Chapter 6
Theory of Estimation

We know from our basic knowledge of statistics that one of the objectives in statistics is to better understand and model the underlying process which generates data. This is known as statistical inference: we infer from information contained in sample properties of the population from which the observations are taken. In multivariate statistical inference, we do exactly the same. The basic ideas were introduced in Sect. 4.5 on sampling theory: we observed the values of a multivariate random variable X and obtained a sample $\mathcal{X} = \{x_i\}_{i=1}^n$. Under random sampling, these observations are considered to be realisations of a sequence of i.i.d. random variables $X_1, \ldots, X_n$ where each X_i is a p-variate random variable which replicates the *parent* or *population* random variable X. In this chapter, for notational convenience, we will no longer differentiate between a random variable X_i and an observation of it, x_i, in our notation. We will simply write x_i and it should be clear from the context whether a random variable or an observed value is meant.

Statistical inference infers from the i.i.d. random sample $\mathcal{X}$ the properties of the population: typically, some unknown characteristic θ of its distribution. In parametric statistics, θ is a k-variate vector $\theta \in \mathbb{R}^k$ characterising the unknown properties of the population pdf $f(x;\theta)$: this could be the mean, the covariance matrix, kurtosis, etc.

The aim will be to estimate θ from the sample $\mathcal{X}$ through estimators $\hat{\theta}$ which are functions of the sample: $\hat{\theta} = \hat{\theta}(\mathcal{X})$. When an estimator $\hat{\theta}$ is proposed, we must derive its sampling distribution to analyse its properties.

In this chapter the basic theoretical tools are developed which are needed to derive estimators and to determine their properties in general situations. We will basically rely on the maximum likelihood theory in our presentation. In many situations, the maximum likelihood estimators (MLEs) indeed share asymptotic optimal properties which make their use easy and appealing.

We will illustrate the multivariate normal population and also the linear regression model where the applications are numerous and the derivations are easy to do. In multivariate setups, the MLE is at times too complicated to be derived

W.K. Härdle, L. Simar, *Applied Multivariate Statistical Analysis*,
DOI 10.1007/978-3-662-45171-7_6

analytically. In such cases, the estimators are obtained using numerical methods (nonlinear optimisation). The general theory and the asymptotic properties of these estimators remain simple and valid. The following Chap. 7 concentrates on hypothesis testing and confidence interval issues.

6.1 The Likelihood Function

Suppose that $\{x_i\}_{i=1}^n$ is an i.i.d. sample from a population with pdf $f(x;\theta)$. The aim is to estimate $\theta \in \mathbb{R}^k$ which is a vector of unknown parameters. The *likelihood function* is defined as the joint density $L(\mathcal{X};\theta)$ of the observations x_i considered as a function of θ:

$$L(\mathcal{X};\theta) = \prod_{i=1}^{n} f(x_i;\theta), \tag{6.1}$$

where $\mathcal{X}$ denotes the sample of the data matrix with the observations $x_1^\top, \ldots, x_n^\top$ in each row. The MLE of θ is defined as

$$\hat{\theta} = \arg\max_{\theta} L(\mathcal{X};\theta).$$

Often it is easier to maximise the *log-likelihood function*

$$\ell(\mathcal{X};\theta) = \log L(\mathcal{X};\theta), \tag{6.2}$$

which is equivalent since the logarithm is a monotone one-to-one function. Hence

$$\hat{\theta} = \arg\max_{\theta} L(\mathcal{X};\theta) = \arg\max_{\theta} \ell(\mathcal{X};\theta).$$

The following examples illustrate cases where the maximisation process can be performed analytically, i.e., we will obtain an explicit analytical expression for $\hat{\theta}$. Unfortunately, in other situations, the maximisation process can be more intricate, involving nonlinear optimisation techniques. In the latter case, given a sample $\mathcal{X}$ and the likelihood function, numerical methods will be used to determine the value of θ maximising $L(\mathcal{X};\theta)$ or $\ell(\mathcal{X};\theta)$. These numerical methods are typically based on Newton–Raphson iterative techniques.

