Chapter 8

Maximum Likelihood Estimation

8.1 Consistency

If $X$ is a random variable (or vector) with density or mass function $f_\theta(x)$ that depends on a parameter $\theta$, then the function $f_\theta(X)$ viewed as a function of $\theta$ is called the likelihood function of $\theta$. We often denote this function by $L(\theta)$. Note that $L(\theta) = f_\theta(X)$ is implicitly a function of $X$, but we suppress this fact in the notation. Since repeated references to the “density or mass function” would be awkward, we will use the term “density” to refer to $f_\theta(x)$ throughout this chapter, even if the distribution function of $X$ is not continuous. (Allowing noncontinuous distributions to have density functions may be made technically rigorous; however, this is a measure theoretic topic beyond the scope of this book.)

Let the set of possible values of $\theta$ be the set $\Omega$. If $L(\theta)$ has a maximizer in $\Omega$, say $\hat{\theta}$, then $\hat{\theta}$ is called a maximum likelihood estimator or MLE. Since the logarithm function is a strictly increasing function, any maximizer of $L(\theta)$ also maximizes $\ell(\theta) \overset{\text{def}}{=} \log L(\theta)$. It is often much easier to maximize $\ell(\theta)$, called the loglikelihood function, than $L(\theta)$.

Example 8.1 Suppose $\Omega = (0, \infty)$ and $X \sim \text{binomial}(n, e^{-\theta})$. Then

$$\ell(\theta) = \log \binom{n}{X} - X \theta + (n - X) \log(1 - e^{-\theta}),$$

so

$$\ell'(\theta) = -X + \frac{X - n}{1 - e^{\theta}}.$$ 

Thus, setting $\ell'(\theta) = 0$ yields $\theta = -\log(X/n)$. It isn’t hard to verify that $\ell''(\theta) < 0$, so that $-\log(X/n)$ is in fact a maximizer of $\ell(\theta)$. 

119
As the preceding example demonstrates, it is not always the case that a MLE exists — for if \( X = 0 \) or \( X = n \), then \( \log(-X/n) \) is not contained in \( \Omega \). This is just one of the technical details that we will consider. Ultimately, we will show that the maximum likelihood estimator is, in many cases, asymptotically normal. However, this is not always the case; in fact, it is not even necessarily true that the MLE is consistent, as shown in Problem 8.1.

We begin the discussion of the consistency of the MLE by defining the so-called Kullback-Leibler information.

**Definition 8.2** If \( f_{\theta_0}(x) \) and \( f_{\theta_1}(x) \) are two densities, the Kullback-Leibler information number equals

\[
K(f_{\theta_0}, f_{\theta_1}) = E_{\theta_0} \log \frac{f_{\theta_0}(X)}{f_{\theta_1}(X)}.
\]

If \( P_{\theta_0}(f_{\theta_0}(X) > 0) \) and \( f_{\theta_1}(X) = 0 \), then \( K(f_{\theta_0}, f_{\theta_1}) \) is defined to be \( \infty \).

We may show that the Kullback-Leibler information must be nonnegative by noting that

\[
E_{\theta_0} \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} = E_{\theta_1} I\{f_{\theta_0}(X) > 0\} \leq 1.
\]

Therefore, by Jensen’s inequality (1.24) and the strict convexity of the function \(-\log x\),

\[
K(f_{\theta_0}, f_{\theta_1}) = E_{\theta_0} - \log \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} \geq -\log E_{\theta_0} \frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} \geq 0,
\]

with equality if and only if \( P_{\theta_0} \{f_{\theta_0}(X) = f_{\theta_1}(X)\} = 1 \). Inequality (8.1) is sometimes called the Shannon-Kolmogorov information inequality.

In the (admittedly somewhat bizarre) case in which the parameter space \( \Omega \) contains only finitely many points, the Shannon-Kolmogorov information inequality may be used to prove the consistency of the maximum likelihood estimator. For the proof of the following theorem, note that if \( X_1, \ldots, X_n \) are independent and identically distributed with density \( f_{\theta_0}(x) \), then the loglikelihood is \( \ell(\theta) = \sum_{i=1}^{n} \log f_{\theta_0}(x_i) \).

**Theorem 8.3** Suppose \( \Omega \) contains finitely many elements and that \( X_1, \ldots, X_n \) are independent and identically distributed with density \( f_{\theta_0}(x) \). Furthermore, suppose that the model is identifiable, which is to say that different values of \( \theta \) lead to different distributions. Then if \( \hat{\theta}_n \) denotes the maximum likelihood estimator, \( \hat{\theta}_n \xrightarrow{P} \theta_0 \).

**Proof:** The Weak Law of Large Numbers (Theorem 2.18) implies that

\[
\frac{1}{n} \sum_{i=1}^{n} \log \frac{f_{\theta}(X_i)}{f_{\theta_0}(X_i)} \xrightarrow{P} E_{\theta_0} \log \frac{f_{\theta}(X_i)}{f_{\theta_0}(X_i)} = -K(f_{\theta_0}, f_{\theta}).
\]
The value of $-K(f_{\theta_0}, f_\theta)$ is strictly negative for $\theta \neq \theta_0$ by the identifiability of the model. Therefore, since $\theta = \theta_n$ is the maximizer of the left hand side of Equation (8.2),

$$P(\hat{\theta}_n \neq \theta_0) = P\left( \max_{\theta \neq \theta_0} \frac{1}{n} \sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} > 0 \right) \leq \sum_{\theta \neq \theta_0} P\left( \frac{1}{n} \sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta_0}(X_i)} > 0 \right) \to 0.$$ 

This implies that $\hat{\theta}_n \xrightarrow{P} \theta_0$. ■

The result of Theorem 8.3 may be extended in several ways; however, it is unfortunately not true in general that a maximum likelihood estimator is consistent, as demonstrated by the example of Problem 8.1.

If we return to the simple Example 8.1, we found that the MLE was found by solving the equation

$$\ell'(\theta) = 0.$$  \hspace{1cm} (8.3)

Equation (8.3) is called the likelihood equation, and naturally a root of the likelihood equation is a good candidate for a maximum likelihood estimator. However, there may be no root and there may be more than one. It turns out the probability that at least one root exists goes to 1 as $n \to \infty$. Consider Example 8.1, in which no MLE exists whenever $X = 0$ or $X = n$. In that case, both $P(X = 0) = (1 - e^{-\theta})^n$ and $P(X = n) = e^{-n\theta}$ go to zero as $n \to \infty$. In the case of multiple roots, one of these roots is typically consistent for $\theta_0$, as stated in the following theorem.

