26. Let $N_1, N_2, \ldots$, be independent Gaussian random variables with means 0 and variances 1, and let $S_1, S_2, \ldots$, be independent Gaussian random variables with means 0 and variances 3. Assume the N_k's and S_k's are independent of each other and consider the hypothesis pair

$$H_0 : Y_k = N_k, \quad k = 1, 2, \ldots$$

versus

$$H_1 : Y_k = S_k + N_k, \quad k = 1, 2, \ldots$$

(a) Repeat part (a) of Exercise 25 for this new model.

(b) Repeat part (b) of Exercise 25 for this new model.

27. Derive Eq. (III.E.22).

28. Show that Eqs. (III.E.24) and (III.E.25) are equivalent.

29. Derive Eq. (III.E.26).

30. Verify Eq. (III.E.54).

IV

Elements of Parameter Estimation

IV.A Introduction

In Chapters II and III we have considered the design of optimum procedures for deciding between two possible statistical situations on the basis of a random observation Y. In many situations arising in practice we are interested not in making a choice between two (or among several) discrete situations, but rather in making a choice among a continuum of possible states of nature. In particular, as in the composite hypothesis-testing problems discussed in Chapter II, we can think of a family of distributions on the observation space, indexed by a parameter or set of parameters. But unlike the case of composite hypothesis testing in which we wish to make a binary decision about the parameter, we wish here to determine as accurately as possible the actual value of the parameter from the observation.

Such problems are known as *parameter* (or *point*) *estimation problems*, and in this chapter we discuss the basic ideas relating to the design of optimum procedures for estimating parameters. As with the hypothesis-testing problem (which incidentally can be thought of as a special case of the parameter estimation problem), a variety of estimation design philosophies can be used, these differing primarily in the amount of prior information known about the parameter and in the performance criteria applied.

In this chapter we discuss two basic approaches to parameter estimation—one, the Bayesian, in which the parameter is assumed to be a random quantity related statistically to the observation, and a second in which the parameter is assumed to be unknown but without being endowed with any probabilistic structure. Of these two approaches, the Bayesian is the most straightforward and so is considered first, in Section IV.B, with nonrandom parameter estimation being considered in the remainder of the chapter.

It should be noted that in this treatment, we consider only the estimation of parameters that are static, i.e., that are constant in time. The estimation of dynamic parameters (i.e., signals) is considered in Chapter V.

IV.B Bayesian Parameter Estimation

Throughout this chapter we assume as a model a family of distributions for the random observation Y, indexed by a parameter θ taking values in a parameter set Λ; i.e., we have the family $\{P_\theta; \theta \in \Lambda\}$, where P_θ denotes a distribution on the observation space $(\Gamma, \mathcal{G})$. We also assume that the parameter set Λ is a subset of $\mathbb{R}^m$ for some m. Within this model the goal of the parameter estimation problem is to find a function $\hat{\theta}$: $\Gamma \to \Lambda$ such that $\hat{\theta}(y)$ is the "best" guess of the true value of θ (i.e., the value of θ for which $Y \sim P_\theta$) based on the observation $Y = y$.

Of course, the solution to this problem depends on the criterion of goodness by which we measure estimation performance; so, as in the hypothesis-testing problem, we begin by assigning costs to our decisions about the parameter. In particular, we suppose that there is a function C: $\Lambda \times \Lambda \to \mathbb{R}$ such that $C[a, \theta]$ is the cost of estimating a true value of θ as a, for a and θ in Λ. Given such a function C we can then associate with an estimator $\hat{\theta}$ a conditional risk or cost averaged over Y for each $\theta \in \Lambda$; i.e., we have

$$R_\theta(\hat{\theta}) = E_\theta\{C[\hat{\theta}(Y), \theta]\}. \tag{IV.B.1}$$

If we now adopt the interpretation that the actual parameter value θ is the realization of a random variable Θ, we can define an average or Bayes risk as

$$r(\hat{\theta}) \triangleq E\{R_\Theta(\hat{\theta})\}, \tag{IV.B.2}$$

and the appropriate design goal is to find an estimator minimizing $r(\hat{\theta})$. Such an estimator is known as a *Bayes estimate* of θ.

Noting that $R_\theta(\hat{\theta}) = E\{C[\hat{\theta}(Y), \Theta] | \Theta = \theta\}$, we have

$$r(\hat{\theta}) = E\{C[\hat{\theta}(Y), \Theta]\} = E\{E\{C[\hat{\theta}(Y), \Theta]|Y\}\}. \tag{IV.B.3}$$

By inspection of (IV.B.3) we see that the Bayes estimate of θ can be found (if it exists) by minimizing, for each $y \in \Gamma$, the posterior cost given $Y = y$:

$$E\{C[\hat{\theta}(y), \Theta]|Y = y\}. \tag{IV.B.4}$$

This is the same procedure as that followed in the Bayesian hypothesis-testing problem (see Section II.E). Note that if we assume that Θ has a conditional density $w(\theta|y)$ given $Y = y$ for each $y \in \Gamma$, then the Bayes estimate $\hat{\theta}(y)$ corresponding to $y \in \Gamma$ can be sought by minimizing

$$\int_\Lambda C[\hat{\theta}(y), \theta] w(\theta|y) \mu(d\theta). \tag{IV.B.5}$$

The following cases illustrate the application of this criterion.

Case IV.B.1: Minimum-Mean-Squared-Error (MMSE) Estimation

For situations in which $\Lambda = \mathbb{R}$ and $E\{\Theta^2\} < \infty$, a commonly used cost function is that given by

$$C[a, \theta] = (a - \theta)^2, \quad (a, \theta) \in \mathbb{R}^2. \tag{IV.B.6}$$

This cost function is a natural one for many situations since it measures the performance of an estimator in terms of the square of the estimation error, $\hat{\theta}(y) - \theta$. The Bayes risk here is $E\{(\hat{\theta}(Y) - \Theta)^2\}$, a quantity known as the *mean-squared error* (MSE). Thus the Bayes estimate in this case is a *minimum-mean-squared-error* (MMSE) *estimator.*

The posterior cost given $Y = y$ is given in this case by

$$\begin{aligned} E\{(\hat{\theta}(y) - \Theta)^2|Y = y\} &= E\{[\hat{\theta}(y)]^2|Y = y\} \\ &\quad - 2E\{\hat{\theta}(y)\Theta|Y = y\} \\ &\quad + E\{\Theta^2|Y = y\} \\ &= [\hat{\theta}(y)]^2 - 2\hat{\theta}(y)E\{\Theta|Y = y\} \\ &\quad + E\{\Theta^2|Y = y\}. \end{aligned} \tag{IV.B.7}$$

The expression in (IV.B.7) is a quadratic function of $\hat{\theta}(y)$, so it achieves its unique minimum at the point where its derivative with respect to $\hat{\theta}(y)$ is zero. On differentiating (IV.B.7) we have that the Bayes estimate, denoted by $\hat{\theta}_{MMSE}$, is given by

$$\hat{\theta}_{MMSE}(y) = E\{\Theta|Y = y\}. \tag{IV.B.8}$$

Thus the MMSE estimate of Θ given $Y = y$ is the conditional mean of Θ given $Y = y$. This is a very basic result to which we will return in subsequent chapters. This estimate is sometimes termed the *conditional mean estimate* (CME).

Case IV.B.2: Minimum-Mean-Absolute-Error (MMAE) Estimation

Another cost function that is sometimes applied in the case $\Lambda = \mathbb{R}$ is the *absolute error*, given by

$$C[a, \theta] = |a - \theta|, \quad (a, \theta) \in \mathbb{R}^2. \tag{IV.B.9}$$

The Bayes risk here is $E\{|\hat{\theta}(Y) - \Theta|\}$, a quantity known as the *mean-absolute error*, so the corresponding Bayes estimate is known as the *minimum-mean-absolute-error* (MMAE) *estimate.*

To derive the MMAE estimate we make use of the fact that if X is a random variable with $P(X \geq 0) = 1$, then $E\{X\} = \int_0^\infty P(X > x)dx$. This result follows essentially by integrating by parts [see, e.g., Breiman (1968)].

Since $|\hat{\theta}(y) - \Theta| \geq 0$, we have from the result above that

$$\begin{aligned} E\{|\hat{\theta}(y) - \Theta| \, | Y = y\} &= \int_0^\infty P(|\hat{\theta}(y) - \Theta| > x | Y = y) dx \\ &= \int_0^\infty P(\Theta > x + \hat{\theta}(y) | Y = y) dx \\ &\quad + \int_0^\infty P(\Theta < -x + \hat{\theta}(y) | Y = y) dx. \end{aligned} \tag{IV.B.10}$$

Substituting $t = x + \hat{\theta}(y)$ in the first integral and $t = -x + \hat{\theta}(y)$ in the second integral on the right of (IV.B.10), we have

$$\begin{aligned} E\{|\hat{\theta}(y) - \Theta| \, | Y = y\} &= \int_{\hat{\theta}(y)}^\infty P(\Theta > t | Y = y) dt \\ &\quad + \int_{-\infty}^{\hat{\theta}(y)} P(\Theta < t | Y = y) dt. \end{aligned} \tag{IV.B.11}$$

With $E\{|\hat{\theta}(y) - \Theta| \, | Y = y\}$ in the form (IV.B.11) we see that it is a differentiable function of $\hat{\theta}(y)$. On differentiating we get

$$\begin{aligned} \frac{\partial}{\partial \hat{\theta}(y)} E\{|\hat{\theta}(y) - \Theta| \, | Y = y\} &= P(\Theta < \hat{\theta}(y) | Y = y) \\ &\quad - P(\Theta > \hat{\theta}(y) | Y = y). \end{aligned} \tag{IV.B.12}$$

From (IV.B.12) we note that this derivative is a nondecreasing function of $\hat{\theta}(y)$ that approaches -1 as $\hat{\theta}(y) \to -\infty$ and $+1$ as $\hat{\theta}(y) \to +\infty$. Thus $E\{|\hat{\theta}(y) - \Theta| \, | Y = y\}$ achieves its minimum over $\hat{\theta}(y)$ at the point (or on the set of points) where its derivative changes sign. That is, the Bayes estimate in this case, denoted by $\hat{\theta}_{ABS}(y)$, is any point such that

$$P(\Theta < t | Y = y) \leq P(\Theta > t | Y = y), \quad t < \hat{\theta}_{ABS}(y)$$

and

$$(\Theta < t | Y = y) \geq P(\Theta > t | Y = y), \quad t > \hat{\theta}_{ABS}(y). \tag{IV.B.13}$$

Note that a point $\hat{\theta}_{ABS}(y)$ satisfying (IV.B.13) is a *median* of the conditional distribution of Θ given $Y = y$. Thus the MMAE estimate is a *conditional median estimate*. This estimate coincides with the MMSE estimate only when the distribution of Θ given $Y = y$ has the same value

for its mean and median. Which of these two is the "better" estimate of Θ depends, of course, on which criterion one adopts.

Case IV.B.3: Maximum *A Posteriori* Probability (MAP) Estimation

Another estimation method that, although not properly a Bayes estimate, fits within the Bayesian framework is maximum *a posteriori* probability (MAP) estimation.

To motivate this method, we assume the case $\Lambda = \mathbb{R}$ and consider the so-called *uniform cost function*,

$$C[a, \theta] = \begin{cases} 0 & \text{if } |a - \theta| \leq \Delta \\ 1 & \text{if } |a - \theta| > \Delta, \end{cases} \tag{IV.B.14}$$

where $\Delta > 0$. For an estimator $\hat{\theta}$ the average posterior cost given $Y = y$ in this case is given by

$$\begin{aligned} E\{C[\hat{\theta}(y), \Theta] | Y = y\} &= P(|\hat{\theta}(y) - \Theta| > \Delta | Y = y) \\ &= 1 - P(|\hat{\theta}(y) - \Theta| \leq \Delta | Y = y). \end{aligned} \tag{IV.B.15}$$

In considering the minimization of (IV.B.15) suppose first that Θ is a discrete random variable taking values in a finite set $\Lambda = \{\theta_0, \ldots, \theta_{M-1}\}$ with $|\theta_i - \theta_j| > \Delta$ for $i \neq j$. Then we have

$$\begin{aligned} E\{C[\hat{\theta}(y), \Theta] | Y = y\} &= 1 - P(\Theta = \hat{\theta}(y) | Y = y) \\ &= 1 - w(\hat{\theta}(y) | y) \quad \text{for } \hat{\theta}(y) \in \Lambda, \end{aligned} \tag{IV.B.16}$$

where $w(\theta | y)$ is the conditional probability mass function of Θ given $Y = y$. We see from (IV.B.16) that the Bayes estimate in this case is given for each $y \in \Gamma$ by any value of θ that maximizes $w(\theta | y)$ over $\theta \in \Lambda$. That is, the Bayes estimate is the value of Θ that has the maximum *a posteriori* probability of occurring given $Y = y$.[1]

Now suppose that $\Lambda = \mathbb{R}$ and Θ is a continuous random variable with conditional density function $w(\theta | y)$ given $Y = y$. In this case the posterior

[1] Of course, this case in which Λ is a finite set is simply an M-ary hypothesis testing problem, and the cost criterion of (IV.B.14) reduces here to $C[a, \theta] = 1$ if $a \neq \theta$ and $C[a, \theta] = 0$ if $a = \theta$, since $|\theta_i - \theta_j| > \Delta$ for $i \neq j$. The Bayes estimate in this case is thus the M-ary Bayes decision rule for uniform cost (as in Exercise 16 of Chapter II).