Example 6.1 Consider a sample $\{x_i\}_{i=1}^n$ from $N_p(\mu, \mathcal{I})$, i.e., from the pdf

$$f(x;\theta) = (2\pi)^{-p/2} \exp\left\{-\frac{1}{2}(x-\theta)^\top (x-\theta)\right\},$$

where $\theta = \mu \in \mathbb{R}^p$ is the mean vector parameter. The log-likelihood is in this case

$$\ell(\mathcal{X};\theta) = \sum_{i=1}^{n} \log\{f(x_i;\theta)\} = \log(2\pi)^{-np/2} - \frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^\top(x_i-\theta). \tag{6.3}$$

The term $(x_i-\theta)^\top(x_i-\theta)$ equals

$$(x_i-\overline{x})^\top(x_i-\overline{x}) + (\overline{x}-\theta)^\top(\overline{x}-\theta) + 2(\overline{x}-\theta)^\top(x_i-\overline{x}).$$

Summing this term over $i = 1,\ldots,n$ we see that

$$\sum_{i=1}^{n}(x_i-\theta)^\top(x_i-\theta) = \sum_{i=1}^{n}(x_i-\overline{x})^\top(x_i-\overline{x}) + n(\overline{x}-\theta)^\top(\overline{x}-\theta).$$

Hence

$$\ell(\mathcal{X};\theta) = \log(2\pi)^{-np/2} - \frac{1}{2}\sum_{i=1}^{n}(x_i-\overline{x})^\top(x_i-\overline{x}) - \frac{n}{2}(\overline{x}-\theta)^\top(\overline{x}-\theta).$$

Only the last term depends on θ and is obviously maximised for

$$\hat{\theta} = \hat{\mu} = \overline{x}.$$

Thus $\overline{x}$ is the MLE of θ for this family of pdfs $f(x,\theta)$.

A more complex example is the following one where we derive the MLEs for μ and Σ.

Example 6.2 Suppose $\{x_i\}_{i=1}^n$ is a sample from a normal distribution $N_p(\mu,\Sigma)$. Here $\theta = (\mu,\Sigma)$ with Σ interpreted as a vector. Due to the symmetry of Σ the unknown parameter θ is in fact $\{p + \frac{1}{2}p(p+1)\}$-dimensional. Then

$$L(\mathcal{X};\theta) = |2\pi\Sigma|^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^\top\Sigma^{-1}(x_i-\mu)\right\} \tag{6.4}$$

and

$$\ell(\mathcal{X};\theta) = -\frac{n}{2}\log|2\pi\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^\top\Sigma^{-1}(x_i-\mu). \tag{6.5}$$

The term $(x_i-\mu)^\top\Sigma^{-1}(x_i-\mu)$ equals

$$(x_i-\overline{x})^\top\Sigma^{-1}(x_i-\overline{x}) + (\overline{x}-\mu)^\top\Sigma^{-1}(\overline{x}-\mu) + 2(\overline{x}-\mu)^\top\Sigma^{-1}(x_i-\overline{x}).$$

Summing this term over $i = 1, \ldots, n$ we see that

$$\sum_{i=1}^{n}(x_i-\mu)^\top\Sigma^{-1}(x_i-\mu) = \sum_{i=1}^{n}(x_i-\overline{x})^\top\Sigma^{-1}(x_i-\overline{x}) + n(\overline{x}-\mu)^\top\Sigma^{-1}(\overline{x}-\mu).$$

Note that from (2.14)

$$\begin{aligned}(x_i-\overline{x})^\top\Sigma^{-1}(x_i-\overline{x}) &= \operatorname{tr}\left\{(x_i-\overline{x})^\top\Sigma^{-1}(x_i-\overline{x})\right\}\\ &= \operatorname{tr}\left\{\Sigma^{-1}(x_i-\overline{x})(x_i-\overline{x})^\top\right\}.\end{aligned}$$