**Theorem 8.4** Suppose that $X_1, \ldots, X_n$ are independent and identically distributed with density $f_{\theta_0}(x)$ for $\theta_0$ in an open interval $\Omega \subset R$, where the model is identifiable (i.e., different values of $\theta \in \Omega$ give different distributions). Furthermore, suppose that the loglikelihood function $\ell(\theta)$ is differentiable and that the support \( \{ x : f_\theta(x) > 0 \} \) does not depend on $\theta$. Then with probability approaching 1 as $n \to \infty$, there exists $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ such that $\ell'(\hat{\theta}_n) = 0$ and $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Stated succinctly, Theorem 8.4 says that under certain regularity conditions, there is a consistent root of the likelihood equation. It is important to note that there is no guarantee that this consistent root is the MLE. However, if the likelihood equation only has a single root, we can be more precise:

**Corollary 8.5** Under the conditions of Theorem 8.4, if for every $n$ there is a unique root of the likelihood equation, and this root is a local maximum, then this root is the MLE and the MLE is consistent.

**Proof:** The only thing that needs to be proved is the assertion that the unique root is the MLE. Denote the unique root by $\hat{\theta}_n$ and suppose there is some other point $\theta$ such that
\( \ell(\theta) \geq \ell(\hat{\theta}_n) \). Then there must be a local minimum between \( \hat{\theta}_n \) and \( \theta \), which contradicts the assertion that \( \hat{\theta}_n \) is the unique root of the likelihood equation. \( \blacksquare \)

Exercises for Section 8.1

**Exercise 8.1** In this problem, we explore an example in which the MLE is not consistent. Suppose that for \( \theta \in (0, 1) \), \( X \) is a continuous random variable with density

\[
 f_\theta(x) = \frac{3(1 - \theta)}{4\delta^3(\theta)} \left[ \delta^2(\theta) - (x - \theta)^2 \right] I\{|x - \theta| \leq \delta(\theta)\} + \frac{\theta}{2} I\{|x| \leq 1\}, \tag{8.4}
\]

where \( \delta(\theta) > 0 \) for all \( \theta \).

(a) Prove that \( f_\theta(x) \) is a legitimate density.

(b) What condition on \( \delta(\theta) \) ensures that \( \{x : f_\theta(x) > 0\} \) does not depend on \( \theta \)?

(c) With \( \delta(\theta) = \exp\{-(1 - \theta)^{-1}\} \), let \( \theta = .125 \). Take samples of sizes \( n \in \{50, 250, 500\} \) from \( f_\theta(x) \). In each case, graph the loglikelihood function and find the MLE. Also, try to identify the consistent root of the likelihood equation in each case.

**Hints:** To generate a sample from \( f_\theta(x) \), note that \( f_\theta(x) \) is a mixture density, which means you can start by generating a standard uniform random variable. If it’s less than \( \theta \), generate a uniform variable on \((-1, 1)\). Otherwise, generate a variable with density \( 3(\delta^2 - x^2)/4\delta^3 \) on \((-\delta, \delta)\) and then add \( \theta \). You should be able to do this by inverting the distribution function or by using appropriately scaled and translated beta variables. Be very careful when graphing the loglikelihood and finding the MLE. In particular, make sure you evaluate the loglikelihood analytically at each of the sample points in \((0, 1)\); if you fail to do this, you’ll miss the point of the problem and you’ll get the MLE incorrect. This is because the correct loglikelihood graph will have tall, extremely narrow spikes.

**Exercise 8.2** In the situation of Problem 8.1, prove that the MLE is inconsistent.

**Exercise 8.3** Suppose that \( X_1, \ldots, X_n \) are independent and identically distributed with density \( f_\theta(x) \), where \( \theta \in (0, \infty) \). For each of the following forms of \( f_\theta(x) \), prove that the likelihood equation has a unique solution and that this solution maximizes the likelihood function.

(a) **Weibull:** For some constant \( a > 0 \),

\[
 f_\theta(x) = a\theta^a x^{a-1} \exp\{-(\theta x)^a\} I\{x > 0\}
\]
(b) Cauchy:

\[ f_\theta(x) = \frac{\theta}{\pi x^2 + \theta^2} \]

(c)

\[ f_\theta(x) = \frac{3\theta^2\sqrt{3}}{2\pi(x^3 + \theta^3)}I\{x > 0\} \]

**Exercise 8.4** Find the MLE and its asymptotic distribution given a random sample of size \( n \) from \( f_\theta(x) = (1 - \theta)x, x = 0, 1, 2, \ldots, \theta \in (0, 1). \)

**Hint:** For the asymptotic distribution, use the central limit theorem.

### 8.2 Asymptotic normality of the MLE

As seen in the preceding section, the MLE is not necessarily even consistent, let alone asymptotically normal, so the title of this section is slightly misleading — however, “Asymptotic normality of the consistent root of the likelihood equation” is a bit too long! It will be necessary to review a few facts regarding Fisher information before we proceed.

**Definition 8.6** *Fisher information:* For a density (or mass) function \( f_\theta(x) \), the Fisher information function is given by

\[ I(\theta) = E_\theta \left\{ \frac{d}{d\theta} \log f_\theta(X) \right\}^2. \tag{8.5} \]

If \( \eta = g(\theta) \) for some invertible and differentiable function \( g(\cdot) \), then since

\[ \frac{d}{d\eta} = \frac{d\theta}{d\eta} \frac{d}{d\theta} = \frac{1}{g'(\theta)} \frac{d}{d\theta} \]

by the chain rule, we conclude that

\[ I(\eta) = \frac{I(\theta)}{\{g'(\theta)\}^2}. \tag{8.6} \]