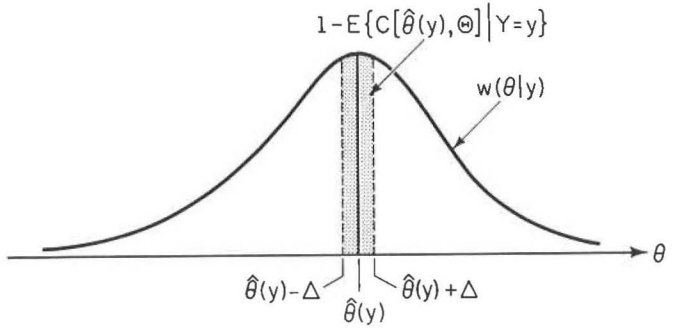

FIGURE IV.B.1. Illustration of MAP estimation.

risk becomes

$$E\{C[\hat{\theta}(y), \Theta]|Y = y\} = 1 - \int_{\hat{\theta}(y)-\Delta}^{\hat{\theta}(y)+\Delta} w(\theta|y)d\theta. \tag{IV.B.17}$$

The quantity in (IV.B.17) is minimized over $\hat{\theta}(y)$ by maximizing the area under $w(\theta|y)$ over the interval $(\hat{\theta}(y)-\Delta, \hat{\theta}(y)+\Delta)$. Referring to Fig. IV.B.1 we see that if $w(\theta|y)$ is a smooth function of θ and if Δ is sufficiently small, this area will be approximately maximized by choosing $\hat{\theta}(y)$ to be a point of maximum of $w(\theta|y)$. That is, for small Δ and smooth $w(\theta|y)$, we have

$$\int_{\hat{\theta}(y)-\Delta}^{\hat{\theta}(y)+\Delta} w(\theta|y)d\theta \cong 2\Delta w(\theta|y)|_{\theta=\hat{\theta}(y)}, \tag{IV.B.18}$$

and the right-hand side is maximized by choosing $\hat{\theta}(y)$ to be the value of θ maximizing $w(\theta|y)$ over Λ.

In either of the cases above, the uniform cost criterion leads to the procedure for estimating Θ as that value maximizing the *a posteriori* (discrete or continuous) density $w(\theta|y)$. [Similarly, with θ discrete but taking on infinitely many values, it can be argued that (IV.B.15) is minimized approximately by choosing $\hat{\theta}(y)$ to maximize the conditional mass function $w(\theta|y)$.] This estimate is known as the *maximum a posteriori probability* (MAP) *estimate* and is denoted by $\hat{\theta}_{MAP}$. Although this estimate often only approximates the Bayes estimate for uniform cost with small Δ, the MAP criterion is widely used to design estimates. A principal reason for this is that MAP estimates are often easier to compute than MMSE, MMAE, or other estimates.

Note that a point at which a density achieves its maximum value is termed a *mode* of the corresponding probability distribution. Thus since

$\hat{\theta}_{MAP}$ estimates Θ by the mode of its conditional distribution, it is a *conditional mode estimate.*

From Cases IV.B.1 through IV.B.3 [and from (IV.B.5)] we see that Bayes estimates for a given situation are determined from the conditional distribution of the parameter given the observations. In particular, the MMSE, MMAE, and MAP estimates are the mean, median, and mode of this distribution, respectively. As in the case of hypothesis testing, we can think of the observation as a means for converting the prior distribution of the parameter into a posterior distribution. In general, Bayes estimators are features of this posterior distribution.

In modeling a given statistical situation we usually start with the family $\{P_\theta; \theta \in \Lambda\}$ of conditional distributions of Y given $\Theta = \theta$, and for the Bayesian formulation we also have a prior distribution for Θ. To obtain the conditional distribution of Θ given Y from the prior and the conditional of Y given Θ we need only to apply Bayes' formula. In particular, supposing that P_θ has density p_θ for each $\theta \in \Lambda$ and that the prior distribution of Θ has density $w(\theta)$, we have that the conditional distribution of Θ given $Y = y$ has density

$$w(\theta|y) = \frac{p_\theta(y)w(\theta)}{\int_\Lambda p_\theta(y)w(\theta)\mu(d\theta)}. \tag{IV.B.19}$$

Note that the denominator of (IV.B.19) is $p(y)$, the unconditioned density of Y.

The Bayes estimates for the three cases above can be obtained straightforwardly from (IV.B.19). Note that the MAP estimate can be obtained without the computation of $p(y)$ since this term will not affect the maximization over θ. That is, $\hat{\theta}_{MAP}(y)$ is found by maximizing $p_\theta(y)w(\theta)$ over $\theta \in \Lambda$. Since the logarithm is an increasing function, $\hat{\theta}_{MAP}(y)$ also maximizes $[\log p_\theta(y) + \log w(\theta)]$ over $\theta \in \Lambda$. If Θ is a continuous random variable given $Y = y$, then for sufficiently smooth p_θ and w, a necessary condition for this maximization is

$$\frac{\partial}{\partial\theta} \log p_\theta(y)|_{\theta=\hat{\theta}_{MAP}(y)} = -\frac{\partial}{\partial\theta} \log w(\theta)|_{\theta=\hat{\theta}_{MAP}(y)}. \tag{IV.B.20}$$

Equation (IV.B.20) is known as the *MAP equation.*

The following two examples serve to illustrate the computation of the MMSE, MMAE, and MAP estimates.

Example IV.B.1: Estimation of the Parameter of an Exponential Distribution

Consider the situation $\Lambda = (0, \infty)$ and $\Gamma = \mathbb{R}$, in which the observations have the following conditional probability density function given $\Theta = \theta$:

$$p_\theta(y) = \begin{cases} \theta e^{-\theta y} & \text{if } y \geq 0 \\ 0 & \text{if } y < 0. \end{cases} \tag{IV.B.21}$$

This is the *exponential density* with parameter θ. The exponential density models many physical phenomena. It is particularly useful in modeling the time intervals between successive events occurring randomly in time, such as messages or data packets arriving at a communications switching station, vehicles arriving at an intersection of roads, photons emitting from a coherent light source, or devices failing in a logic circuit. The parameter θ in this model can be interpreted as the rate of such occurrences, and thus we can think of the estimation problem here as that of estimating the rate of occurrences of such events from an observation of the time between successive occurrences of them.

Suppose that our prior information about Θ is that it also has an exponential distribution with density

$$w(\theta) = \begin{cases} \alpha e^{-\alpha\theta} & \text{if } \theta \geq 0 \\ 0 & \text{if } \theta < 0, \end{cases} \tag{IV.B.22}$$

where $\alpha > 0$ is known. We can then find the posterior distribution of Θ given $Y = y$ from (IV.B.19). We have

$$\begin{aligned} w(\theta|y) &= \frac{\alpha\theta e^{-(\alpha+y)\theta}}{\int_0^\infty \alpha\theta e^{-(\alpha+y)\theta} d\theta} \\ &= (\alpha+y)^2 \theta e^{-\theta(\alpha+y)}, \end{aligned} \tag{IV.B.23}$$

for $\theta \geq 0$ and $y \geq 0$, and $w(\theta|y) = 0$ otherwise.

The MMSE estimate is the mean of (IV.B.23) and thus is given by

$$\begin{aligned} \hat{\theta}_{\text{MMSE}}(y) &= \int_0^\infty \theta w(\theta|y) d\theta = (\alpha+y)^2 \int_0^\infty \theta^2 e^{-\theta(\alpha+y)} d\theta \\ &= \frac{2}{\alpha+y}. \end{aligned} \tag{IV.B.24}$$

Note that for fixed α, this estimate of Θ varies inversely with y. This is intuitively reasonable from the foregoing interpretation of the exponential model since a large interarrival time (large y) would be evidence of a low rate (small θ). This behavior is tempered to a degree depending on the value of α since the estimate is never greater than $2/\alpha$. Note that a small value of α implies that Θ is distributed diffusely [i.e., $w(\theta)$ is relatively spread out] and the corresponding estimate allows larger values of Θ if implied by the observation. Alternatively, a large value of α implies that Θ is close to zero with high probability, so the estimate is never large in this case.

The minimum value of the MSE can be computed straightforwardly in this case. First, we note from (IV.B.3) that the Bayes risk is the average of the posterior cost, so that

$$\begin{aligned} \text{MMSE} = r(\hat{\theta}_{\text{MMSE}}) &= E\{E\{(\hat{\theta}_{\text{MMSE}}(Y) - \Theta)^2|Y\}\} \\ &= E\{E\{(\Theta - E\{\Theta|Y\})^2|Y\}\} \\ &= E\{\text{Var}(\Theta|Y)\}. \end{aligned} \tag{IV.B.25}$$

Thus the minimum MSE is the average of the conditional variance of Θ given Y. Since

$$\text{Var}(\Theta|Y=y) = E\{\Theta^2|Y=y\} - E^2\{\Theta|Y=y\},$$

we have

$$\begin{aligned} \text{Var}(\Theta|Y=y) &= \int_0^\infty \theta^2 w(\theta|y) d\theta - [\hat{\theta}_{\text{MMSE}}(y)]^2 \\ &= (\alpha+y)^2 \int_0^\infty \theta^3 e^{-\theta(\alpha+y)} d\theta - \frac{4}{(\alpha+y)^2} \\ &= \frac{2}{(\alpha+y)^2}. \end{aligned}$$

Thus

$$\begin{aligned} \text{MMSE} = E\left\{\frac{2}{(\alpha+Y)^2}\right\} &= \int_0^\infty \frac{2}{(\alpha+y)^2} p(y) dy \\ &= \int_0^\infty \frac{2\alpha}{(\alpha+y)^4} dy \\ &= \frac{2}{3\alpha^2}, \end{aligned} \tag{IV.B.26}$$

where we have used $p(y) = \int_0^\infty \alpha\theta e^{-(\alpha+y)\theta} d\theta = \alpha/(\alpha+y)^2$, as in (IV.B.23).

The MMAE estimate, $\hat{\theta}_{ABS}(y)$, is the median of $w(\theta|y)$. Since Θ is continuous given $Y = y$, we can find $\hat{\theta}_{ABS}(y)$ by solving the equation

$$\int_{\hat{\theta}_{ABS}(y)}^\infty w(\theta|y) d\theta = \tfrac{1}{2}. \tag{IV.B.27}$$

Inserting (IV.B.23) and integrating yields

$$[1 + (\alpha+y)\hat{\theta}_{ABS}(y)] e^{-(\alpha+y)\hat{\theta}_{ABS}(y)} = \tfrac{1}{2}, \tag{IV.B.28}$$

so that we have

$$\hat{\theta}_{ABS}(y) = \frac{T_o}{\alpha+y}, \tag{IV.B.29}$$

with T_o being the solution to $(1+T_o)e^{-T_o} = 1/2$, which is given by $T_o \cong 1.68$. Comparing (IV.B.29) with (IV.B.24), we see the same general

behavior as the MMSE estimate, these differing only in the constant in the numerator. The minimum Bayes risk for this situation can be computed similarly to that for the MMSE estimate, and this computation is left as an exercise.

The MAP estimate of Θ can also be obtained easily in this case. Noting that

$$\begin{aligned}\frac{\partial}{\partial\theta}[\log p_\theta(y) + \log w(\theta)] &= \frac{\partial}{\partial\theta}(\log\theta - \theta y + \log\alpha - \alpha\theta)\\ &= \theta^{-1} - (\alpha + y)\end{aligned}$$

and

$$\frac{\partial^2}{\partial\theta^2}[\log p_\theta(y) + \log w(\theta)] = -\theta^{-2} < 0,$$

we see that $w(\theta|y)$ has its unique maximum at

$$\hat{\theta}_{MAP}(y) = \frac{1}{\alpha + y}. \tag{IV.B.30}$$

Thus we again get an estimate differing from the MMSE estimate by only a scale factor.