Therefore, by summing over the index i we finally arrive at

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\mu)^\top\Sigma^{-1}(x_i-\mu) &= \operatorname{tr}\left\{\Sigma^{-1}\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})^\top\right\}\\ &\quad +n(\overline{x}-\mu)^\top\Sigma^{-1}(\overline{x}-\mu)\\ &= \operatorname{tr}\{\Sigma^{-1}n\mathcal{S}\} + n(\overline{x}-\mu)^\top\Sigma^{-1}(\overline{x}-\mu).\end{aligned}$$

Thus the log-likelihood function for $N_p(\mu, \Sigma)$ is

$$\ell(\mathcal{X};\theta) = -\frac{n}{2}\log|2\pi\Sigma| - \frac{n}{2}\operatorname{tr}\{\Sigma^{-1}\mathcal{S}\} - \frac{n}{2}(\overline{x}-\mu)^\top\Sigma^{-1}(\overline{x}-\mu). \tag{6.6}$$

We can easily see that the third term is maximised by $\mu = \bar{x}$. In fact the MLEs are given by

$$\hat{\mu} = \overline{x}, \quad \hat{\Sigma} = \mathcal{S}.$$

The derivation of $\hat{\Sigma}$ is a lot more complicated. It involves derivatives with respect to matrices with their notational complexities and will not be presented here; for more elaborate proof, see Mardia, Kent and Bibby (1979, pp. 103–104). Note that the unbiased covariance estimator $\mathcal{S}_u = \frac{n}{n-1}\mathcal{S}$ is not the MLE of Σ!

Example 6.3 Consider the linear regression model $y_i = \beta^\top x_i + \varepsilon_i$ for $i = 1, \ldots, n$, where ε_i is i.i.d. and $N(0, \sigma^2)$ and where $x_i \in \mathbb{R}^p$. Here $\theta = (\beta^\top, \sigma)$ is a $(p+1)$-dimensional parameter vector. Denote

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathcal{X} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix}.$$

Then

$$L(y, \mathcal{X}; \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta^\top x_i)^2\right\}$$

and

$$\begin{aligned}\ell(y, \mathcal{X}; \theta) &= \log\left\{\frac{1}{(2\pi)^{n/2}\sigma^n}\right\} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta^\top x_i)^2 \\ &= -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}(y - \mathcal{X}\beta)^\top(y - \mathcal{X}\beta) \\ &= -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}(y^\top y + \beta^\top\mathcal{X}^\top\mathcal{X}\beta - 2\beta^\top\mathcal{X}^\top y).\end{aligned}$$

Differentiating w.r.t. the parameters yields

$$\begin{aligned}\frac{\partial}{\partial\beta}\ell &= -\frac{1}{2\sigma^2}(2\mathcal{X}^\top\mathcal{X}\beta - 2\mathcal{X}^\top y) \\ \frac{\partial}{\partial\sigma}\ell &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\{(y - \mathcal{X}\beta)^\top(y - \mathcal{X}\beta)\}.\end{aligned}$$

Note that $\frac{\partial}{\partial\beta}\ell$ denotes the vector of the derivatives w.r.t. all components of β (the gradient). Since the first equation only depends on β, we start with deriving $\hat{\beta}$.

$$\mathcal{X}^\top\mathcal{X}\hat{\beta} = \mathcal{X}^\top y, \quad \text{hence} \quad \hat{\beta} = (\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{X}^\top y$$

Plugging $\hat{\beta}$ into the second equation gives

$$\frac{n}{\hat{\sigma}} = \frac{1}{\hat{\sigma}^3}(y - \mathcal{X}\hat{\beta})^\top(y - \mathcal{X}\hat{\beta}), \quad \text{hence} \quad \hat{\sigma}^2 = \frac{1}{n}||y - \mathcal{X}\hat{\beta}||^2,$$

where $||\cdot||^2$ denotes the Euclidean vector norm from Sect. 2.6. We see that the MLE $\hat{\beta}$ is identical with the least squares estimator (3.52). The variance estimator

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}^\top x_i)^2$$

is nothing else than the residual sum of squares (RSS) from (3.37) generalised to the case of multivariate x_i. Note that when the x_i are considered to be fixed, we have

$$\mathsf{E}(y) = \mathcal{X}\beta \text{ and } \mathsf{Var}(y) = \sigma^2\mathcal{I}_n.$$