Loosely speaking, \( I(\theta) \) is the amount of information about \( \theta \) contained in a single observation from the density \( f_\theta(x) \). However, this interpretation doesn’t always make sense — for example, it is possible to have \( I(\theta) = 0 \) for a very informative observation. See Exercise 8.5.
Although we do not dwell on this fact in this course, expectation may be viewed as integration. Suppose that \( f_\theta(x) \) is twice differentiable with respect to \( \theta \) and that the operations of differentiation and integration may be interchanged in the following sense:

\[
\frac{d}{d \theta} \mathbb{E}_\theta \frac{f_\theta(X)}{f_\theta(X)} = \mathbb{E}_\theta \frac{\frac{d}{d \theta} f_\theta(X)}{f_\theta(X)},
\]

and

\[
\frac{d^2}{d \theta^2} \mathbb{E}_\theta \frac{f_\theta(X)}{f_\theta(X)} = \mathbb{E}_\theta \frac{\frac{d^2}{d \theta^2} f_\theta(X)}{f_\theta(X)}.
\]

Equations (8.7) and (8.8) give two additional expressions for \( I(\theta) \). From Equation (8.7) follows

\[
I(\theta) = \text{Var}_\theta \left\{ \frac{d}{d \theta} \log f_\theta(X) \right\},
\]

and Equation (8.8) implies

\[
I(\theta) = -\mathbb{E}_\theta \left\{ \frac{d^2}{d \theta^2} \log f_\theta(X) \right\}.
\]

In many cases, Equation (8.10) is the easiest form of the information to work with.

Equations (8.9) and (8.10) make clear a helpful property of the information, namely that for independent random variables, the information about \( \theta \) contained in the joint sample is simply the sum of the individual information components. In particular, if we have a simple random sample of size \( n \) from \( f_\theta(x) \), then the information about \( \theta \) equals \( nI(\theta) \).

The reason that we need the Fisher information is that we will show that under certain regularity conditions,

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\rightarrow} N\left\{ 0, \frac{1}{I(\theta_0)} \right\},
\]

where \( \hat{\theta}_n \) is the consistent root of the likelihood equation guaranteed to exist by Theorem 8.4.

**Example 8.7** Suppose that \( X_1, \ldots, X_n \) are independent Poisson(\( \theta_0 \)) random variables. Then the likelihood equation has a unique root, namely \( \hat{\theta}_n = \bar{X}_n \), and we know that by the central limit theorem \( \sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\rightarrow} N(0, \theta_0) \). Furthermore, the Fisher information for a single observation in this case is

\[- \mathbb{E}_\theta \left\{ \frac{d}{d \theta} f_\theta(X) \right\} = \mathbb{E}_\theta \frac{X}{\theta^2} = \frac{1}{\theta}.\]

Thus, in this example, Equation (8.11) holds.
Rather than stating all of the regularity conditions necessary to prove Equation (8.11), we work backwards, figuring out the conditions as we go through the proof. The first step is to expand $\ell'(\hat{\theta}_n)$ in a Taylor series around $\theta_0$:

$$\ell'(\hat{\theta}_n) = \ell'(\theta_0) + (\hat{\theta}_n - \theta_0) \left\{ \ell''(\theta_0) + \frac{(\hat{\theta}_n - \theta_0)\ell'''(\theta^*)}{2} \right\}. \tag{8.12}$$

In order for the second derivative $\ell''(\theta_0)$ to exist, it is enough to assume that Equation (8.8) holds (we will need this equation later anyway). Rewriting Equation (8.12) gives

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \sqrt{n} \left\{ \ell'(\hat{\theta}_n) - \ell'(\theta_0) \right\} = \frac{1}{\sqrt{n}} \left\{ \ell'(\theta_0) - \ell'(\hat{\theta}_n) \right\} - \frac{1}{n} \ell''(\theta_0) - (\hat{\theta}_n - \theta_0)\ell'''(\theta^*)/(2n). \tag{8.13}$$

Let’s consider the pieces of Equation (8.13) individually. If the conditions of Theorem 8.4 are met, then $\ell'(\hat{\theta}_n) \xrightarrow{P} 0$. If Equation (8.7) holds and $I(\theta_0) < \infty$, then

$$\frac{1}{\sqrt{n}} \ell'(\theta_0) = \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} \frac{d}{d\theta} \log f_{\theta_0}(X_i) \right)^d \xrightarrow{d} N\{0, I(\theta_0)\}$$

by the central limit theorem and Equation (8.9). If Equation (8.8) holds, then

$$\frac{1}{n} \ell''(\theta_0) = \frac{1}{n} \sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f_{\theta_0}(X_i) \xrightarrow{P} -I(\theta_0)$$

by the weak law of large numbers and Equation (8.10).

Finally, if we assume that the third derivative $\ell'''(\theta)$ is uniformly bounded in a neighborhood of $\theta_0$, say by the constant $K_0$, we obtain

$$\left| \frac{(\hat{\theta}_n - \theta_0)\ell'''(\theta^*)}{2n} \right| \leq \left| \frac{(\hat{\theta}_n - \theta_0)K_0}{2n} \right| \xrightarrow{P} 0.$$ 

In conclusion, if all of the our assumptions hold, then the numerator of (8.13) converges in distribution to $N\{0, I(\theta_0)\}$ by Slutsky’s theorem. Furthermore, the denominator in (8.13) converges to $I(\theta_0)$, so a second use of Slutsky’s theorem gives the following theorem.

**Theorem 8.8** Suppose that the conditions of Theorem 8.4 are satisfied, and let $\hat{\theta}_n$ denote a consistent root of the likelihood equation. Assume also that Equations (8.7) and (8.8) hold and that $0 < I(\theta_0) < \infty$. Finally, assume that $\ell(\theta)$ has three derivatives in a neighborhood of $\theta_0$ and that $\ell'''(\theta)$ is uniformly bounded in this neighborhood. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N \left\{ 0, \frac{1}{I(\theta_0)} \right\}.$$
Sometimes, it is not possible to find an exact zero of $\ell'(\theta)$. One way to get a numerical approximation to a zero of $\ell'(\theta)$ is to use Newton’s method, in which we start at a point $\theta_0$ and then set

$$
\theta_1 = \theta_0 - \frac{\ell'(\theta_0)}{\ell''(\theta_0)}.
$$

Ordinarily, after finding $\theta_1$ we would set $\theta_0$ equal to $\theta_1$ and apply Equation (8.14) iteratively. However, we may show that by using a single step of Newton’s method, starting from a $\sqrt{n}$-consistent estimator of $\theta_0$, we may obtain an estimator with the same asymptotic distribution as $\hat{\theta}_n$. The proof of the following theorem is left as an exercise:

**Theorem 8.9** Suppose that $\hat{\theta}_n$ is any $\sqrt{n}$-consistent estimator of $\theta_0$ (i.e., $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is bounded in probability). Then under the conditions of Theorem 8.8, if we set

$$
\delta_n = \hat{\theta}_n - \frac{\ell'(\hat{\theta}_n)}{\ell''(\hat{\theta}_n)},
$$

then

$$
\sqrt{n}(\delta_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right).
$$

**Exercises for Section 8.2**

**Exercise 8.5** Suppose that $X$ is a normal random variable with mean $\theta^3$ and known variance $\sigma^2$. Calculate $I(\theta)$, then argue that the Fisher information can be zero in a case in which there is information about $\theta$ in the observed value of $X$.