In Example IV.B.1 the three estimation criteria considered lead to three different estimators for Θ. To decide which one to use, one must decide which of the three corresponding cost functions penalizes the estimation error in the way most suitable for the application of interest. In many problems of interest one does not need to make such a choice because the three estimates coincide. The following is an example of such a situation.

Example IV.B.2: Estimation of Signal Amplitude

Consider the case $\Gamma = \mathbb{R}^n$ and $\Lambda = \mathbb{R}$ with

$$Y_k = N_k + \Theta s_k, \quad k = 1, \ldots, n, \tag{IV.B.31}$$

where $\underline{N} \sim \mathcal{N}(0, \Sigma)$, $\underline{s}$ is known, $\Theta \sim \mathcal{N}(\mu, v^2)$, and $\underline{N}$ and Θ are independent. Note that this problem corresponds to the estimation of the unknown amplitude of an otherwise known signal observed in the presence of additive noise.

Given $\Theta = \theta$, we have that $\underline{Y} \sim \mathcal{N}(\theta\underline{s}, \Sigma)$. Thus the posterior density for Θ is

$$w(\theta|\underline{y}) =$$

$$\frac{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-(1/2)(\underline{y}-\theta\underline{s})^T\Sigma^{-1}(\underline{y}-\theta\underline{s})} \frac{1}{\sqrt{2\pi}v} e^{-(\theta-\mu)^2/2v^2}}{\int_{-\infty}^{\infty} \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-(1/2)(\underline{y}-\theta\underline{s})^T\Sigma^{-1}(\underline{y}-\theta\underline{s})} \frac{1}{\sqrt{2\pi}v} e^{-(\theta-\mu)^2/2v^2} d\theta}$$

$$= K(\underline{y}) \exp\left\{-\frac{\theta^2}{2}(d^2 + 1/v^2) + \theta(\underline{s}^T\Sigma^{-1}\underline{y} + \frac{\mu}{v^2})\right\}, \tag{IV.B.32}$$

where as in Chapter III we define $d^2 = \underline{s}^T\Sigma^{-1}\underline{s}$ and where $K(\underline{y})$ is a function depending on $\underline{y}$ but not on θ. Note from (IV.B.32) that $w(\theta|\underline{y})$ is the exponential of a quadratic term in θ, so that it must be a Gaussian density. If $w(\theta|\underline{y})$ were $\mathcal{N}(m, q^2)$, we would have

$$\begin{aligned}w(\theta|\underline{y}) &= \frac{1}{\sqrt{2\pi}q} e^{-(\theta-m)^2/2q^2}\\ &= \frac{e^{-m^2/2q^2}}{\sqrt{2\pi}q} e^{-\theta^2/2q^2 + \theta m/q^2}.\end{aligned} \tag{IV.B.33}$$

Comparing (IV.B.32) with (IV.B.33), we see that given $\underline{Y} = \underline{y}, \Theta \sim \mathcal{N}(m, q^2)$ with

$$q^2 = (d^2 + 1/v^2)^{-1}$$

and

$$m = (d^2 + 1/v^2)^{-1}(\underline{s}^T\Sigma^{-1}\underline{y} + \mu/v^2),$$

and $K(\underline{y})$ becomes $e^{-m^2/2q^2}/\sqrt{2\pi}q$.

Since the first parameter of the Gaussian density is its mean, we immediately have that the conditional mean estimate of Θ is

$$\begin{aligned}\hat{\theta}_{MMSE}(\underline{y}) &= \frac{\underline{s}^T\Sigma^{-1}\underline{y} + \mu/v^2}{d^2 + 1/v^2}\\ &= \frac{v^2d^2\hat{\theta}_1(\underline{y}) + \mu}{v^2d^2 + 1},\end{aligned} \tag{IV.B.34}$$

where $\hat{\theta}_1(\underline{y}) \triangleq \underline{s}^T\Sigma^{-1}\underline{y}/d^2$. Moreover, the minimum-mean-squared error is

$$\text{MMSE} = E\{\text{Var}(\Theta|\underline{Y})\} = \frac{1}{d^2 + 1/v^2} = \frac{v^2}{v^2d^2 + 1}, \tag{IV.B.35}$$

since $\text{Var}(\Theta|\underline{Y}) = (d^2 + 1/v^2)^{-1}$ which does not depend on $\underline{Y}$. Also, since the Gaussian density is symmetric about its mean and it achieves its maximum at its mean, the conditional median and conditional mode both equal the conditional mean; i.e., we have $\hat{\theta}_{\text{ABS}} = \hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{MMSE}}$.

The behavior of this estimate well illustrates the nature of Bayesian estimation. Note that v^2 determines the accuracy of our prior knowledge about Θ; that is, the smaller v^2 is, the more accurately we know θ in the absence of observations. On the other hand, in view of the discussion of coherent detection in Gaussian noise in Chapter III, the quantity d^2 is a measure of the quality with which $\underline{s}$ can be distinguished from the $\mathcal{N}(0, \mathbf{\Sigma})$ noise. That is, d^2 is a measure of the accuracy of our observations in terms of producing information about the signal—large d^2 corresponds to high-quality observations and small d^2 to low-quality observations in this sense.

With these ideas in mind, consider the estimate $\hat{\theta}_{\text{MMSE}}$ of (IV.B.34). If v^2d^2 is very small relative to the other quantities in this estimate, we have $\hat{\theta}_{\text{MMSE}}(\underline{y}) \cong \mu$. This occurs when the prior knowledge is very accurate relative to the observations (i.e., v^2 is small relative to $1/d^2$), so the estimator ignores the observations and chooses the mean of the prior distribution as its estimate. Note that the MMSE in this case is approximately v^2, the prior variance. On the other hand, if v^2d^2 is large, then $\hat{\theta}_{\text{MMSE}}(\underline{y}) \cong \hat{\theta}_1(\underline{y})$, an estimate that depends only on the observations and does not incorporate the prior information at all. The latter situation is also reasonable since with v^2 large relative to $1/d^2$, we are better off trusting the observations rather than the prior information. The MMSE in the latter case is approximately $1/d^2$.[2] Between these two extremes the optimum estimator balances the prior knowledge and the observations, and the corresponding MMSE reflects this balance.

It is interesting to consider the particular case $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ and $\underline{s} = \underline{1} \triangleq (1, 1, \ldots, 1)^T$ in the context of the discussion above. In this case our observations are

$$Y_k = N_k + \Theta, \quad k = 1, \ldots, n,$$

with $N_1, \ldots, N_n$ i.i.d. $\mathcal{N}(0, \sigma^2)$. The quantity $v^2d^2 = nv^2/\sigma^2$ and $\hat{\theta}_1(\underline{y}) = \bar{y} \triangleq (1/n)\sum_{k=1}^{n} y_k$, the sample mean. If we have no observations ($n = 0$), we simply estimate Θ as its prior mean μ, but as we take more observations (increase n) the sample mean $\bar{y}$ becomes more reliable and we place more weight on it. In the limit as $n \to \infty$ we disregard the prior mean entirely and adopt the sample mean as our estimate. The scale of this behavior is controlled by the ratio v^2/σ^2 (note that σ^2 determines the accuracy of each observation).

[2] Note that in the absence of any observations the MMSE estimate is μ and the MMSE is v^2, which corresponds to the approximate conditions for v^2d^2 small. It turns out (as we shall see in the following sections) that the estimate $\hat{\theta}_1(\underline{y})$ and the accuracy $1/d^2$ are optimum in the absence of any prior information. Thus these two extremes are quite reasonable.

In the discussion above we have concentrated on the estimation of a single real parameter. However, in many problems arising in practice we wish to estimate several parameters simultaneously. The Bayesian formulation, of course, applies equally well to the vector-parameter situation, and in the following discussion we treat this case.

Case IV.B.4: Estimation of Vector Parameters

We consider now the case in which $\Lambda = \mathbb{R}^m$. To follow the Bayesian procedure for designing an estimate of $\underline{\Theta}$ we must specify a cost function $C : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$. It is sometimes meaningful to use a cost function of the form

$$C[\underline{a}, \underline{\theta}] = \sum_{i=1}^{m} C_i[a_i, \theta_i], \tag{IV.B.36}$$

where C_i is a cost function associated with the estimation of the ith component of the parameter. If we have a cost function of this form, the conditional posterior cost for an estimate $\underline{\hat{\theta}}$ is given by

$$E\{C[\underline{\hat{\theta}}(y), \underline{\Theta}]|Y = y\} = \sum_{i=1}^{m} E\{C_i[\hat{\theta}_i(y), \Theta_i]|Y = y\}, \tag{IV.B.37}$$

so that we essentially have m scalar estimation problems to solve. That is, $\hat{\theta}_i(y)$ [the ith component $\hat{\theta}(y)$] is chosen to minimize $E\{C_i[\hat{\theta}_i(y), \Theta_i]|Y = y\}$.

An example of a useful cost function that decomposes as in (IV.B.36) is the square of the Euclidean norm of the error:

$$C[\underline{a}, \underline{\theta}] = \| \underline{a} - \underline{\theta} \|^2 = \sum_{i=1}^{m} (a_i - \theta_i)^2. \tag{IV.B.38}$$

It follows from Case IV.B.1 that for this cost function the ith component of the Bayes estimate is $E\{\Theta_i|Y = y\}$; i.e., the Bayes estimate is

$$\underline{\hat{\theta}}_B(y) = E\{\underline{\Theta}|Y = y\}, \tag{IV.B.39}$$

the conditional mean of $\underline{\Theta}$ given $Y = y$.

Another example of a cost function satisfying (IV.B.36) is the following

$$C[\underline{a}, \underline{\theta}] = \sum_{i=1}^{m} |a_i - \theta_i|. \tag{IV.B.40}$$

This function provides an alternative to $\| \underline{a} - \underline{\theta} \|$ as a measure of the distance between $\underline{a}$ and $\underline{\theta}$. From Case IV.B.2 we see that this cost function leads to the estimate whose ith component is the conditional median of Θ_i given $Y = y$.

To extend the concept of MAP estimation to vector parameters we might consider a cost function of the form (IV.B.36), in which $C_i[a_i, \theta_i]$ is the uniform cost function of (IV.B.14). This leads to the vector estimator that has as its ith component the conditional mode of Θ_i given $Y = y$. However, this decomposed cost function is not the most meaningful extension of the uniform cost function to the vector case. More meaningful is

$$C[\underline{a}, \underline{\theta}] = \begin{cases} 1 & \text{if } \max_{1\le i\le m} |a_i - \theta_i| > \Delta \\ 0 & \text{if } \max_{1\le i\le m} |a_i - \theta_i| \le \Delta, \end{cases} \tag{IV.B.41}$$

for which we have

$$E\{C[\hat{\underline{\theta}}(y), \underline{\Theta}]|Y = y\} = 1 - P(|\hat{\theta}_1(Y) - \Theta_1| \le \Delta, \ldots, |\hat{\theta}_m(Y) - \Theta_m| \le \Delta | Y = y). \tag{IV.B.42}$$

From (IV.B.42) we can argue the approximate optimality of estimating $\underline{\Theta}$ as its conditional mode given $Y = y$, a quantity that differs in general from the vector whose ith component is the conditional mode of Θ_i given $Y = y$ obtained from decomposing the cost. The estimate that chooses the conditional mode of $\underline{\Theta}$ given $Y = y$ is the MAP estimate for the vector-parameter case. Note that the region where $\max_{1\le i\le m} |a_i - \theta_i| \le \Delta$ is an m-dimensional cube centered at $\underline{\theta}$ with side length 2Δ. We could define similar cost functions by replacing this cube with other shapes (e.g., an m-dimensional ball, $\| \underline{a} - \underline{\theta} \| \le \Delta$); however, the approximate optimality of the MAP estimate would still be implied within the appropriate smoothness conditions.

A further useful cost function of interest in estimating vector parameters is a generalization of the squared-error norm. In particular, it is of interest to consider cost functions of the form

$$C[\underline{a}, \underline{\theta}] = (\underline{a} - \underline{\theta})^T \mathbf{A} (\underline{a} - \underline{\theta}), \tag{IV.B.43}$$

where $\mathbf{A}$ is a symmetric, positive-definite matrix. Note that this cost function allows for joint weightings of errors in different parameters, a desirable feature for some applications since the accuracy of our estimate of one of the real parameters forming $\underline{\Theta}$ may have an impact on how well we need to know other parameters.