Then, using the properties of moments from Sect. 4.2 we have

$$\mathsf{E}(\hat{\beta}) = (\mathcal{X}^\top \mathcal{X})^{-1} \mathcal{X}^\top \mathsf{E}(y) = \beta, \tag{6.7}$$

$$\mathsf{Var}(\hat{\beta}) = \sigma^2 (\mathcal{X}^\top \mathcal{X})^{-1}. \tag{6.8}$$

Summary

- If $\{x_i\}_{i=1}^n$ is an i.i.d. sample from a distribution with pdf $f(x;\theta)$, then $L(\mathcal{X};\theta) = \prod_{i=1}^n f(x_i;\theta)$ is the likelihood function. The MLE is that value of θ which maximises $L(\mathcal{X};\theta)$. Equivalently one can maximise the log-likelihood $\ell(\mathcal{X};\theta)$.
- The MLEs of μ and Σ from a $N_p(\mu,\Sigma)$ distribution are $\hat{\mu} = \overline{x}$ and $\hat{\Sigma} = \mathcal{S}$. Note that the MLE of Σ is not unbiased.
- The MLEs of β and σ in the linear model $y = \mathcal{X}\beta + \varepsilon$, $\varepsilon \sim N_n(0,\sigma^2\mathcal{I})$ are given by the least squares estimator $\hat{\beta} = (\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{X}^\top y$ and $\hat{\sigma}^2 = \frac{1}{n}||y - \mathcal{X}\hat{\beta}||^2$. $\mathsf{E}(\hat{\beta}) = \beta$ and $\mathsf{Var}(\hat{\beta}) = \sigma^2(\mathcal{X}^\top\mathcal{X})^{-1}$.

6.2 The Cramer–Rao Lower Bound

As pointed out above, an important question in estimation theory is whether an estimator $\hat{\theta}$ has certain desired properties, in particular, if it converges to the unknown parameter θ it is supposed to estimate. One typical property we want for an estimator is unbiasedness, meaning that on the average, the estimator hits its target: $\mathsf{E}(\hat{\theta}) = \theta$. We have seen for instance (see Example 6.2) that $\overline{x}$ is an unbiased estimator of μ and $\mathcal{S}$ is a biased estimator of Σ in finite samples. If we restrict ourselves to unbiased estimation, then the natural question is whether the estimator shares some optimality properties in terms of its sampling variance. Since we focus on unbiasedness, we look for an estimator with the smallest possible variance.

In this context, the Cramer–Rao lower bound will give the minimal achievable variance for any unbiased estimator. This result is valid under very general regularity conditions (discussed below). One of the most important applications of the Cramer–Rao lower bound is that it provides the asymptotic optimality property of MLEs. The Cramer–Rao theorem involves the *score function* and its properties which will be derived first.

The score function $s(\mathcal{X};\theta)$ is the derivative of the log likelihood function w.r.t. $\theta \in \mathbb{R}^k$

$$s(\mathcal{X};\theta) = \frac{\partial}{\partial\theta}\ell(\mathcal{X};\theta) = \frac{1}{L(\mathcal{X};\theta)}\frac{\partial}{\partial\theta}L(\mathcal{X};\theta). \tag{6.9}$$

The covariance matrix $\mathcal{F}_n = \mathsf{Var}\{s(\mathcal{X};\theta)\}$ is called the *Fisher information matrix*. In what follows, we will give some interesting properties of score functions.

Theorem 6.1 *If $s = s(\mathcal{X};\theta)$ is the score function and if $\hat{\theta} = t = t(\mathcal{X},\theta)$ is any function of $\mathcal{X}$ and θ, then under regularity conditions*

$$\mathsf{E}(st^\top) = \frac{\partial}{\partial\theta}\,\mathsf{E}(t^\top) - \mathsf{E}\left(\frac{\partial t^\top}{\partial\theta}\right). \tag{6.10}$$

The proof is left as an exercise (see Exercise 6.9). The regularity conditions required for this theorem are rather technical and ensure that the expressions (expectations and derivations) appearing in (6.10) are well defined. In particular, the support of the density $f(x;\theta)$ should not depend on θ. The next corollary is a direct consequence.