**Exercise 8.6** (a) Show that under the conditions of Theorem 8.8, then if $\hat{\theta}_n$ is a consistent root of the likelihood equation, $P_{\theta_0}(\hat{\theta}_n$ is a local maximum) $\to 1$.

(b) Using the result of part (a), show that for any two sequences $\hat{\theta}_{1n}$ and $\hat{\theta}_{2n}$ of consistent roots of the likelihood equation, $P_{\theta_0}(\hat{\theta}_{1n} = \hat{\theta}_{2n}) \to 1$.

**Exercise 8.7** Prove Theorem 8.9.

**Hint:** Start with $\sqrt{n}(\delta_n - \theta_0) = \sqrt{n}(\delta_n - \hat{\theta}_n) + \sqrt{n}(\hat{\theta}_n - \theta_0)$, then expand $\ell'(\hat{\theta}_n)$ in a Taylor series about $\theta_0$ and rewrite $\sqrt{n}(\hat{\theta}_n - \theta_0)$ using this expansion.

**Exercise 8.8** Suppose that the following is a random sample from a logistic density with distribution function $F_{\theta}(x) = (1 + \exp\{\theta - x\})^{-1}$ (I’ll cheat and tell you that I used $\theta = 2$.)

126
(a) Evaluate the unique root of the likelihood equation numerically. Then, taking the sample median as our known $\sqrt{n}$-consistent estimator $\tilde{\theta}_n$ of $\theta$, evaluate the estimator $\delta_n$ in Equation (8.15) numerically.

(b) Find the asymptotic distributions of $\sqrt{n}(\tilde{\theta}_n - 2)$ and $\sqrt{n}(\delta_n - 2)$. Then, simulate 200 samples of size $n = 15$ from the logistic distribution with $\theta = 2$. Find the sample variances of the resulting sample medians and $\delta_n$-estimators. How well does the asymptotic theory match reality?

### 8.3 Asymptotic Efficiency and Superefficiency

In Theorem 8.8, we showed that a consistent root $\hat{\theta}_n$ of the likelihood equation satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right).$$

In Theorem 8.9, we stated that if $\tilde{\theta}_n$ is a $\sqrt{n}$-consistent estimator of $\theta_0$ and $\delta_n = \tilde{\theta}_n - \ell'(\tilde{\theta}_n)/\ell''(\tilde{\theta}_n)$, then

$$\sqrt{n}(\delta_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right).$$

(8.16)

Note the similarity in the last two asymptotic limit statements. There seems to be something special about the limiting variance $1/I(\theta_0)$, and in fact this is true.

Much like the Cramér-Rao lower bound states that (under some regularity conditions) an unbiased estimator of $\theta_0$ cannot have a variance smaller than $1/I(\theta_0)$, the following result is true:

#### Theorem 8.10

Suppose that the conditions of theorem 8.8 are satisfied and that $\delta_n$ is an estimator satisfying

$$\sqrt{n}(\delta_n - \theta_0) \xrightarrow{d} N\{0, v(\theta_0)\}$$

for all $\theta_0$, where $v(\theta)$ is continuous. Then $v(\theta) \geq 1/I(\theta)$ for all $\theta$.

In other words, $1/I(\theta)$ is in a sense the smallest possible asymptotic variance for a $\sqrt{n}$-consistent estimator. For this reason, we refer to any estimator $\delta_n$ satisfying (8.16) for all $\theta_0$ an efficient estimator.
One condition in Theorem 8.10 that may be a bit puzzling is the condition that \( v(\theta) \) be continuous. If this condition is dropped, then a well-known counterexample, due to Hodges, exists:

**Example 8.11** Suppose that \( \delta_n \) is an efficient estimator of \( \theta_0 \). Then if we define

\[
\delta^*_n = \begin{cases} 
0 & \text{if } n(\delta_n)^4 < 1; \\
\delta_n & \text{otherwise,}
\end{cases}
\]

it is possible to show (see Exercise 8.9) that \( \delta^*_n \) is superefficient in the sense that

\[
\sqrt{n}(\delta^*_n - \theta_0) \overset{d}{\to} N\left(0, \frac{1}{I(\theta_0)}\right)
\]

for all \( \theta_0 \neq 0 \) but \( \sqrt{n}(\delta^*_n - \theta_0) \overset{d}{\to} 0 \) if \( \theta_0 = 0 \).

Just as the invariance property of maximum likelihood estimation states that the MLE of a function of \( \theta \) equals the same function applied to the MLE of \( \theta \), a function of an efficient estimator is itself efficient:

**Theorem 8.12** If \( \delta_n \) is efficient for \( \theta_0 \), and if \( g(\theta) \) is a differentiable and invertible function with \( g'(\theta_0) \neq 0 \), \( g(\delta_n) \) is efficient for \( g(\theta_0) \).

The proof of the above theorem follows immediately from the delta method, since the Fisher information for \( g(\theta) \) is \( I(\theta)/\{g'(\theta)\}^2 \) by Equation (8.6).

