})^{-T}$, then beyond this point $L_n(\underline{X};\theta)$ decays ie. $L_n(\underline{X};\theta) = \theta^{-T}$ for $\theta \geq X_{(T)}$. Hence the maximum of the likelihood is $\hat{\theta}_n = \max_{1\leq t\leq T} X_i$. The sampling properties of $\widehat{\theta}_n$ were calculated in Exercise 2.3.

The shifted exponential

Let us consider the shifted exponential distribution

$$f(x;\theta,\phi) = \frac{1}{\theta}\exp\left(-\frac{(x-\phi)}{\theta}\right) \quad x \geq \phi,$$

which is only well defined for $\theta, \phi > 0$. We first observe when $\phi = 0$ we have the usual exponential function, ϕ is simply a shift parameter. It can be shown that the usual regularity conditions (Assumption 1.3.1) will not be satisfied. This means the Cramer-Rao bound does not hold in this case and the limiting variance of the mle estimators will not be the inverse of the Fisher information matrix.

The likelihood for this example is

$$L_n(\underline{X}; \theta, \phi) = \frac{1}{\theta^n} \prod_{i=1}^{n} \exp\left(-\frac{(X_i - \phi)}{\theta}\right) I(\phi \leq X_i).$$

We see that we cannot obtain the maximum of $L_n(\underline{X}; \theta, \phi)$ by differentiating. Instead let us consider what happens to $L_n(\underline{X}; \theta, \phi)$ for different values of ϕ. We see that for $\phi > X_i$ for any t, the likelihood is zero. But at $\phi = X_{(1)}$ (smallest value), the likelihood is $\frac{1}{\theta^n} \prod_{i=1}^{n} \exp(-\frac{(X_{(t)} - X_{(1)})}{\theta}$. But for $\phi < X_{(1)}$, $L_n(\underline{X}; \theta, \phi)$ starts to decrease because $(X_{(t)} - \phi) > (X_{(t)} - X_{(1)})$, hence the likelihood decreases. Thus the MLE for ϕ is $\widehat{\phi}_n = X_{(1)}$, notice that this estimator is completely independent of θ. To obtain the mle of θ, differentiate and solve $\frac{\partial L_n(\underline{X}; \theta, \phi)}{d\theta} \rfloor_{\widehat{\phi}_n = X_{(1)}} = 0$. We obtain $\widehat{\theta}_n = \bar{X} - \widehat{\phi}_n$. For a reality check, we recall that when $\phi = 0$ then the MLE of θ is $\widehat{\theta}_n = \bar{X}$.

We now derive the distribution of $\widehat{\phi}_n - \phi = X_{(1)} - \phi$ (in this case we can actually obtain the finite sample distribution). To make the calculation easier we observe that X_i can be rewritten as $X_i = \phi + E_i$, where $\{E_i\}$ are iid random variables with the standard exponential distribution starting at zero: $f(x; \theta, 0) = \theta^{-1} \exp(-x/\theta)$. Therefore the distribution function of $\widehat{\phi}_n - \phi = \min_i E_i$

$$\begin{aligned} P(\widehat{\phi}_n - \phi \leq x) &= P(\min_i(E_i) \leq x) = 1 - P(\min_i(E_i) > x) \\ &= 1 - [\exp(-x/\theta)]^n. \end{aligned}$$

Therefore the density of $\widehat{\phi}_n - \phi$ is $\frac{\theta}{n} \exp(-nx/\theta)$, which is an exponential with parameter n/θ. Using this, we observe that the mean of $\widehat{\phi}_n - \phi$ is θ/n and the variance is θ^2/n^2. In this case when we standardize $(\widehat{\phi}_n - \phi)$ we need to do so with n (and not the classical $\sqrt{n}$). When we do this we observe that the distribution of $n(\widehat{\phi}_n - \phi)$ is exponential with parameter θ^{-1} (since the sum of n iid exponentials with parameter θ^{-1} is exponential with parameter $n\theta^{-1}$).

In summary, we observe that $\widehat{\phi}_n$ is a biased estimator of ϕ, but the bias decreases as $n \to \infty$. Morover, the variance is quite amazing. Unlike standard estimators where the variance decreases at the rate $1/n$, the variance of $\widehat{\phi}_n$ decreases at the rate $1/n^2$.