Corollary 6.1 *If $s = s(\mathcal{X};\theta)$ is the score function, and $\hat{\theta} = t = t(\mathcal{X})$ is any unbiased estimator of θ (i.e., $\mathsf{E}(t) = \theta$), then*

$$\mathsf{E}(st^\top) = \mathsf{Cov}(s,t) = \mathcal{I}_k. \tag{6.11}$$

Note that the score function has mean zero (see Exercise 6.10).

$$\mathsf{E}\{s(\mathcal{X};\theta)\} = 0. \tag{6.12}$$

Hence, $\mathsf{E}(ss^\top) = \mathsf{Var}(s) = \mathcal{F}_n$ and by setting $s = t$ in Theorem 6.1 it follows that

$$\mathcal{F}_n = -\,\mathsf{E}\left\{\frac{\partial^2}{\partial\theta\partial\theta^\top}\ell(\mathcal{X};\theta)\right\}.$$

Remark 6.1 If $x_1,\ldots,x_n$ are i.i.d., $\mathcal{F}_n = n\mathcal{F}_1$ where $\mathcal{F}_1$ is the Fisher information matrix for sample size $n = 1$.

Example 6.4 Consider an i.i.d. sample $\{x_i\}_{i=1}^n$ from $N_p(\theta,\mathcal{I})$. In this case the parameter θ is the mean μ. It follows from (6.3) that

$$\begin{aligned} s(\mathcal{X};\theta) &= \frac{\partial}{\partial\theta}\,\ell(\mathcal{X};\theta) \\ &= -\frac{1}{2}\frac{\partial}{\partial\theta}\left\{\sum_{i=1}^n (x_i-\theta)^\top(x_i-\theta)\right\} \\ &= n(\overline{x}-\theta). \end{aligned}$$

Hence, the information matrix is

$$\mathcal{F}_n = \mathsf{Var}\{n(\overline{x} - \theta)\} = n\mathcal{I}_p.$$

How well can we estimate θ? The answer is given in the following theorem which is from Cramer and Rao. As pointed out above, this theorem gives a lower bound for unbiased estimators. Hence, all estimators, which are unbiased **and** attain this lower bound, are *minimum variance estimators*.

Theorem 6.2 (Cramer–Rao) *If $\hat{\theta} = t = t(\mathcal{X})$ is any unbiased estimator for θ, then under regularity conditions*

$$\mathsf{Var}(t) \geq \mathcal{F}_n^{-1}, \tag{6.13}$$

where

$$\mathcal{F}_n = \mathsf{E}\{s(\mathcal{X};\theta)s(\mathcal{X};\theta)^\top\} = \mathsf{Var}\{s(\mathcal{X};\theta)\} \tag{6.14}$$

is the Fisher information matrix.

Proof Consider the correlation $\rho_{Y,Z}$ between Y and Z where $Y = a^\top t$, $Z = c^\top s$. Here s is the score and the vectors $a, c \in \mathbb{R}^p$. By Corollary 6.1 $\mathsf{Cov}(s,t) = \mathcal{I}$ and thus

$$\begin{aligned} \mathsf{Cov}(Y,Z) &= a^\top \mathsf{Cov}(t,s)c = a^\top c \\ \mathsf{Var}(Z) &= c^\top \mathsf{Var}(s)c = c^\top \mathcal{F}_n c. \end{aligned}$$

Hence,

$$\rho^2_{Y,Z} = \frac{\mathsf{Cov}^2(Y,Z)}{\mathsf{Var}(Y)\,\mathsf{Var}(Z)} = \frac{(a^\top c)^2}{a^\top \mathsf{Var}(t)a \cdot c^\top \mathcal{F}_n c} \leq 1. \tag{6.15}$$