We have already noted that (under suitable regularity conditions) if \( \tilde{\theta}_n \) is a \( \sqrt{n} \)-consistent estimator of \( \theta_0 \) and

\[
\delta_n = \tilde{\theta}_n - \frac{\ell'(\tilde{\theta}_n)}{\ell''(\tilde{\theta}_n)}, \quad (8.17)
\]

then \( \delta_n \) is an efficient estimator of \( \theta_0 \). Alternatively, we may set

\[
\delta^*_n = \tilde{\theta}_n + \frac{\ell'(\tilde{\theta}_n)}{nI(\tilde{\theta}_n)} \quad (8.18)
\]

and \( \delta^*_n \) is also an efficient estimator of \( \theta_0 \). Problem 8.7 asked you to prove the former fact regarding \( \delta_n \); the latter fact regarding \( \delta^*_n \) is proved in nearly the same way because

\[
-\frac{1}{n}\ell''(\tilde{\theta}_n) \overset{P}{\to} I(\theta_0).
\]

In Equation (8.17), as already remarked earlier, \( \delta_n \) results from a single step of Newton’s method; in Equation (8.18), \( \delta^*_n \) results from a similar method called Fisher scoring. As is clear from comparing Equations (8.17) and (8.18), scoring differs from Newton’s method in that the Fisher information is used in place of the negative second
derivative of the loglikelihood function. In some examples, scoring and Newton’s method are equivalent.

A note on terminology: The derivative of \( \ell(\theta) \) is sometimes called the **score function**. Furthermore, \( nI(\theta) \) and \(-\ell''(\theta)\) are sometimes referred to as the **expected information** and the **observed information**, respectively.

**Example 8.13** Suppose \( X_1, \ldots, X_n \) are independent from a Cauchy location family with density

\[
f_\theta(x) = \frac{1}{\pi \{1 + (x - \theta)^2\}}.
\]

Then

\[
\ell'(\theta) = \sum_{i=1}^{n} \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2},
\]

so the likelihood equation is very difficult to solve. However, an efficient estimator may still be created by starting with some \( \sqrt{n} \)-consistent estimator \( \hat{\theta}_n \), say the sample median, and using either Equation (8.17) or Equation (8.18). In the latter case, we obtain the simple estimator

\[
\delta_n^* = \hat{\theta}_n + \frac{2}{n} \ell'(\hat{\theta}_n),
\]

verification of which is the subject of Problem 8.10.

For the remainder of this section, we turn our attention to Bayes estimators, which give yet another source of efficient estimators. A Bayes estimator is the expected value of the posterior distribution, which of course depends on the prior chosen. Although we do not prove this fact here (see Ferguson §21 for details), any Bayes estimator is efficient under some very general conditions. The conditions are essentially those of Theorem 8.8 along with the stipulation that the prior density is positive everywhere on \( \Omega \). (Note that if the prior density is not positive on \( \Omega \), the Bayes estimator may not even be consistent.)

**Example 8.14** Consider the binomial distribution with beta prior, say \( X \sim \text{binomial}(n, p) \) and \( p \sim \text{beta}(a, b) \). Then the posterior density of \( p \) is proportional to the product of the likelihood and the prior, which (ignoring multiplicative constants not involving \( p \)) equals

\[
p^{a-1}(1-p)^{b-1} \times p^X(1-p)^{n-X}.
\]

Therefore, the posterior distribution of \( p \) is \( \text{beta}(a + X, b + n - X) \). The Bayes estimator is the expectation of this distribution, which equals \( (a + X)/(a + b + n) \).
If we let $\gamma_n$ denote the Bayes estimator here, then

$$\sqrt{n}(\gamma_n - p) = \sqrt{n} \left( \frac{X}{n} - p \right) + \sqrt{n} \left( \gamma_n - \frac{X}{n} \right).$$

We may see that the rightmost term converges to zero in probability by writing

$$\sqrt{n} \left( \gamma_n - \frac{X}{n} \right) = \sqrt{n} \left( \frac{a}{a + b + n} \right) \left( a + (a + b) \frac{X}{n} \right),$$

since $a + (a + b)X/n \xrightarrow{P} a + (a + b)p$ by the weak law of large numbers. In other words, the Bayes estimator in this example has the same limiting distribution as the MLE, $X/n$. It is possible to verify that the MLE is efficient in this case.

The central question when constructing a Bayes estimator is how to choose the prior distribution. We consider one class of prior distributions, called Jeffreys priors. Note that these priors are named for Harold Jeffreys, so it is incorrect to write “Jeffrey’s priors”. For a Bayes estimator $\hat{\theta}_n$ of $\theta_0$, we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N \left( 0, \frac{1}{I(\theta_0)} \right).$$

Since the limiting variance of $\hat{\theta}_n$ depends on $I(\theta_0)$, if $I(\theta)$ is not a constant, then some values of $\theta$ may be estimated more precisely than others. In analogy with the idea of variance-stabilizing transformations seen in Section 5.2, we might consider a reparameterization $\eta = g(\theta)$ such that $\{g'(\theta)\}^2/I(\theta)$ is a constant. More precisely, if $g'(\theta) = c\sqrt{I(\theta)}$, then

$$\sqrt{n} \left( g(\hat{\theta}_n) - \eta_0 \right) \xrightarrow{d} N(0, c^2).$$

So as not to influence the estimation of $\eta$, we choose as the Jeffreys prior a uniform prior on $\eta$. Therefore, the Jeffreys prior density on $\theta$ is proportional to $g'(\theta)$, which is proportional to $\sqrt{I(\theta)}$. Note that this may lead to an improper prior.

**Example 8.15** In the case of Example 8.14, we may verify that for a Bernoulli($p$) observation,

$$I(p) = -E \frac{d^2}{dp^2} \{X \log p + (1 - X) \log(1 - p)\} = E \left( \frac{X}{p^2} + \frac{1 - X}{(1 - p)^2} \right) = \frac{1}{p(1 - p)}.$$

Thus, the Jeffreys prior on $p$ in this case has a density proportional to $p^{-1/2}(1 - p)^{-1/2}$. In other words, the prior is beta($\frac{1}{2}$, $\frac{1}{2}$). Therefore, the Bayes estimator corresponding to the Jeffreys prior is

$$\gamma_n = \frac{X + \frac{1}{2}}{n + 1}.$$
Exercises for Section 8.3

Exercise 8.9 Verify the claim made in Example 8.11.

Exercise 8.10 If \( f_\theta(x) \) forms a location family, so that \( f_\theta(x) = f(x - \theta) \) for some density \( f(x) \), then the Fisher information \( I(\theta) \) is a constant (you may assume this fact without proof).

(a) Verify that for the Cauchy location family,

\[
f_\theta(x) = \frac{1}{\pi \{1 + (x - \theta)^2\}},
\]

we have \( I(\theta) = \frac{1}{2} \).