To derive the Bayes estimate for (IV.B.43), we write

$$\begin{aligned} &E\{(\hat{\underline{\theta}}(y) - \underline{\Theta})^T \mathbf{A} (\hat{\underline{\theta}}(y) - \underline{\Theta}) | Y = y\} \\ &\quad = [\hat{\underline{\theta}}(y)]^T \mathbf{A} \hat{\underline{\theta}}(y) - 2[\hat{\underline{\theta}}(y)]^T \mathbf{A} E\{\underline{\Theta}|Y = y\} \\ &\quad + E\{\underline{\Theta}^T \mathbf{A} \Theta | Y = y\}. \end{aligned} \tag{IV.B.44}$$

Since the function of (IV.B.44) is quadratic in $\hat{\underline{\theta}}(y)$, it achieves its minimum at the point at which its gradient with respect to $\hat{\underline{\theta}}(y)$ vanishes. We have straightforwardly that

$$\nabla_{\hat{\underline{\theta}}(y)} E\{C[\hat{\underline{\theta}}(y), \underline{\Theta}]|Y = y\} = 2\mathbf{A}\hat{\underline{\theta}}(y) - 2\mathbf{A}E\{\underline{\Theta}|Y = y\}. \tag{IV.B.45}$$

So the Bayes estimate, $\hat{\underline{\theta}}_B$, for (IV.B.43) satisfies

$$2\mathbf{A}\hat{\underline{\theta}}_B(y) = 2\mathbf{A}E\{\underline{\Theta}|Y = y\}. \tag{IV.B.46}$$

Premultiplying (IV.B.46) by $(1/2)\mathbf{A}^{-1}$ yields that $\hat{\underline{\theta}}_B(y) = E\{\underline{\Theta}|Y = y\}$.

Thus we see that the quadratic cost criterion of (IV.B.43) yields the conditional mean vector as a Bayes estimate regardless of choice of $\mathbf{A}$. The resulting Bayes risk of course does depend on $\mathbf{A}$ and it is straightforward to show (see Exercise 10) that for this case

$$r(\hat{\theta}_B) = \text{tr}\{\mathbf{A}E\{\text{Cov}(\underline{\Theta}|Y)\}\}, \tag{IV.B.47}$$

where $\text{tr}\{\cdot\}$ denotes the trace operator (i.e., summation of the diagonal terms) and where $\text{Cov}(\underline{\Theta}|Y)$ is the conditional covariance matrix of $\underline{\Theta}$ given $Y = y$. Note that the squared-error norm is the special case of (IV.B.47) with $\mathbf{A} = \mathbf{I}$, so in the latter case $r(\hat{\underline{\theta}}_B)$ is simply the trace of $E\{\text{Cov}(\underline{\Theta}|Y)\}$.

Example IV.B.3: Estimation of a Gaussian Vector from a Jointly Gaussian Observation

Consider the situation in which $\Gamma = \mathbb{R}^n, \Lambda = \mathbb{R}^m$, and Y and $\underline{\Theta}$ are jointly Gaussian with mean vectors $\underline{\mu}_Y$ and $\underline{\mu}_\Theta$, covariance matrices $\mathbf{\Sigma}_Y$ and $\mathbf{\Sigma}_\Theta$, and cross-covariance matrix $\mathbf{\Sigma}_{Y\Theta} \triangleq E\{(\underline{Y} - \underline{\mu}_Y)(\underline{\Theta} - \underline{\mu}_\Theta)^T\}$; that is, we assume that

$$\begin{pmatrix} \underline{Y} \\ \underline{\Theta} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \underline{\mu}_Y \\ \underline{\mu}_\Theta \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_Y & \mathbf{\Sigma}_{Y\Theta} \\ \mathbf{\Sigma}_{\Theta Y} & \mathbf{\Sigma}_\Theta \end{pmatrix} \right) \tag{IV.B.48}$$

with $\mathbf{\Sigma}_{\Theta Y} = \mathbf{\Sigma}_{Y\Theta}^T$.

Within this model it is straightforward to show that the conditional distribution of $\underline{\Theta}$ given $\underline{Y} = \underline{y}$ is also Gaussian, with conditional mean $\hat{\underline{\mu}}(y)$ given by

$$\hat{\underline{\mu}}(\underline{y}) = \underline{\mu}_\Theta + \mathbf{\Sigma}_{\Theta Y} \mathbf{\Sigma}_Y^{-1} (\underline{y} - \underline{\mu}_Y) \tag{IV.B.49a}$$

and with conditional covariance matrix $\hat{\mathbf{\Sigma}}$ given by

$$\hat{\mathbf{\Sigma}} = \mathbf{\Sigma}_\Theta - \mathbf{\Sigma}_{\Theta Y} \mathbf{\Sigma}_Y^{-1} \mathbf{\Sigma}_{Y\Theta}. \tag{IV.B.49b}$$

From this property we can find all of the optimum estimates discussed in Case IV.B.4. In particular, we note immediately that the conditional-mean estimate is equal to $\hat{\underline{\mu}}(\underline{y})$ of (IV.B.49a). Also, since the multivariate

Gaussian density has its mode at its mean, the MAP estimate is given by $\hat{\underline{\mu}}(\underline{y})$ as well. Moreover, since $\underline{\Theta}$ being Gaussian given $\underline{Y} = \underline{y}$ implies that Θ_i is marginally Gaussian given $\underline{Y} = \underline{y}$, the marginal mode and median of Θ_i given $\underline{Y} = \underline{y}$ occur at $\hat{\mu}_i(\underline{y})$, the ith component of $\hat{\underline{\mu}}(\underline{y})$. Thus $\hat{\underline{\mu}}(\underline{y})$ provides the optimum estimate in all the senses discussed under Case IV.B.4. It should be noted that this estimate is linear (or, more properly, *affine*) in $\underline{y}$, so that it is easily computed if $\mathbf{\Sigma}_Y^{-1}$ can be determined efficiently. We will comment further on this issue later.

The minimum Bayes risk can also be computed easily for the quadratic cost function of (IV.B.43) via (IV.B.47). In particular we note that $\mathrm{Cov}(\underline{\Theta}|\underline{Y}) = \hat{\mathbf{\Sigma}}$, which does not depend on $\underline{Y}$. Thus $E\{\mathrm{Cov}(\underline{\Theta}|\underline{Y})\} = \hat{\mathbf{\Sigma}}$ and the minimum Bayes risk becomes

$$r(\hat{\underline{\theta}}_B) = tr\{\mathbf{A}\hat{\mathbf{\Sigma}}\} = tr\{\mathbf{A}\mathbf{\Sigma}_\theta\} - tr\{\mathbf{A}\mathbf{\Sigma}_{\Theta Y}\mathbf{\Sigma}_Y^{-1}\mathbf{\Sigma}_{Y\Theta}\}. \tag{IV.B.50}$$

Note also that $\hat{\mathbf{\Sigma}} = E\{(\underline{\Theta} - \hat{\underline{\theta}}_B(\underline{Y}))(\underline{\Theta} - \hat{\underline{\theta}}_B(\underline{Y}))^T\}$, so that $\hat{\mathbf{\Sigma}}$ is the covariance matrix of the estimation error, $\underline{\Theta} - \hat{\underline{\theta}}_B(\underline{Y})$.

A special case of interest of this general Gaussian problem arises from the so-called *linear observation model*:

$$\underline{Y} = \mathbf{H}\underline{\Theta} + \underline{N}, \tag{IV.B.51}$$

where $\underline{\Theta} \sim \mathcal{N}(\underline{\mu}_\theta, \mathbf{\Sigma}_\Theta)$, $\underline{N} \sim \mathcal{N}(\underline{0}, \mathbf{\Sigma})$, $\mathbf{H}$ is a fixed $n \times m$ matrix, and $\underline{\Theta}$ and $\underline{N}$ are independent. Such models arise in many applications. For example, the model of Example IV.B.2 in which we wish to estimate signal amplitude is of this form with $m = 1$ and $\mathbf{H} = \underline{s}$. Furthermore, if we think of $\Theta_1, \ldots, \Theta_m$ as being samples of a stochastic signal, then

$$Y_k = \sum_{j=1}^{m} h_{k,j}\Theta_j + N_k, \quad k = 1, \ldots, n \tag{IV.B.52}$$

is an observation sequence consisting of linearly filtered signal plus additive noise—a situation arising, for example, when a signal is observed through a channel with finite bandwidth or other linearly distorting characteristic. In this case the estimation of $\underline{\Theta}$ is known as the problem of *equalizing* the channel. A further applications of the model is discussed in Chapter V in the context of Kalman-Bucy filtering.

In this model it is straightforward to show that $\underline{Y}$ and $\underline{\Theta}$ are jointly Gaussian with $\underline{\mu}_\Theta$ and $\mathbf{\Sigma}_\Theta$ given, $\underline{\mu}_Y = \mathbf{H}\underline{\mu}_\theta$, $\mathbf{\Sigma}_Y = \mathbf{H}\mathbf{\Sigma}_\theta\mathbf{H}^T + \mathbf{\Sigma}$, and $\mathbf{\Sigma}_{\Theta Y} = \mathbf{\Sigma}_\theta\mathbf{H}^T$. Thus we get the Bayes estimate

$$\hat{\underline{\mu}}(\underline{y}) = \underline{\mu}_\theta + \mathbf{\Sigma}_\Theta\mathbf{H}^T(\mathbf{H}\mathbf{\Sigma}_\theta\mathbf{H}^T + \mathbf{\Sigma})^{-1}(\underline{y} - \mathbf{H}\underline{\mu}_\theta) \tag{IV.B.53}$$

and the error covariance matrix

$$\hat{\mathbf{\Sigma}} = \mathbf{\Sigma}_\theta - \mathbf{\Sigma}_\theta\mathbf{H}^T(\mathbf{H}\mathbf{\Sigma}_\theta\mathbf{H}^T + \mathbf{\Sigma})^{-1}\mathbf{H}\mathbf{\Sigma}_\theta. \tag{IV.B.54}$$

With regard to the computation of (IV.B.53) we note that it involves the inversion of an $n \times n$ matrix, a computation whose complexity is of the order of n^3 unless the matrix has some special structure. This computational complexity can sometimes be reduced by making use of the following simple matrix identity.

$$\begin{aligned}&\mathbf{\Sigma}_\theta\mathbf{H}^T(\mathbf{H}\mathbf{\Sigma}_\theta\mathbf{H}^T + \mathbf{\Sigma})^{-1} \\ &\quad = (\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{H} + \mathbf{\Sigma}_\theta^{-1})^{-1}\mathbf{H}^T\mathbf{\Sigma}^{-1}.\end{aligned} \tag{IV.B.55}$$

If $\mathbf{\Sigma}^{-1}$ is known (e.g., if $\mathbf{\Sigma} = \sigma^2\mathbf{I}$) and $m < n$, the matrix on the right-hand side of (IV.B.55) is easier to compute than that on the left.

In Chapter V, (IV.B.53) and (IV.B.54) will be used to derive the Kalman-Bucy filter. It is also interesting to rework Example IV.B.2 in this general context. In this case we have $m = 1$, $\mathbf{H} = \underline{s}$, $\underline{\mu}_\theta = \mu$, and $\mathbf{\Sigma}_\theta = v^2$. Inserting these quantities into (IV.B.53) and (IV.B.54) and applying (IV.B.55), we get

$$\begin{aligned}\hat{\underline{\mu}}(\underline{y}) &= \mu + (\underline{s}^T\mathbf{\Sigma}^{-1}\underline{s} + 1/v^2)^{-1}\underline{s}^T\mathbf{\Sigma}^{-1}(\underline{y} - \underline{s}\mu) \\ &= \frac{v^2d^2\hat{\theta}_1(\underline{y}) + \mu}{v^2d^2 + 1}\end{aligned}$$

and

$$\begin{aligned}r(\hat{\underline{\mu}}) &= \hat{\mathbf{\Sigma}} = v^2 - (\underline{s}^T\mathbf{\Sigma}^{-1}\underline{s} + v^{-2})^{-1}\underline{s}^T\mathbf{\Sigma}^{-1}\underline{s}v^2 \\ &= \frac{v^2}{v^2d^2 + 1}\end{aligned}$$

as in (IV.B.34) and (IV.B.35).

IV.C Nonrandom Parameter Estimation: General Structure

In Section IV.B we considered the problem of estimating a random parameter indexing a class of distributions on the observation space. A related problem is that in which we have a parameter (indexing the class of observation statistics) that is not modeled as a random variable but, nevertheless, is unknown. In particular, we may not have enough prior information about the parameter to assign a prior probability distribution to it, but yet we wish to treat the estimation of such parameters in an organized manner.

Suppose, then, that we have an observation $Y \in \Gamma$ and that the distribution of Y is a member of a class of distributions on $(\Gamma, \mathcal{G})$ indexed by a parameter θ lying in some set Λ. As before, we denote this set of distributions by $\{P_\theta; \theta \in \Lambda\}$. Assume for now that the parameter θ is real-valued.