In particular, this holds for any $c \neq 0$. Therefore it holds also for the maximum of the left-hand side of (6.15) with respect to c. Since

$$\max_c \frac{c^\top aa^\top c}{c^\top \mathcal{F}_n c} = \max_{c^\top \mathcal{F}_n c = 1} c^\top aa^\top c$$

and

$$\max_{c^\top \mathcal{F}_n c = 1} c^\top aa^\top c = a^\top \mathcal{F}_n^{-1} a$$

by our maximisation Theorem 2.5, we have

$$\frac{a^\top \mathcal{F}_n^{-1} a}{a^\top \mathsf{Var}(t) a} \leq 1 \quad \forall\, a \in \mathbb{R}^p, \quad a \neq 0,$$

i.e.,

$$a^\top \{\mathsf{Var}(t) - \mathcal{F}_n^{-1}\} a \geq 0 \quad \forall\, a \in \mathbb{R}^p, \quad a \neq 0,$$

which is equivalent to $\mathsf{Var}(t) \geq \mathcal{F}_n^{-1}$. □

MLEs attain the lower bound if the sample size n goes to infinity. The next Theorem 6.3 states this and, in addition, gives the asymptotic sampling distribution of the maximum likelihood estimation, which turns out to be multinormal.

Theorem 6.3 *Suppose that the sample $\{x_i\}_{i=1}^n$ is i.i.d. If $\hat{\theta}$ is the MLE for $\theta \in \mathbb{R}^k$, i.e., $\hat{\theta} = \arg\max_\theta L(\mathcal{X}; \theta)$, then under some regularity conditions, as $n \to \infty$:*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} N_k(0, \mathcal{F}_1^{-1}) \tag{6.16}$$

where $\mathcal{F}_1$ denotes the Fisher information for sample size $n = 1$.

As a consequence of Theorem 6.3 we see that under regularity conditions the MLE is asymptotically unbiased, efficient (minimum variance) and normally distributed. Also it is a consistent estimator of θ.

Note that from property (5.4) of the multinormal it follows that asymptotically

$$n(\hat{\theta} - \theta)^\top \mathcal{F}_1 (\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \chi_p^2. \tag{6.17}$$

If $\hat{\mathcal{F}}_1$ is a consistent estimator of $\mathcal{F}_1$ (e.g. $\hat{\mathcal{F}}_1 = \mathcal{F}_1(\hat{\theta})$), we have equivalently

$$n(\hat{\theta} - \theta)^\top \hat{\mathcal{F}}_1 (\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \chi_p^2. \tag{6.18}$$

This expression is sometimes useful in testing hypotheses about θ and in constructing confidence regions for θ in a very general setup. These issues will be raised in more detail in the next chapter, but from (6.18) it can be seen, for instance, that when n is large,

$$\mathrm{P}\left\{ n(\hat{\theta} - \theta)^\top \hat{\mathcal{F}}_1 (\hat{\theta} - \theta) \leq \chi^2_{1-\alpha;p} \right\} \approx 1 - \alpha,$$

where $\chi^2_{\nu;p}$ denotes the ν-quantile of a χ_p^2 random variable. So, the ellipsoid $n(\hat{\theta} - \theta)^\top \hat{\mathcal{F}}_1 (\hat{\theta} - \theta) \leq \chi^2_{1-\alpha;p}$ provides in $\mathbb{R}^p$ an asymptotic $(1-\alpha)$-confidence region for θ.

Summary

↪ The score function is the derivative $s(\mathcal{X};\theta) = \frac{\partial}{\partial\theta}\ell(\mathcal{X};\theta)$ of the log-likelihood with respect to θ. The covariance matrix of $s(\mathcal{X};\theta)$ is the Fisher information matrix.

↪ The score function has mean zero: $\mathsf{E}\{s(\mathcal{X};\theta)\} = 0$.

↪ The Cramer–Rao bound says that any unbiased estimator $\hat{\theta} = t = t(\mathcal{X})$ has a variance that is bounded from below by the inverse of the Fisher information. Thus, an unbiased estimator, which attains this lower bound, is a minimum variance estimator.

↪ For i.i.d. data $\{x_i\}_{i=1}^n$ the Fisher information matrix is: $\mathcal{F}_n = n\mathcal{F}_1$.