(b) For 500 samples of size \( n = 51 \) from a standard Cauchy distribution, calculate the sample median \( \hat{\theta}_n \) and the efficient estimator \( \delta_n^* \) of Equation (8.19). Compare the variances of \( \hat{\theta}_n \) and \( \delta_n^* \) with their theoretical asymptotic limits.

Exercise 8.11 (a) Derive the Jeffreys prior on \( \theta \) for a random sample from Poisson(\( \theta \)). Is this prior proper or improper?

(b) What is the Bayes estimator of \( \theta \) for the Jeffreys prior? Verify directly that this estimator is efficient.

Exercise 8.12 (a) Derive the Jeffreys prior on \( \sigma^2 \) for a random sample from \( N(0, \sigma^2) \). Is this prior proper or improper?

(b) What is the Bayes estimator of \( \sigma^2 \) for the Jeffreys prior? Verify directly that this estimator is efficient.

8.4 The multiparameter case

Suppose now that the parameter is the vector \( \theta = (\theta_1, \ldots, \theta_k) \). If \( X \sim f_\theta(x) \), then \( I(\theta) \), the information matrix, is the \( k \times k \) matrix

\[
I(\theta) = \mathbb{E}_\theta \{h_\theta(X)h_\theta^T(X)\},
\]

where \( h_\theta(x) = \nabla_\theta[\log f_\theta(x)] \). This is a rare instance in which it’s probably clearer to use component-wise notation than vector notation:

Definition 8.16 Given a \( k \)-dimensional parameter vector \( \theta \) and a density function \( f_\theta(x) \), the Fisher information matrix \( I(\theta) \) is the \( k \times k \) matrix whose \((i, j)\) element
equals
\[ I_{ij}(\theta) = \mathbb{E}_\theta \left\{ \frac{\partial}{\partial \theta_i} \log f_\theta(X) \frac{\partial}{\partial \theta_j} \log f_\theta(X) \right\}, \]
as long as the above quantity is defined for all \((i, j)\).

Note that the one-dimensional Definition 8.6 is a special case of Definition 8.16.

Example 8.17 Let \(\theta = (\mu, \sigma^2)\) and suppose \(X \sim N(\mu, \sigma^2)\). Then
\[
\log f_\theta(x) = -\frac{1}{2} \log \sigma^2 - \frac{(x - \mu)^2}{2\sigma^2} - \log \sqrt{2\pi},
\]
so
\[
\frac{\partial}{\partial \mu} \log f_\theta(x) = \frac{x - \mu}{\sigma^2} \quad \text{and} \quad \frac{\partial}{\partial \sigma^2} \log f_\theta(x) = \frac{1}{2\sigma^2} \left( \frac{(x - \mu)^2}{\sigma^2} - 1 \right).
\]
Thus, the entries in the information matrix are as follows:
\[
I_{11}(\theta) = \mathbb{E}_\theta \left( \frac{(X - \mu)^2}{\sigma^4} \right) = \frac{1}{\sigma^2},
\]
\[
I_{21}(\theta) = I_{12}(\theta) = \mathbb{E}_\theta \left( \frac{(X - \mu)^3}{2\sigma^6} - \frac{X - \mu}{2\sigma^4} \right) = 0,
\]
\[
I_{22}(\theta) = \mathbb{E}_\theta \left( \frac{1}{4\sigma^2} - \frac{(X - \mu)^2}{2\sigma^6} + \frac{(X - \mu)^4}{4\sigma^8} \right) = \frac{1}{4\sigma^4} - \frac{\sigma^2}{2\sigma^6} + \frac{3\sigma^4}{4\sigma^8} = \frac{1}{2\sigma^4}.
\]

As in the one-dimensional case, Definition 8.16 is often not the easiest form of \(I(\theta)\) to work with. This fact is illustrated by Example 8.17, in which the calculation of the information matrix requires the evaluation of a fourth moment as well as a lot of algebra. However, in analogy with Equations (8.7) and (8.8), if
\[
0 = \frac{\partial}{\partial \theta_i} \mathbb{E}_\theta \frac{f_\theta(X)}{f_\theta(X)} = \mathbb{E}_\theta \frac{\partial}{\partial \theta_i} f_\theta(X) f_\theta(X) \quad \text{and} \quad 0 = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \mathbb{E}_\theta \frac{f_\theta(X)}{f_\theta(X)} = \mathbb{E}_\theta \frac{\partial^2}{\partial \theta_i \partial \theta_j} f_\theta(X) f_\theta(X) \tag{8.20}
\]
for all \(i\) and \(j\), then the following alternative forms of \(I(\theta)\) are valid:
\[
I_{ij}(\theta) = \text{Cov}_\theta \left( \frac{\partial}{\partial \theta_i} \log f_\theta(X), \frac{\partial}{\partial \theta_j} \log f_\theta(X) \right) \tag{8.21}
\]
\[
= -\mathbb{E}_\theta \left( \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f_\theta(X) \right). \tag{8.22}
\]
Example 8.18  In the normal case of Example 8.17, the information matrix is perhaps a bit easier to compute using Equation (8.22), since we obtain

\[
\frac{\partial^2}{\partial \mu^2} \log f_\theta(x) = \frac{1}{\sigma^2}, \quad \frac{\partial^2}{\partial \mu \partial \sigma^2} \log f_\theta(x) = \frac{x-\mu}{\sigma^4},
\]

and

\[
\frac{\partial^2}{\partial (\sigma^2)^2} \log f_\theta(x) = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6},
\]

taking expectations gives

\[
I(\theta) = \begin{pmatrix}
\frac{1}{\sigma^2} & 0 \\
0 & \frac{1}{2\sigma^4}
\end{pmatrix}
\]

as before but without requiring any fourth moments.