We do not know anything about the true value of θ other than the fact that it lies in Λ, and simply stated, the problem we would like to solve is: Given the observation $Y = y$, what is the best estimate of θ? In view of the procedures developed in Section IV.B, we might begin to answer this question by seeking an estimate $\hat{\theta}(y)$ that minimizes some average performance criterion. Throughout the remainder of this chapter we consider exclusively the squared-error cost, although some results discussed here apply straightforwardly to other cost assignments as well. In the absence of a prior on Λ, the only averaging of cost that can be done is with respect to the distribution of Y given θ; i.e., we can use only the conditional risk function $R_\theta(\hat{\theta}) \triangleq E_\theta\{(\hat{\theta}(Y) - \theta)^2\}, \theta \in \Lambda$.

As was seen in the hypothesis-testing case in Chapter II, we cannot generally expect to minimize $R_\theta(\hat{\theta})$ uniformly for $\theta \in \Lambda$. This is easily seen for the squared-error cost since for any particular value of θ, say θ_o, the conditional mean-squared error can be made zero by choosing $\hat{\theta}(y)$ to be identically θ_o for all observations $y \in \Gamma$; but such an estimate would perform poorly if θ_o were not near the true value of θ. Thus it is obvious that the conditional mean-squared error is not by itself a suitable design criterion for an estimator of a nonrandom parameter unless the class of estimators is somehow restricted to contain only reasonable estimators [e.g., to exclude estimators such as $\hat{\theta}(y) \equiv \theta_o$].

A reasonable restriction to place on an estimate of θ is that its expected value equal the true parameter value; i.e., that

$$E_\theta\{\hat{\theta}(Y)\} = \theta, \quad \theta \in \Lambda. \tag{IV.C.1}$$

Such an estimate is termed *unbiased*. Within this restriction, the conditional mean-squared error becomes the variance of the estimate under P_θ, and an unbiased estimate minimizing the mean-squared error for each $\theta \in \Lambda$ is termed a *minimum-variance unbiased estimator* (MVUE).

In this section we consider the general structure of nonrandom parameter estimation problems with a goal of characterizing MVUEs.

We begin with the concept of sufficiency, defined as follows (until otherwise noted, we now assume that Λ is general, i.e., not necessarily a subset of $\mathbb{R}$).

Definition IV.C.1: Sufficiency

Suppose that Δ is an arbitrary set and $\mathcal{D}$ is an event class on Δ. A function $T : (\Gamma, \mathcal{G}) \to (\Delta, \mathcal{D})$ is said to be a *sufficient statistic* for $\{P_\theta; \theta \in \Lambda\}$ if the distribution of Y conditioned on $T(Y)$ when $Y \sim P_\theta$ does not depend on θ for $\theta \in \Lambda$. (When $\{P_\theta; \theta \in \Lambda\}$ is understood, we may simply say that T *is sufficient for* θ.)

Note that θ affects the observations only through its distribution P_θ. So we can only learn about θ by viewing the statistical behavior of Y. Thus

if knowing $T(Y)$ removes any further dependence on θ of the distribution of Y, we can conclude that $T(Y)$ contains all the information in Y that is useful for estimating θ—thus the origin of the term "sufficient."

Note that any one-to-one mapping of the observations is trivially sufficient for θ, so there are always many sufficient statistics for any given estimation model. However, it is desirable to find a sufficient statistic that reduces the observations as much as possible. In this context we make the following definition.

Definition IV.C.2: Minimal Sufficiency

A function T on $(\Gamma, \mathcal{G})$ is said to be *minimal sufficient* for $\{P_\theta; \theta \in \Lambda\}$ if it is a function of every other sufficient statistic for $\{P_\theta; \theta \in \Lambda\}$.

In other words, a minimal sufficient statistic represents the furthest that the observation can be reduced without destroying information about θ. Unfortunately, minimal sufficient statistics do not exist for many estimation problems, and they are often difficult to identify when they do exist.

On the other hand, it is often very easy to find useful (although not necessarily minimal) sufficient statistics by way of the following result.

Proposition IV.C.1: The Factorization Theorem

Suppose that $\{P_\theta; \theta \in \Lambda\}$ has a corresponding family of densities $\{p_\theta; \theta \in \Lambda\}$. A statistic T is sufficient for θ if and only if there are functions g_θ and h such that

$$p_\theta(y) = g_\theta[T(y)]h(y) \tag{IV.C.2}$$

for all $y \in \Gamma$ and $\theta \in \Lambda$.

Proof: We prove this result only for the case in which Γ is discrete. This case illustrates the general idea of this proposition without introducing technicalities required for the general case. A proof of the general case can be found in Lehmann (1986).

Suppose that Γ is discrete and $\{p_\theta; \theta \in \Lambda\}$ satisfies (IV.C.2) for a function T. Let $p_\theta(y|t)$ denote the density of Y given $T(Y) = t$ when $Y \sim P_\theta$. By the Bayes formula we have

$$\begin{aligned} p_\theta(y|t) &\triangleq P_\theta(Y = y|T(Y) = t) \\ &= \frac{P_\theta(T(Y) = t|Y = y)P_\theta(Y = y)}{P_\theta(T(Y) = t)}. \end{aligned} \tag{IV.C.3}$$

Since $P_\theta(T(Y) = t|Y = y)$ equals 1 if $T(y) = t$ and 0 if $T(y) \neq t$, and since $P_\theta(Y = y) = p_\theta(y)$, (IV.C.3) becomes

$$p_\theta(y|t) = \begin{cases} p_\theta(y)/P_\theta(T(Y) = t) & \text{if } T(y) = t \\ 0 & \text{if } T(y) \neq t. \end{cases} \tag{IV.C.4}$$

Now $P_\theta(T(Y)=t)=\sum_{\{y|T(y)=t\}} p_\theta(y)$. Thus from (IV.C.2), we have

$$\begin{aligned} P_\theta(T(Y)=t) &= \sum_{\{y|T(y)=t\}} g_\theta[T(y)]h(y) \\ &= g_\theta(t) \sum_{\{y|T(y)=t\}} h(y), \end{aligned}$$

and we also have $p_\theta(y)=g_\theta[T(y)]h(y)=g_\theta(t)h(y)$. From (IV.C.4) we then have

$$p_\theta(y|t)=\begin{cases} h(y)/\sum_{\{y|T(y)=t\}} h(y), & \text{if } T(y)=t \\ 0, & \text{if } T(y)\neq t. \end{cases}$$

Since this expression does not depend on θ, T is a sufficient statistic for $\{P_\theta;\theta\in\Lambda\}$. This proves that T is sufficient if (IV.C.2) holds.

To prove that T is sufficient only if (IV.C.2) holds, let T be any sufficient statistic for θ. From (IV.C.4) we can write

$$p_\theta(y)=p_\theta[y|T(y)]P_\theta[T(Y)=T(y)]. \tag{IV.C.5}$$

Since T is sufficient for $\theta, p_\theta[y|T(y)]$ depends only on y and not on θ. Also, $P_\theta[T(Y)=T(y)]$ is a function only of $T(y)$ and θ. On defining $h(y)\triangleq p_\theta[y|T(y)]$ and $g_\theta[T(y)]\triangleq P_\theta[T(Y)=T(y)]$, we see that (IV.C.5) implies the factorization of (IV.C.2). This completes the proof of this proposition for the discrete Γ case. □

To illustrate Proposition IV.C.1, we consider the following simple example.

Example IV.C.1: A Sufficient Statistic for Hypothesis Testing

Consider the hypothesis-testing problem $\Lambda=\{0,1\}$ with densities p_0 and p_1. Noting that

$$p_\theta(y)=\begin{cases} p_0(y) & \text{if } \theta=0 \\ \frac{p_1(y)}{p_0(y)}p_0(y) & \text{if } \theta=1, \end{cases}$$

we can see the factorization $p_\theta(y)=g_\theta[T(y)]h(y)$ with $h(y)=p_0(y), T(y)=p_1(y)/p_0(y)\triangleq L(y)$, and $g_\theta(t)$ defined by

$$g_\theta(t)=\begin{cases} 1 & \text{if } \theta=0 \\ t & \text{if } \theta=1. \end{cases}$$

Thus we see that the likelihood ratio $L(y)$ is a sufficient statistic for the binary hypothesis-testing problem. It is a very useful sufficient statistic because it is one-dimensional regardless of the nature of Γ. Of course, we have already seen that all of the optimum tests for $\Lambda=\{0,1\}$ defined in Chapter II depend on the observation y only through this sufficient statistic $L(y)$.

The usefulness of sufficient statistics in seeking good unbiased estimators of real parameters can be seen partly from the following result. Here we allow Λ to be arbitrary, but we suppose that we wish to estimate some real-valued function g of θ.

Proposition IV.C.2: The Rao-Blackwell Theorem

Suppose that $\hat{g}(y)$ is an unbiased estimate of $g(\theta)$ and that T is sufficient for θ. Define $\tilde{g}[T(y)]$ by

$$\tilde{g}[T(y)]=E_\theta\{\hat{g}(Y)|T(Y)=T(y)\}.$$

Then $\tilde{g}[T(Y)]$ is also an unbiased estimate of $g(\theta)$. Furthermore,

$$\mathrm{Var}_\theta(\tilde{g}[T(Y)])\leq \mathrm{Var}_\theta[\hat{g}(Y)],$$

with equality if and only if $P_\theta(\hat{g}(Y)=\tilde{g}[T(Y)])=1$.

Proof: We remark first that the expectation defining $\tilde{g}$ does not depend on θ by virtue of the sufficiency of T [i.e., given $T(Y)$, the distribution of Y, and hence the mean of $\hat{g}(Y)$, does not depend on θ]. To see that $\tilde{g}$ is unbiased, we note that

$$\begin{aligned} E_\theta\{\tilde{g}[T(Y)]\} &= E_\theta\{E_\theta\{\hat{g}(Y)|T(Y)\}\} \\ &= E_\theta\{\hat{g}(Y)\}=g(\theta), \end{aligned}$$

where we have used the fact that $E\{E\{X|Z\}\}=E\{X\}$ to get the second equality and the unbiasedness of $\hat{g}$ to get the third equality.

To see that $\mathrm{Var}_\theta(\tilde{g}[T(Y)])\leq \mathrm{Var}_\theta[\hat{g}(Y)]$, we first note that

$$\mathrm{Var}_\theta(\tilde{g}[T(Y)])=E_\theta\{[\tilde{g}[T(Y)]]^2\}-g^2(\theta)$$

and

$$\mathrm{Var}_\theta(\hat{g}(Y))=E_\theta\{[\hat{g}(Y)]^2\}-g^2(\theta).$$

So we only need to show that $E_\theta\{[\tilde{g}[T(Y)]]^2\}\leq E_\theta\{[\hat{g}(Y)]^2\}$. We have

$$\begin{aligned} E_\theta\{(\tilde{g}[T(Y)])^2\} &= E_\theta\{[E_\theta\{\hat{g}(Y)|T(Y)\}]^2\} \\ &\leq E_\theta\{E_\theta\{[\hat{g}(Y)]^2|T(Y)\}\}=E_\theta\{[\hat{g}(Y)]^2\}, \end{aligned} \tag{IV.C.6}$$

where the inequality follows from applying Jensen's inequality to get $[E_\theta\{\hat{g}(Y)|T(Y)\}]^2\leq E_\theta\{[\hat{g}(Y)]^2|T(Y)\}$, and the final equality follows from

iterated expectations. Note that we have equality in Jensen's inequality here if and only if $P_\theta[\hat{g}(Y) = E_\theta\{\hat{g}(Y)|T(Y)\}|T(Y)] = 1$. Since $\tilde{g}[T(Y)] \triangleq E_\theta\{\hat{g}(Y)|T(Y)\}$, this is equivalent to the condition $P_\theta[\hat{g}(Y) = \tilde{g}[T(Y)]] = 1$. This completes the proof of Proposition IV.C.2. □

From the Rao-Blackwell theorem we see that with a sufficient statistic T we can improve any unbiased estimator that is not already a function of T by conditioning it on $T(Y)$. Furthermore, this theorem implies that if T is sufficient for θ and if there is only one function of T that is an unbiased estimate of $g(\theta)$, that function is an MVUE for $g(\theta)$. To see this, suppose that $g^\star[T(y)]$ is the only function of $T(y)$ for which $E_\theta\{g^\star[T(Y)]\} = g(\theta)$. Let $\hat{g}(y)$ be any unbiased estimator of $g(\theta)$. Then, by the Rao-Blackwell theorem, $\tilde{g}[T(y)] \triangleq E_\theta\{\hat{g}(Y)|T(Y) = T(y)\}$ is unbiased for $g(\theta)$ and it is a function of $T(y)$. So by uniqueness of $g^\star$, we must have $g^\star = \tilde{g}$. The Rao-Blackwell theorem also asserts that $\mathrm{Var}_\theta(\tilde{g}[T(Y)]) \leq \mathrm{Var}_\theta[\hat{g}(Y)]$. Since $\hat{g}$ is arbitrary, we see that $\mathrm{Var}_\theta(g^\star[T(Y)]) \leq \mathrm{Var}_\theta[\hat{g}(Y)]$ for any unbiased estimate of $g(\theta)$; in other words, $g^\star[T(y)]$ is an MVUE of $g(\theta)$.