↪ MLEs attain the lower bound in an asymptotic sense, i.e.,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} N_k(0, \mathcal{F}_1^{-1})$$

if $\hat{\theta}$ is the MLE for $\theta \in \mathbb{R}^k$, i.e., $\hat{\theta} = \arg\max_{\theta} L(\mathcal{X};\theta)$.

6.3 Exercises

Exercise 6.1 *Consider a uniform distribution on the interval* $[0,\theta]$. *What is the MLE of* θ? *(Hint: the maximisation here cannot be performed by means of derivatives. Here the support of* x *depends on* θ.*)*

Exercise 6.2 *Consider an i.i.d. sample of size n from the bivariate population with pdf* $f(x_1,x_2) = (\theta_1\theta_2)^{-1}\exp(-x_1/\theta_1 - x_2/\theta_2)$, $x_1,x_2 > 0$. *Compute the MLE of* $\theta = (\theta_1,\theta_2)$. *Find the Cramer–Rao lower bound. Is it possible to derive a minimal variance unbiased estimator of* θ?

Exercise 6.3 *Show that the MLE of Example 6.1,* $\hat{\mu} = \overline{x}$, *is a minimal variance estimator for any finite sample size n (i.e., without applying Theorem 6.3).*

Exercise 6.4 *We know from Example 6.4 that the MLE of Example 6.1 has* $\mathcal{F}_1 = \mathcal{I}_p$. *This leads to*

$$\sqrt{n}(\overline{x} - \mu) \xrightarrow{\mathcal{L}} N_p(0, \mathcal{I})$$

by Theorem 6.3. Can you give an analogous result for the square $\overline{x}^2$ *for the case* $p = 1$?

Exercise 6.5 *Consider an i.i.d. sample of size n from the bivariate population with pdf* $f(x_1, x_2) = (\theta_1^2\theta_2 x_2)^{-1} \exp(-x_1/\theta_1 x_2 - x_2/\theta_1\theta_2)$, $x_1, x_2 > 0$. *Compute the MLE of* $\theta = (\theta_1, \theta_2)$. *Find the Cramer–Rao lower bound and the asymptotic variance of* $\hat{\theta}$.

Exercise 6.6 *Consider a sample* $\{x_i\}_{i=1}^n$ *from* $N_p(\mu, \Sigma_0)$ *where* Σ_0 *is known. Compute the Cramer–Rao lower bound for* μ. *Can you derive a minimal unbiased estimator for* μ*?*

Exercise 6.7 *Let* $X \sim N_p(\mu, \Sigma)$ *where* Σ *is unknown but we know* $\Sigma = \text{diag}(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{pp})$. *From an i.i.d. sample of size n, find the MLE of* μ *and of* Σ.

Exercise 6.8 *Reconsider the setup of the previous exercise. Suppose that*

$$\Sigma = \text{diag}(\sigma_{11}, \sigma_{22}, \ldots, \sigma_{pp}).$$

Can you derive in this case the Cramer–Rao lower bound for $\theta^\top = (\mu_1 \ldots \mu_p, \sigma_{11} \ldots \sigma_{pp})$*?*

Exercise 6.9 *Prove Theorem 6.1. Hint: start from* $\frac{\partial}{\partial\theta}\mathsf{E}(t^\top) = \frac{\partial}{\partial\theta}\int t^\top(\mathcal{X};\theta) L(\mathcal{X};\theta)d\mathcal{X}$, *then permute integral and derivatives and note that* $s(\mathcal{X};\theta) = \frac{1}{L(\mathcal{X};\theta)}\frac{\partial}{\partial\theta}L(\mathcal{X};\theta)$.

Exercise 6.10 *Prove expression (6.12).*
(Hint: start from $\mathsf{E}\{s(\mathcal{X};\theta)\} = \int \frac{1}{L(\mathcal{X};\theta)}\frac{\partial}{\partial\theta}L(\mathcal{X};\theta)L(\mathcal{X};\theta)\partial\mathcal{X}$ *and then permute integral and derivative.)*