By Equation (8.21), the Fisher information matrix is nonnegative definite, just as the Fisher information is nonnegative in the one-parameter case. A further fact the generalizes into the multiparameter case is the additivity of information: If \(X\) and \(Y\) are independent, then \(I_X(\theta) + I_Y(\theta) = I_{(X,Y)}(\theta)\). Finally, suppose that \(\eta = g(\theta)\) is a reparameterization, where \(g(\theta)\) is invertible and differentiable. Then if \(J\) is the Jacobian matrix of the inverse transformation (i.e., \(J_{ij} = \partial \theta_i / \partial \eta_j\)), then the information about \(\eta\) is

\[
I(\eta) = J^T I(\theta) J.
\]

As we’ll see later, almost all of the same efficiency results that applied to the one-parameter case apply to the multiparameter case as well. In particular, we will see that an efficient estimator \(\hat{\theta}\) is one that satisfies

\[
\sqrt{n}(\hat{\theta} - \theta^0) \overset{d}{\rightarrow} N_k\{0, I(\theta^0)^{-1}\}.
\]

Note that the formula for the information under a reparameterization implies that if \(\hat{\eta}\) is an efficient estimator for \(\eta^0\) and the matrix \(J\) is invertible, then \(g^{-1}(\hat{\eta})\) is an efficient estimator for \(\theta^0 = g^{-1}(\eta^0)\), since

\[
\sqrt{n}\{g^{-1}(\hat{\eta}) - g^{-1}(\eta^0)\} \overset{d}{\rightarrow} N_k\{0, J(J^T I(\theta^0) J)^{-1} J\},
\]

and the covariance matrix above equals \(I(\theta^0)^{-1}\).

As in the one-parameter case, the likelihood equation is obtained by setting the derivative of \(\ell(\theta)\) equal to zero. In the multiparameter case, though, the gradient \(\nabla \ell(\theta)\) is a \(1 \times k\) vector, so the likelihood equation becomes \(\nabla \ell(\theta) = 0\). Since there are really \(k\) univariate equations
implied by this likelihood equation, it is common to refer to the likelihood equations (plural), which are
\[ \frac{\partial}{\partial \theta_i} \ell(\theta) = 0 \quad \text{for } i = 1, \ldots, k. \]

In the multiparameter case, we have essentially the same theorems as in the one-parameter case.

**Theorem 8.19** Suppose that \( X_1, \ldots, X_n \) are independent and identically distributed random variables (or vectors) with density \( f_{\theta^0}(x) \) for \( \theta^0 \) in an open subset \( \Omega \) of \( \mathbb{R}^k \), where distinct values of \( \theta^0 \) yield distinct distributions for \( X_1 \) (i.e., the model is identifiable). Furthermore, suppose that the support set \( \{ x : f_\theta(x) > 0 \} \) does not depend on \( \theta \). Then with probability approaching 1 as \( n \to \infty \), there exists \( \hat{\theta} \) such that \( \nabla \ell(\hat{\theta}) = 0 \) and \( \hat{\theta} \overset{P}{\to} \theta^0 \).

As in the one-parameter case, we shall refer to the \( \hat{\theta} \) guaranteed by Theorem 8.19 as a consistent root of the likelihood equations. Unlike Theorem 8.4, however, Corollary 8.5 does not generalize to the multiparameter case because it is possible that \( \hat{\theta} \) is the unique solution of the likelihood equations and a local maximum but not the MLE. The best we can say is the following:

**Corollary 8.20** Under the conditions of Theorem 8.19, if there is a unique root of the likelihood equations, then this root is consistent for \( \theta^0 \).

The asymptotic normality of a consistent root of the likelihood equation holds in the multi-parameter case just as in the single-parameter case:

**Theorem 8.21** Suppose that the conditions of Theorem 8.19 are satisfied and that \( \hat{\theta} \) denotes a consistent root of the likelihood equations. Assume also that Equation (8.20) is satisfied for all \( i \) and \( j \) and that \( I(\theta^0) \) is positive definite with finite entries. Finally, assume that \( \partial^3 \ell(\theta) / \partial \theta_i \partial \theta_j \partial \theta_k \) exists and is bounded in a neighborhood of \( \theta^0 \) for all \( i, j, k \). Then
\[ \sqrt{n}(\hat{\theta} - \theta^0) \overset{d}{\to} N_k\{0, I^{-1}(\theta^0)\}. \]

As in the one-parameter case, there is some terminology associated with the derivatives of the loglikelihood function. The gradient vector \( \nabla \ell(\theta) \) is called the score vector. The negative second derivative \( -\nabla^2 \ell(\theta) \) is called the observed information, and \( nI(\theta) \) is sometimes called the expected information. And the second derivative of a real-valued function of a \( k \)-vector, such as the loglikelihood function \( \ell(\theta) \), is called the Hessian matrix.

Newton’s method (often called the Newton-Raphson method in the multivariate case) and scoring work just as they do in the one-parameter case. Starting from the point \( \tilde{\theta} \), one step
of Newton-Raphson gives

\[ \delta = \tilde{\theta} - \left\{ \nabla^2 \ell(\tilde{\theta}) \right\}^{-1} \nabla \ell(\tilde{\theta}) \]  

(8.23)

and one step of scoring gives

\[ \delta^* = \tilde{\theta} + \frac{1}{n} I^{-1}(\tilde{\theta}) \nabla \ell(\tilde{\theta}). \]  

(8.24)

**Theorem 8.22** Under the assumptions of Theorem 8.21, if \( \tilde{\theta} \) is a \( \sqrt{n} \)-consistent estimator of \( \theta^0 \), then the one-step Newton-Raphson estimator \( \delta \) defined in Equation (8.23) satisfies

\[ \sqrt{n}(\delta - \theta^0) \xrightarrow{d} N_k\{0, I^{-1}(\theta^0)\} \]

and the one-step scoring estimator \( \delta^* \) defined in Equation (8.24) satisfies

\[ \sqrt{n}(\delta^* - \theta^0) \xrightarrow{d} N_k\{0, I^{-1}(\theta^0)\}. \]

As in the one-parameter case, we define an efficient estimator \( \hat{\theta} \) as one that satisfies

\[ \sqrt{n}(\hat{\theta} - \theta^0) \xrightarrow{d} N_k\{0, I^{-1}(\theta^0)\}. \]

This definition is justified by the fact that if \( \delta \) is any estimator satisfying

\[ \sqrt{n}(\delta - \theta) \xrightarrow{d} N_k\{0, \Sigma(\theta)\}, \]

where \( I^{-1}(\theta) \) and \( \Sigma(\theta) \) are continuous, then \( \Sigma(\theta) - I^{-1}(\theta) \) is nonnegative definite for all \( \theta \).

Finally, we note that Bayes estimators are efficient, just as in the one-parameter case. This means that the same three types of estimators that are efficient in the one-parameter case — the consistent root of the likelihood equation, the one-step scoring and Newton-Raphson estimators, and Bayes estimators — are also efficient in the multiparameter case.