Thus we see that an MVUE of $g(\theta)$ can be constructed if we can find a sufficient statistic T with such a unique unbiased estimate $g^\star[T(y)]$. Toward the end of finding such a statistic, we introduce the notion of completeness.

Definition IV.C.3: Completeness

The family $\{P_\theta; \theta \in \Lambda\}$ is said to be *complete* if the condition $E_\theta\{f(Y)\} = 0$ for all $\theta \in \Lambda$ implies that $P_\theta[f(Y) = 0] = 1$ for all $\theta \in \Lambda$.

This notion of completeness is very similar to the notion of completeness of a set of vectors in $\mathbb{R}^n$. To see this, consider the situation in which Γ is a finite set $\{\gamma_1, \ldots, \gamma_n\}$. In this case, for any function f on Γ we can write

$$E_\theta\{f(Y)\} = \underline{f}^T \underline{p}_\theta,$$

where $\underline{f} = [f(\gamma_1), f(\gamma_2), \ldots, f(\gamma_n)]^T$ and $\underline{p}_\theta = [p_\theta(\gamma_1), p_\theta(\gamma_2), \ldots, p_\theta(\gamma_n)]^T$. Assuming that $p_\theta(\gamma_i) > 0$ for all $\theta \in \Lambda$ and $i = 1, \ldots, n$, the completeness of $\{p_\theta; \theta \in \Lambda\}$ is defined by the condition that $\underline{f}^T \underline{p}_\theta = 0$ for all $\theta \in \Lambda$ implies that $\underline{f}$ is the n-vector of all zeros. That is, $\{P_\theta; \theta \in \Lambda\}$ is complete if $\underline{0}$ is the only vector that is orthogonal to all the vectors $\{\underline{p}_\theta; \theta \in \Lambda\}$. This, of course, is the ordinary notion of completeness of the set of vectors $\{\underline{p}_\theta; \theta \in \Lambda\}$ in $\mathbb{R}^n$. (Recall that a complete set of vectors in $\mathbb{R}^n$ is said to *span* $\mathbb{R}^n$.) Similar analogies hold for more general observation spaces.

To illustrate the notion of completeness further, consider the following example.

Example IV.C.2: Completeness of the Binomial Distribution

Suppose that $\Gamma = \{0, 1, \ldots, n\}$, $\Lambda = (0, 1)$, and

$$p_\theta(y) = \frac{n!}{y!(n-y)!}\theta^y(1-\theta)^{n-y}, \quad y = 0, \ldots, n, \quad 0 < \theta < 1.$$

For any function f on Γ we have

$$\begin{aligned} E_\theta\{f(Y)\} &= \sum_{y=0}^{n} \frac{n!}{y!(n-y)!} f(y)\theta^y(1-\theta)^{n-y}, \\ &= (1-\theta)^n \sum_{y=0}^{n} a_y x^y, \end{aligned}$$

where

$$a_y \triangleq \frac{n!}{y!(n-y)!} f(y), \quad \text{for } y = 0, \ldots, n,$$

and

$$x \triangleq \theta/(1-\theta).$$

The condition $E_\theta\{f(Y)\} = 0$ for all $\theta \in \Lambda$ is equivalent to the condition

$$\sum_{y=0}^{n} a_y x^y = 0, \text{ for all } x > 0. \tag{IV.C.7}$$

The function $\sum_{y=0}^{n} a_y x^y$ is an nth-order polynominal and thus has at most n zeros unless all its coefficients are zero. It follows that (IV.C.7) can be satisfied only with $f(y) = 0, y = 0, \ldots, n$. So $\{p_\theta; \theta \in \Lambda\}$ is complete. Note that completeness is retained here for any Λ containing at least $(n+1)$ nonzero parameter values.

The notions of completeness and sufficiency are closely related. To see this, suppose that T is sufficient for the complete family $\{P_\theta; \theta \in \Lambda\}$, and for convenience assume that $E_\theta\{|Y|\} < \infty$ for each $\theta \in \Lambda$. Define a function $f(y)$ by

$$f(y) = y - E_\theta\{Y|T(Y) = T(y)\}.$$

Note that f does not depend on θ since T is sufficient. For each $\theta \in \Lambda$ we have

$$\begin{aligned} E_\theta\{f(Y)\} &= E_\theta\{Y\} - E_\theta\{E_\theta\{Y|T(Y)\}\} \\ &= E_\theta\{Y\} - E_\theta\{Y\} = 0. \end{aligned}$$

Thus the completeness of $\{P_\theta; \theta \in \Lambda\}$ implies that $P_\theta[Y = E_\theta\{Y|T(Y)\}] = 1$ for all $\theta \in \Lambda$ or, in effect, that $y = E_\theta\{Y|T(Y) = T(y)\}$. Since

$E_\theta\{Y|T(Y) = T(y)\}$ is a function of $T(y)$, the latter condition implies that y itself is a function of $T(y)$. Since $T(y)$ is obviously a function of y, we see that $T(y)$ must be a one-to-one function of y; that is T is a trivial sufficient statistic. We conclude then that *if* $\{P_\theta; \theta \in \Lambda\}$ *is complete, then there is no nontrivial sufficient statistic for* θ; i.e., the observation Y cannot be reduced without destroying information about θ.

Completeness is a useful concept in characterizing MVUEs. To see this, suppose that T is sufficient for θ, and let Q_θ denote the distribution of $T(Y)$ when $Y \sim P_\theta$. If $\{Q_\theta; \theta \in \Lambda\}$ is complete, then T is said to be a *complete sufficient statistic.*[3] Suppose that T is complete and let $\tilde{g}[T(y)]$ and $g^\star[T(y)]$ be any functions of $T(y)$ that are unbiased estimators of $g(\theta)$. We have

$$\begin{aligned} E_\theta\{\tilde{g}[T(Y)] - g^\star[T(Y)]\} &= E_\theta\{\tilde{g}[T(Y)]\} - E_\theta\{g^\star[T(Y)]\} \\ &= g(\theta) - g(\theta) = 0 \end{aligned}$$

for all $\theta \in \Lambda$. Thus, by the completeness of T, we see that $P_\theta(\tilde{g}[T(Y)] = g^\star[T(Y)]) = 1$ for all $\theta \in \Lambda$, i.e., that $\tilde{g}[T(y)]$ and $g^\star[T(y)]$ are the same estimator. Thus since $\tilde{g}$ and $g^\star$ were chosen arbitrarily, we see that *any unbiased estimator that is a function of a complete sufficient statistic is unique in this respect and thus is an MVUE.*

We thus see a procedure for seeking MVUEs:

1. Find a complete sufficient statistic T for $\{P_\theta; \theta \in \Lambda\}$.
2. Find *any* unbiased estimator $\hat{g}(y)$ of $g(\theta)$.
3. Then $\tilde{g}[T(y)] \triangleq E_\theta\{\hat{g}(Y)|T(Y) = T(y)\}$ is an MVUE of $g(\theta)$.

Of these stages the first appears to be the least straightforward, since the second step is often fairly easy and the third is accomplished directly by probability calculus. However, for many models of interest in practice the first step turns out to be quite easy. To develop this, we first present the following definition.

Definition IV.C.4: Exponential Families

A class of distributions $\{P_\theta; \theta \in \Lambda\}$ is said to be an *exponential family* if there are real-valued functions $C, Q_1, \ldots, Q_m, T_1, \ldots, T_m$, and h such that

[3] In view of the discussion above, we see that the observation cannot be reduced beyond $T(Y)$ without destroying information about θ. In fact, T must be a minimal sufficient statistic for $\{P_\theta; \theta \in \Lambda\}$. This follows since, for any sufficient statistic T', we must have $T = E_\theta\{T|T'\}$ by completeness. Thus, T is a function of T'.

P_θ has density

$$p_\theta(y) = C(\theta) \exp\left\{\sum_{l=1}^{m} Q_l(\theta) T_l(y)\right\} h(y), \qquad \text{(IV.C.8)}$$

for all $\theta \in \Lambda$ and $y \in \Gamma$.

Many distributions encountered in practice can be put into the form of exponential families, including Gaussian, Poisson, Laplacian, binomial, geometric, and certain multivariate forms of these. Exponential families play an important role in the theory of minimum-variance unbiased estimation by virtue of the following result.

Proposition IV.C.3: The Completeness Theorem for Exponential Families

Suppose that $\Gamma = \mathbb{R}^n, \Lambda \subset \mathbb{R}^m$ and that each P_θ has density p_θ given by

$$p_\theta(y) = C(\theta) \exp\left\{\sum_{l=1}^{m} \theta_l T_l(y)\right\} h(y), \qquad \text{(IV.C.9)}$$

where $C, T_1, \ldots, T_m$, and h are real-valued functions.[4] Then $T(y) = [T_1(y), \ldots, T_m(y)]$ is a complete sufficient statistic for $\{P_\theta; \theta \in \Lambda\}$ if Λ contains a m-dimensional rectangle.

Outline of Proof: A complete proof of Proposition IV.C.3 can be found in Lehmann (1986). The steps in this proof can be outlined as follows.

We first note that T is sufficient for θ by the factorization theorem (Proposition IV.C.1), so we need only show completeness of T. With Y distributed according to (IV.C.9), it is straightforward to show that $T(Y)$ will have a density (on $\mathbb{R}^m$) of the form

$$g_\theta(t) = C(\theta) \exp\left\{\sum_{l=1}^{m} \theta_l t_l\right\} h_T(t), \qquad \text{(IV.C.10)}$$

where h_T is a real-valued function of t. Suppose that f is a real-valued function on $\mathbb{R}^m$ such that $E_\theta\{f[T(Y)]\} = 0$. We have

$$E_\theta\{f[T(Y)]\} = C(\theta) \int_{\mathbf{R}^m} f(t) \exp\left\{\sum_{l=1}^{m} \theta_l t_l\right\} h_T(t) \mu(dt). \qquad \text{(IV.C.11)}$$

[4] Note that (IV.C.8) can be reparameterized to be put into the form of (IV.C.9).

Suppose that Λ contains an m-dimensional rectangle $J = \{\theta | a_1 \leq \theta_1 \leq b_1, a_2 \leq \theta_2 \leq b_2, \ldots, a_m \leq \theta_m \leq b_m\}$. By simple translation of the parameters we can always choose this rectangle to be of the form $J' = \{\theta | -1 \leq \theta_1 \leq 1, -1 \leq \theta_2 \leq 1, \ldots, -1 \leq \theta_m \leq 1\}$. Consider (IV.C.11) as a function of a complex variable by replacing θ_l with $\theta_l + iu_l, l = 1, \ldots, m$. It can be shown that this function is analytic in the region $C = \{\theta + iu | -1 \leq \theta_l \leq 1, -\infty < u_l < \infty, l = 1, \ldots, m\}$, and thus the condition that it be zero for all real arguments in J' implies that it is zero throughout the strip C. In particular, this function is zero in the region $C' = \{\theta + iu | \theta_l = 0, -\infty < u_l < \infty, l = 1, \ldots, m\}$, i.e., we have

$$C(\theta) \int_{\mathbf{R}^m} f(t) \exp\left\{ i \sum_{l=1}^{m} u_l t_l \right\} h_T(t) \mu(dt) = 0, \qquad \text{(IV.C.12)}$$

for all $u \in \mathbb{R}^m$. Note that the function on the left of (IV.C.12) is a multidimensional Fourier transform. This being identically zero for all $\theta \in \Lambda$ implies that the function being transformed is zero for all $\theta \in \Lambda$, or equivalently that $P_\theta\{f(Y) = 0\} = 1$ for all $\theta \in \Lambda$. This implies in turn that T is complete, and thus completes the proof of the proposition. □

To illustrate the use of Proposition IV.C.3, we consider the following example.