Even though $\widehat{\phi}_n$ behaves in a nonstandard way, the estimator $\widehat{\theta}_n$ is completely standard. If ϕ were known then the regularity conditions are satisfied. Furthermore, since $[\widehat{\phi}_n - \phi] = O_p(n^{-1})$ then the difference between the likelihoods with known and estimated ϕ are almost the same; i.e. $\mathcal{L}_n(\theta, \phi) \approx \mathcal{L}_n(\theta, \widehat{\phi}_n)$. Therefore the sampling properties of $\widehat{\theta}_n$ are asymptotically equivalent to the sampling properties of the MLE if ϕ were known.

See Davison (2002), page 145, example 4.43 for more details.

Note that in many problems in inference one replaces the observed likelihood with the unobserved likelihood and show that the difference is "asymptotically negligible". If this can be shown then the sampling properties of estimators involving the observed and unobserved likelihoods are asymptotically equivalent.

Example 4.3.1 *Let us suppose that $\{X_i\}$ are iid exponentially distributed random variables with density $f(x) = \frac{1}{\lambda}\exp(-x/\lambda)$. Suppose that we only observe $\{X_i\}$, if $X_i > c$ (else X_i is not observed).*

(i) Show that the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is a biased estimator of λ.

(ii) Suppose that λ and c are unknown, obtain the log-likelihood of $\{X_i\}_{i=1}^n$ and the maximum likelihood estimators of λ and c.

Solution

(i) It is easy to see that $\mathrm{E}(\bar{X}) = \mathrm{E}(X_i|X_i > c)$, thus

$$\begin{aligned}
\mathrm{E}(X_i|X_i > c) &= \int_0^\infty x\frac{f(x)I(X \geq c)}{P(X > c)}dx \\
&= \int_c^\infty x\frac{f(x)I(X \geq c)}{P(X > c)}dx = \frac{1}{e^{-c/\lambda}}\int_c^\infty xf(x)dx \\
&= \frac{\lambda e^{-c/\lambda}(\frac{c}{\lambda} + 1)}{e^{-c/\lambda}} = \lambda + c.
\end{aligned}$$

Thus $\mathrm{E}(\bar{X}) = \lambda + c$ and not the desired λ.

(ii) We observe that the density of X_i given $X_i > c$ is $f(x|X_i > c) = \frac{f(x)I(X>c)}{P(X>c)} = \lambda^{-1}\exp(-1/\lambda(X-c))I(X \geq c)$; this is close to a shifted exponential and the density does not satisfy the regularity conditions.

Based on this the log-likelihood $\{X_i\}$ is

$$\begin{aligned}\mathcal{L}_n(\lambda) &= \sum_{i=1}^{n} \big\{ \log f(X_i) + \log I(X_i \geq c) - \log P(X_i > c) \big\} \\ &= \sum_{i=1}^{n} \big\{ -\log \lambda - \frac{1}{\lambda}(X_i - c) + \log I(X_i \geq c) \big\}.\end{aligned}$$

Hence we want to find the λ and c which maximises the above. Here we can use the idea of profiling to estimate the parameters - it does not matter which parameter we profile out. Suppose we fix, λ, and maximise the above with respect to c, in this case it is easier to maximise the actual likelihood:

$$L_\lambda(c) = \prod_{i=1}^{n} \frac{1}{\lambda} \exp(-(X_i - c)/\lambda) I(X_i > c).$$

By drawing L with respect to c, we can see that it is maximum at $\min X_{(i)}$ (for all λ), thus the MLE of c is $\widehat{c} = \min_i X_i$. Now we can estimate λ. Putting $\widehat{c}$ back into the log-likelihood gives

$$\sum_{i=1}^{n} \big\{ -\log \lambda - \frac{1}{\lambda}(X_i - \widehat{c}) + \log I(X_i \geq \widehat{c}) \big\}.$$

Differentiating the above with respect to λ gives $\sum_{i=1}^{n}(X_i - \widehat{c}) = \lambda n$. Thus $\widehat{\lambda}_n = \frac{1}{n}\sum_{i=1}^{n} X_i - \widehat{c}$. Thus $\widehat{c} = \min_i X_i$ $\widehat{\lambda}_n = \frac{1}{n}\sum_{i=1}^{n} X_i - \widehat{c}_n$, are the MLE estimators of c and λ respectively.