**Exercises for Section 8.4**

**Exercise 8.13** Let \( X \sim \text{multinomial}(1, p) \), where \( p \) is a \( k \)-vector for \( k > 2 \). Let \( p^* = (p_1, \ldots, p_{k-1}) \). Find \( I(p^*) \).
Exercise 8.14  Suppose that $\theta \in \mathbb{R} \times \mathbb{R}_+$ (that is, $\theta_1 \in \mathbb{R}$ and $\theta_2 \in (0, \infty)$) and

$$f_\theta(x) = \frac{1}{\theta_2} f \left( \frac{x - \theta_1}{\theta_2} \right)$$

for some continuous, differentiable density $f(x)$ that is symmetric about the origin. Find $I(\theta)$.

Exercise 8.15  Prove Theorem 8.21.

Hint:  Use Theorem 1.38.

Exercise 8.16  Prove Theorem 8.22.

Hint:  Use Theorem 1.38.

Exercise 8.17  The multivariate generalization of a beta distribution is a Dirichlet distribution, which is the natural prior distribution for a multinomial likelihood. If $p$ is a random $(k+1)$-vector constrained so that $p_i > 0$ for all $i$ and $\sum_{i=1}^{k+1} p_i = 1$, then $(p_1, \ldots, p_k)$ has a Dirichlet distribution with parameters $a_1 > 0, \ldots, a_{k+1} > 0$ if its density is proportional to

$$p_1^{a_1-1} \cdots p_k^{a_k-1} (1 - p_1 - \cdots - p_k)^{a_{k+1}-1} I \left\{ \min_i p_i > 0 \right\} I \left\{ \sum_{i=1}^{k} p_i < 1 \right\}.$$

Prove that if $G_1, \ldots, G_{k+1}$ are independent random variables with $G_i \sim \text{gamma}(a_i, 1)$, then

$$\frac{1}{G_1 + \cdots + G_{k+1}} (G_1, \ldots, G_k)$$

has a Dirichlet distribution with parameters $a_1, \ldots, a_{k+1}$.

8.5  Nuisance parameters

This section is the one intrinsically multivariate section in this chapter; it does not have an analogue in the one-parameter setting. Here we consider how efficiency of an estimator is affected by the presence of nuisance parameters.

Suppose $\theta$ is the parameter vector but $\theta_1$ is the only parameter of interest, so that $\theta_2, \ldots, \theta_{k}$ are nuisance parameters. We are interested in how the asymptotic precision with which we may estimate $\theta_1$ is influenced by the presence of the nuisance parameters. In other words, if
\( \hat{\theta} \) is efficient for \( \theta \), then how does \( \hat{\theta}_1 \) as an estimator of \( \theta_1 \) compare to an efficient estimator of \( \theta_1 \), say \( \theta^* \), under the assumption that all of the nuisance parameters are known?

Assume \( I(\theta) \) is positive definite. Let \( \sigma_{ij} \) denote the \((i,j)\) entry of \( I(\theta) \) and let \( \gamma_{ij} \) denote the \((i,j)\) entry of \( I^{-1}(\theta) \). If all of the nuisance parameters are known, then \( I(\hat{\theta}_1) = \sigma_{11} \), which means that the asymptotic variance of \( \sqrt{n}(\theta^* - \hat{\theta}_1) \) is \( 1/\sigma_{11} \). On the other hand, if the nuisance parameters are not known then the asymptotic variance of \( \sqrt{n}(\hat{\theta}_1 - \theta_1) \) is \( \gamma_{11} \). Of interest here is the comparison between \( \gamma_{11} \) and \( 1/\sigma_{11} \).

The following theorem may be interpreted to mean that the presence of nuisance parameters always increases the variance of an efficient estimator.

**Theorem 8.23** \( \gamma_{11} \geq 1/\sigma_{11} \), with equality if and only if \( \gamma_{12} = \cdots = \gamma_{1k} = 0 \).

**Proof:** Partition \( I(\theta) \) as follows:

\[
I(\theta) = \begin{pmatrix} \sigma_{11} & \rho^T \\ \rho & \Sigma \end{pmatrix},
\]

where \( \rho \) and \( \Sigma \) are \((k-1) \times 1\) and \((k-1) \times (k-1)\), respectively. Let \( \tau = \sigma_{11} - \rho^T \Sigma^{-1} \rho \). We may verify that if \( \tau > 0 \), then

\[
I^{-1}(\theta) = \frac{1}{\tau} \begin{pmatrix} 1 & -\rho^T \Sigma^{-1} \\ -\Sigma^{-1} \rho & \Sigma^{-1} + \tau \Sigma^{-1} \end{pmatrix}.
\]

This proves the result, because the positive definiteness of \( I(\theta) \) implies that \( \Sigma^{-1} \) is positive definite, which means that

\[
\gamma_{11} = \frac{1}{\sigma_{11} - \rho^T \Sigma^{-1} \rho} \geq \frac{1}{\sigma_{11}},
\]

with equality if and only if \( \rho = 0 \). Thus, it remains only to show that \( \tau > 0 \). But this is immediate from the positive definiteness of \( I(\theta) \), since if we set

\[
v = \begin{pmatrix} 1 \\ -\rho^T \Sigma^{-1} \end{pmatrix},
\]

then \( \tau = v^T I(\theta) v \).

The above result shows that it is important to take nuisance parameters into account in estimation. However, it is not necessary to estimate the entire parameter vector all at once, since \((\hat{\theta}_1, \ldots, \hat{\theta}_k)\) is efficient for \( \theta \) if and only if each of the \( \hat{\theta}_i \) is efficient for \( \theta_i \) in the presence of the other nuisance parameters (see problem 8.18).
Exercises for Section 8.5

Exercise 8.18  Letting $\gamma_{ij}$ denote the $(i, j)$ entry of $I^{-1}(\theta)$, we say that $\hat{\theta}_i$ is efficient for $\theta_i$ in the presence of the nuisance parameters $\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k$ if the asymptotic variance of $\sqrt{n}(\hat{\theta}_i - \theta_i)$ is $\gamma_{ii}$.

Prove that $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$ is efficient for $\theta$ if and only if for all $i$, the estimator $\hat{\theta}_i$ is efficient for $\theta_i$ in the presence of nuisance parameters $\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k$. 