Example IV.C.3: Minimum-Variance Unbiased Estimation of Signal Amplitude

Consider the model

$$Y_k = N_k + \mu s_k, \quad k = 1, \ldots, n$$

where $N_1, \ldots, N_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ noise samples, $\underline{s} = (s_1, \ldots, s_n)^T$ is a known signal, and μ is a signal amplitude parameter. Assume for now that σ^2 is known and that we wish to estimate the amplitude parameter μ. The density of $\underline{Y}$ is given by

$$\begin{aligned} &\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{k=1}^{n} (y_k - \mu s_k)^2 \right\} \\ &\quad = C(\theta_1) \exp\{\theta_1 T_1(\underline{y})\} h(\underline{y}), \qquad \text{(IV.C.13)} \end{aligned}$$

where we have defined

$$\begin{aligned} \theta_1 &= \mu/\sigma^2, \\ T_1(\underline{y}) &= \sum_{k=1}^{n} s_k y_k, \\ C(\theta_1) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{\theta_1^2 \sigma^2}{2} \sum_{k=1}^{n} s_k^2 \right\}, \end{aligned}$$

and

$$h(\underline{y}) = \exp\left\{ -\frac{1}{2\sigma^2} \sum_{k=1}^{n} y_k^2 \right\}.$$

Assuming that μ is an arbitrary real number, the parameter set is $\Lambda = \{\theta_1 | -\infty < \theta_1 < \infty\} = \mathbb{R}$. A one-dimensional rectangle is an interval, and Λ obviously contains an interval, so from Proposition IV.C.3 and (IV.C.13), we see that $T_1(\underline{y})$ is a complete sufficient statistic for θ_1.

We wish to estimate $\mu = g(\theta) = \sigma^2 \theta_1$. Note that $E_\theta\{Y_1\} = \mu s_1$. So, assuming that $s_1 \neq 0$, the estimate $\hat{g}(\underline{y}) = y_1/s_1$ is an unbiased estimator of $g(\theta)$. Thus since T_1 is complete, the estimate

$$\tilde{g}[T_1(\underline{y})] = E_\theta\{\hat{g}(\underline{Y}) | T_1(\underline{Y}) = T_1(\underline{y})\} \qquad \text{(IV.C.14)}$$

is an MVUE. To compute (IV.C.14) we note that $\hat{g}(\underline{Y})$ and $T_1(\underline{Y})$ are both linear functions of $\underline{Y}$, which is Gaussian. Thus $\hat{g}(\underline{Y})$ and $T_1(\underline{Y})$ are jointly Gaussian. It is easy to see that

$$\begin{aligned} E_\theta\{\hat{g}(\underline{Y})\} &= \mu, \\ E_\theta\{T_1(\underline{Y})\} &= n\mu\overline{s^2} \\ \mathrm{Var}_\theta\{\hat{g}(\underline{Y})\} &= \sigma^2/s_1^2, \\ \mathrm{Var}_\theta\{T_1(\underline{Y})\} &= n\sigma^2\overline{s^2}, \end{aligned}$$

and

$$\mathrm{Cov}_\theta[\hat{g}(\underline{Y}), T_1(\underline{Y})] = \sigma^2,$$

where we have defined $\overline{s^2} \triangleq (1/n) \sum_{k=1}^{n} s_k^2$. So, applying the results of Section IV.B, we can write this conditional mean of (IV.C.14) as

$$\begin{aligned} \tilde{g}[T_1(\underline{y})] &= E_\theta\{\hat{g}(\underline{Y})\} + \mathrm{Cov}_\theta[\hat{g}(\underline{Y}), T_1(\underline{Y})] \\ &\quad \times [\mathrm{Var}_\theta[T_1(\underline{Y})]]^{-1} [T_1(\underline{y}) - E_\theta\{T_1(\underline{Y})\}] \\ &= \mu + \sigma^2 (n\sigma^2\overline{s^2})^{-1} [T_1(\underline{y}) - n\mu\overline{s^2}] \\ &= T_1(\underline{y})/n\overline{s^2} = \left(\sum_{k=1}^{n} s_k y_k \right) \Big/ n\overline{s^2}. \qquad \text{(IV.C.15)} \end{aligned}$$

Thus we have constructed an MVUE for the signal amplitude μ. The variance of this estimator is

$$\mathrm{Var}_\theta(\tilde{g}[T_1(\underline{Y})]) = \sigma^2/n\overline{s^2}. \qquad \text{(IV.C.16)}$$

Suppose now that both μ and σ^2 are unknown, with μ ranging over $\mathbb{R}$ and σ^2 ranging over $(0, \infty)$, and that we would like to estimate both of these parameters. We see from (IV.C.16) that estimating σ^2 gives us an estimate of the accuracy of our amplitude estimate. Note that $h(\underline{y})$ as

defined in (IV.C.13) is a function of σ^2, so that (IV.C.13) as written is not a correct exponential family if σ^2 is not known. However, we can rewrite the density as

$$\begin{aligned}\frac{1}{(2\pi\sigma^2)^{n/2}} \quad & \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(y_k-\mu s_k)^2\right\} \\ = C(\theta) \quad & \exp\left\{\theta_1 T_1(\underline{y})+\theta_2 T_2(\underline{y})\right\} h(\underline{y}),\end{aligned} \tag{IV.C.17}$$

where θ_1 and T_1 are as in (IV.C.13), but, we now define $\theta=(\theta_1,\theta_2)$,

$$\begin{aligned}\theta_2 &= -\frac{1}{2\sigma^2}, \\ T_2(\underline{y}) &= \sum_{k=1}^{n} y_k^2, \\ C(\theta) &= \left(-\frac{\theta_2}{\pi}\right)^{n/2}\exp\left\{\frac{\theta_1^2}{4\theta_2}\sum_{k=1}^{n}s_k^2\right\},\end{aligned}$$

and

$$h(\underline{y})\equiv 1.$$

The range $\{(\mu,\sigma^2)|\mu\in\mathbb{R},\sigma^2>0\}$ corresponds to $\Lambda=\{(\theta_1,\theta_2)|\theta_1\in\mathbb{R},\theta_2<0\}$, which certainly contains a rectangle. Thus $T=(T_1,T_2)$ is a complete sufficient statistic for θ.

We wish to estimate $\mu=g_1(\theta)\triangleq -\theta_1/2\theta_2$ and $\sigma^2=g_2(\theta)\triangleq -1/2\theta_2$. Note that the estimate found in (IV.C.15) is computed without knowledge of σ^2, it is unbiased, and it is a function of $T_1(\underline{y})$ [and hence of $T(\underline{y})$]. Thus it is an MVUE of μ even when σ^2 is not known.

To find an MVUE of σ^2 we can first seek an unbiased estimator of σ^2 and then condition it on $T(\underline{y})$. It is simpler in this case, however, to look directly for an unbiased function of T. In particular, we note that since $T_1(\underline{Y})\sim\mathcal{N}(n\mu\overline{s^2},n\sigma^2\overline{s^2})$, we have

$$\begin{aligned}E_\theta\{T_1^2(\underline{Y})\} &\equiv \mathrm{Var}_\theta[T_1(\underline{Y})]+(E_\theta\{T_1(\underline{Y})\})^2 \\ &= n\sigma^2\overline{s^2}+n^2\mu^2(\overline{s^2})^2.\end{aligned}$$

Also, we have that

$$\begin{aligned}E_\theta\{T_2(\underline{Y})\} &= \sum_{k=1}^{n}E_\theta\{Y_k^2\}=\sum_{k=1}^{n}(\sigma^2+\mu^2 s_k^2) \\ &= n\sigma^2+n\mu^2\overline{s^2}.\end{aligned}$$

From these two results we see that the quantity $[T_2(\underline{Y})-T_1^2(\underline{Y})/n\overline{s^2}]$ has mean

$$E_\theta\{T_2(\underline{Y})\}-E_\theta\{T_1^2(\underline{Y})/n\overline{s^2}\}=(n-1)\sigma^2. \tag{IV.C.18}$$

Thus the function $\tilde{g}_2[T(\underline{y})]=[T_2(\underline{y})-T_1^2(\underline{y})/n\overline{s^2}]/(n-1)$ is an unbiased estimator of σ^2, and by the completeness of T it is an MVUE. We can rewrite $\tilde{g}_2$ as

$$\tilde{g}_2[T(\underline{y})]=\frac{1}{n-1}\sum_{k=1}^{n}(y_k-\hat{\mu}s_k)^2\triangleq\hat{\sigma}^2, \tag{IV.C.19}$$

where $\hat{\mu}$ is the MVUE of μ from (IV.C.15). Note that $\hat{n}_k\triangleq y_k-\hat{\mu}s_k$ is an estimate of the noise in the kth sample, so $\hat{\sigma}^2$ estimates the variance (which equals the second moment) of the noise by $[1/(n-1)]\sum_{k=1}^{n}(\hat{n}_k)^2$. Note that a more natural estimator for the second moment would be $(1/n)\sum_{k=1}^{n}(\hat{n}_k)^2$; but as we see from the analysis above, the latter estimate is biased. Further discussion of this point is included in Section IV.D.

The theory outlined in the paragraphs above provides a means for seeking minimum-variance unbiased estimators. For many models of interest, however, the structure required for applying results such as Proposition IV.C.3 is not present. Thus we are often faced with the problem of proposing an estimator and evaluating its performance (i.e., its bias and variance) in the absence of any knowledge about the optimality of the estimator. In such cases it is useful to have a standard to which estimators can be compared; i.e., it would be useful to know the fundamental limitations on estimator performance imposed by a given model. Such a standard is provided in part by the following result.

Proposition IV.C.4: The Information Inequality

Suppose that $\hat{\theta}$ is an estimate of the parameter θ in a family $\{P_\theta;\theta\in\Lambda\}$ and that the following conditions hold:

(1) Λ is an open interval.

(2) The family $\{P_\theta;\theta\in\Lambda\}$ has a corresponding family of densities $\{p_\theta;\theta\in\Lambda\}$, all of the members of which have the same support.[5]

(3) $\partial p_\theta(y)/\partial\theta$ exists and is finite for all $\theta\in\Lambda$ and all y in the support of p_θ.

(4) $\partial\int_\Gamma h(y)p_\theta(y)\mu(dy)/\partial\theta$ exists and equals $\int_\Gamma h(y)[\partial p_\theta(y)/\partial\theta]\mu(dy)$, for all $\theta\in\Lambda$, for $h(y)=\hat{\theta}(y)$ and $h(y)=1$.

Then

$$\mathrm{Var}_\theta[\hat{\theta}(Y)]\geq\frac{\left[\frac{\partial}{\partial\theta}E_\theta\{\hat{\theta}(Y)\}\right]^2}{I_\theta} \tag{IV.C.20}$$

[5] That is, the set $\{y|p_\theta(y)>0\}$ is the same for all $\theta\in\Lambda$.

where

$$I_\theta \triangleq E_\theta \left\{ \left(\frac{\partial}{\partial \theta} \log p_\theta(Y) \right)^2 \right\}. \tag{IV.C.21}$$

Furthermore, if the following condition also holds:

(5) $\partial^2 p_\theta(y)/\partial\theta^2$ exists for all $\theta \in \Lambda$ and y in the support of p_θ and

$$\int \frac{\partial^2}{\partial\theta^2} p_\theta(y)\mu(dy) = \frac{\partial^2}{\partial\theta^2} \int p_\theta(y)\mu(dy),$$

then I_θ can be computed via

$$I_\theta = -E_\theta \left\{ \frac{\partial^2}{\partial\theta^2} \log p_\theta(Y) \right\}. \tag{IV.C.22}$$

Proof: The proof of this result follows straightforwardly from the Schwarz inequality. In particular, we have that

$$E_\theta\{\hat{\theta}(Y)\} = \int_\Gamma \hat{\theta}(y) p_\theta(y)\mu(dy). \tag{IV.C.23}$$

On differentiating (IV.C.23) and applying condition (4), we have

$$\frac{\partial}{\partial\theta} E_\theta\{\hat{\theta}(Y)\} = \int_\Gamma \hat{\theta}(y) \frac{\partial}{\partial\theta} p_\theta(y)\mu(dy).$$

Condition (4) also implies that

$$\int_\Gamma \frac{\partial}{\partial\theta} p_\theta(y)\mu(dy) = \frac{\partial}{\partial\theta} \int_\Gamma p_\theta(y)\mu(dy) = \frac{\partial}{\partial\theta}(1) = 0,$$

so that we have

$$\begin{aligned} \frac{\partial}{\partial\theta} E_\theta\{\hat{\theta}(Y)\} &= \int_\Gamma (\hat{\theta}(y) - E_\theta\{\hat{\theta}(Y)\}) \frac{\partial}{\partial\theta} p_\theta(y)\mu(dy) \\ &= \int_\Gamma (\hat{\theta}(y) - E_\theta\{\hat{\theta}(Y)\}) \left[\frac{\partial}{\partial\theta} \log p_\theta(y) \right] p_\theta(y)\mu(dy) \\ &= E_\theta \left\{ [\hat{\theta}(Y) - E_\theta\{\hat{\theta}(Y)\}] \left[\frac{\partial}{\partial\theta} \log p_\theta(Y) \right] \right\}, \end{aligned} \tag{IV.C.24}$$

where the second equality follows from the fact that $\partial \log p_\theta(y)/\partial\theta = [\partial p_\theta(y)/\partial\theta]/p_\theta(y)$. Applying the Schwarz inequality to (IV.C.24), we have

$$\left(\frac{\partial}{\partial\theta} E_\theta\{\hat{\theta}(Y)\} \right)^2 \leq E_\theta\{[\hat{\theta}(Y) - E_\theta\{\hat{\theta}(Y)\}]^2\} I_\theta, \tag{IV.C.25}$$

where I_θ is from (IV.C.21). Noting that $E_\theta\{[\hat{\theta}(Y) - E_\theta\{\hat{\theta}(Y)\}]^2\} = \mathrm{Var}_\theta[\hat{\theta}(Y)]$, (IV.C.20) follows.

To see (IV.C.22), we note that

$$\begin{aligned} \frac{\partial^2}{\partial\theta^2} \log p_\theta(Y) &= \left(\frac{\partial^2}{\partial\theta^2} p_\theta(Y)/p_\theta(Y) \right) \\ &\quad - \left(\frac{\partial}{\partial\theta} \log p_\theta(Y) \right)^2. \end{aligned} \tag{IV.C.26}$$

Taking $E_\theta\{\cdot\}$ on both sides of (IV.C.26) and rearranging yields

$$I_\theta = -E_\theta \left(\frac{\partial^2}{\partial\theta^2} \log p_\theta(Y) \right) - \int_\Gamma \frac{\partial^2}{\partial\theta^2} p_\theta(y)\mu(dy).$$

Using condition (5) we have

$$\int_\Gamma \frac{\partial^2}{\partial\theta^2} p_\theta(y)\mu(dy) = \frac{\partial^2}{\partial\theta^2} \int_\Gamma p_\theta(y)\mu(dy) = \frac{\partial^2}{\partial\theta^2}(1) = 0,$$

and (IV.C.22) follows. □

The quantity I_θ defined in (IV.C.21) is known as *Fisher's information* for estimating θ from Y, and (IV.C.20) is called the *information inequality*. The higher this information measure is for a given model, the better is the lower bound on estimation accuracy provided by the information inequality. The existence of an estimate that achieves equality in the information inequality is possible only under special circumstances [see, e.g., Lehmann (1983) and the discussion below]. For the particular case in which $\hat{\theta}$ is unbiased $[E_\theta\{\hat{\theta}(Y)\} = \theta]$, the information inequality reduces to

$$\mathrm{Var}_\theta[\hat{\theta}(Y)] \geq \frac{1}{I_\theta}, \tag{IV.C.27}$$

a result known as the *Cramér-Rao lower bound* (CRLB).

Examples illustrating the information inequality in specific estimation problems will be discussed in the following section. The following general example illustrates further the role of exponential families in parameter estimation.

Example IV.C.4: The Information Inequality for Exponential Families

Suppose that Λ is an open interval and $p_\theta(y)$ is given by

$$p_\theta(y) = C(\theta) e^{g(\theta)T(y)} h(y), \tag{IV.C.28}$$

where C, g, T, and h are real-valued functions of their arguments and where $g(\theta)$ has derivative $g'(\theta)$. Assuming that $E_\theta\{|T(Y)|\} < \infty$ and

$$\frac{\partial}{\partial\theta}\int_\Gamma e^{g(\theta)T(y)}h(y)\mu(dy) = \int_\Gamma \frac{\partial}{\partial\theta}e^{g(\theta)T(y)}h(y)\mu(dy),$$

conditions (1)-(4) of Proposition IV.C.4 hold. Since $p_\theta(y)$ must integrate to unity, we can write $C(\theta) = [\int_\Gamma e^{g(\theta)T(y)}h(y)\mu(dy)]^{-1}$.

To compute I_θ for this family of densities, we write

$$\begin{aligned}\log p_\theta(y) &= g(\theta)T(y) + \log h(y) \\ &\quad - \log\left[\int_\Gamma e^{g(\theta)T(y)}h(y)\mu(dy)\right].\end{aligned}$$

On differentiating we have

$$\begin{aligned}\frac{\partial}{\partial\theta}\log p_\theta(y) &= g'(\theta)T(y) - \frac{g'(\theta)\int_\Gamma T(y)e^{g(\theta)T(y)}h(y)\mu(dy)}{\int_\Gamma e^{g(\theta)T(y)}h(y)\mu(dy)} \\ &= g'(\theta)[T(y) - E_\theta\{T(Y)\}].\end{aligned}$$

Thus

$$\begin{aligned}I_\theta \triangleq E_\theta\left\{\left(\frac{\partial}{\partial\theta}\log p_\theta(Y)\right)^2\right\} &= [g'(\theta)]^2 E_\theta\{[T(Y) - E_\theta\{T(Y)\}]^2\} \\ &= [g'(\theta)]^2\mathrm{Var}_\theta[T(Y)],\end{aligned}$$

and the information inequality in this case is

$$\mathrm{Var}_\theta[\hat\theta(Y)] \geq \frac{\left[\frac{\partial}{\partial\theta}E_\theta\{\hat\theta(Y)\}\right]^2}{[g'(\theta)]^2\mathrm{Var}_\theta[T(Y)]}. \tag{IV.C.29}$$

Suppose that we consider $T(y)$ itself as an estimator of θ. Then we have

$$E_\theta\{T(Y)\} = \frac{\int_\Gamma T(y)e^{g(\theta)T(y)}h(y)\mu(dy)}{\int_\Gamma e^{g(\theta)T(y)}h(y)\mu(dy)}. \tag{IV.C.30}$$

On differentiating (IV.C.30) we have straightforwardly that

$$\frac{\partial}{\partial\theta}E_\theta\{T(Y)\} = g'(\theta)\mathrm{Var}_\theta[T(Y)],$$

and thus (IV.C.29) implies that the lower bound in the information inequality equals

$$\frac{[\frac{\partial}{\partial\theta}E_\theta\{T(Y)\}]^2}{[g'(\theta)]^2\mathrm{Var}_\theta[T(Y)]} = \mathrm{Var}_\theta[T(Y)]. \tag{IV.C.31}$$

From (IV.C.31) we see that $T(Y)$ achieves the information lower bound, so it has minimum variance among all estimators $\hat\theta$ satisfying $\partial E_\theta\{\hat\theta(Y)\}/\partial\theta = \partial E_\theta\{T(Y)\}/\partial\theta$. In particular, if T is unbiased for θ, then it is an MVUE, a fact that we know already from the fact that T is a complete sufficient statistic for θ in this case.

We see that the exponential form (IV.C.28) is sufficient for the variance of T to achieve the information lower bound within the regularity assumed above. It turns out that this form is also necessary for achieving the lower bound for all $\theta \in \Lambda$, again within regularity conditions. In particular, we note that an estimator $\hat\theta$ has variance equal to the information lower bound for all $\theta \in \Lambda$ if and only if we have equality in the Schwarz inequality applied in (IV.C.25). This, in turn, will happen if and only if

$$\frac{\partial}{\partial\theta}\log p_\theta(Y) = k(\theta)[\hat\theta(Y) - E_\theta\{\hat\theta(Y)\}]$$

with probability 1 under P_θ, for some $k(\theta)$. Letting (a, b) denote Λ and $f(\theta)$ denote $E_\theta\{\hat\theta(Y)\}$, we thus conclude that $\hat\theta$ achieves the information bound if and only if

$$p_\theta(y) = h(y)\exp\left\{\int_a^\theta k(\sigma)[\hat\theta(y) - f(\sigma)]d\sigma\right\}, \quad y \in \Gamma, \tag{IV.C.32}$$

where $h(y)$ does not depend on θ. Equation (IV.C.32) will be recognized as the exponential form of (IV.C.28) with h as given,

$$\begin{aligned}C(\theta) &= \exp\{-\int_a^\theta k(\sigma)f(\sigma)d\sigma\}, \\ g(\theta) &= \int_a^\theta k(\sigma)d\sigma,\end{aligned}$$

and

$$T(y) = \hat\theta(y).$$

[Note that $k(\theta)$ must be equal to $I_\theta/[\partial E_\theta\{\hat\theta(Y)\}/\partial\theta]$ in this situation, as can be seen from substituting (IV.C.32) into (IV.C.24).] Thus we conclude that, within regularity, *the information lower bound is achieved by $\hat\theta$ if and only if $\hat\theta(y) = T(y)$ in a one-parameter exponential family.*

IV.D Maximum-Likelihood Estimation

For many observation models arising in practice, it is not possible to apply the results of Section IV.C to find MVUEs, either because of intractability

of the required analysis or because of the lack of a useful complete sufficient statistic. For such models, an alternative method for seeking good estimators is needed. One very commonly used method of designing estimators is the maximum-likelihood method, which is the subject of this section.

To motivate maximum-likelihood estimation, we first consider MAP estimation in which we seek $\hat{\theta}_{MAP}(y)$ given by

$$\hat{\theta}_{MAP}(y) = \arg\{\max_{\theta\in\Lambda} p_\theta(y)w(\theta)\}. \tag{IV.D.1}$$

In the absence of any prior information about the parameter, we might assume that it is uniformly distributed in its range [i.e., $w(\theta)$ is constant on Λ] since this represents more or less a worst-case prior. In this case, the MAP estimate for a given $y \in \Gamma$ is any value of θ that maximizes $p_\theta(y)$ over Λ. Since $p_\theta(y)$ as a function of θ is sometimes called the *likelihood function* [hence, $p_1(y)/p_0(y)$ is the likelihood ratio], this estimate is called the *maximum likelihood estimate* (MLE). Denoting this estimate by $\hat{\theta}_{ML}$, we have

$$\hat{\theta}_{ML}(y) = \arg\{\max_{\theta\in\Lambda} p_\theta(y)\}. \tag{IV.D.2}$$

There are two things wrong with the above argument. First, it is not always possible to construct a uniform distribution on Λ, since Λ may not be a bounded set. Second, and more important, assuming a uniform prior for the parameter is different from assuming that the prior is unknown or that the parameter is not a random variable. However, the maximum-likelihood estimate turns out to be very useful in many situations, and as we will see in this section, its use can be motivated in other, more direct, ways. Moreover, finding the value of θ that makes the observations most likely is a legitimate criterion on its own.

Maximizing $p_\theta(y)$ is equivalent to maximizing $\log p_\theta(y)$, and assuming sufficient smoothness of this function, a necessary condition for the maximum-likelihood estimate is

$$\frac{\partial}{\partial\theta}\log p_\theta(y)\Big|_{\theta=\hat{\theta}_{ML}(y)} = 0. \tag{IV.D.3}$$

Equation (IV.D.3) is known as the *likelihood equation* , and we will see that its solutions have useful properties even when they are not maxima of $p_\theta(y)$.

For example, suppose we have equality in the Cramer-Rao lower bound (IV.C.27); i.e., suppose that $\hat{\theta}$ is an unbiased estimate of θ with $\mathrm{Var}_\theta[\hat{\theta}(Y)] = 1/I_\theta$. (Note that such a $\hat{\theta}$ is an MVUE of θ.) Then, from (IV.C.32), we see that $\log p_\theta(y)$ must be of the form

$$\log p_\theta(y) = \int_a^\theta I_\sigma[\hat{\theta}(y) - \sigma]d\sigma + \log h(y), \tag{IV.D.4}$$

where we have used the facts that $f(\theta) = \theta$ and $k(\theta) = I_\theta/f'(\theta)$. From (IV.D.4), the likelihood equation becomes

$$\frac{\partial}{\partial\theta}\log p_\theta(y)\Big|_{\theta=\hat{\theta}_{ML}(y)} = I_\theta[\hat{\theta}(y) - \theta]\Big|_{\theta=\hat{\theta}_{ML}(y)} = 0, \tag{IV.D.5}$$

which has the solution $\hat{\theta}_{ML}(y) = \hat{\theta}(y)$. Thus we conclude that *if $\hat{\theta}$ achieves the CRLB, it is the solution to the likelihood equation.* In other words, only solutions to the likelihood equation can achieve the CRLB. Unfortunately, it is not always true that solutions to the likelihood equation will achieve the CRLB or even that they are unbiased. [However, when $\log p_\theta$ has the form (IV.D.4), this will happen.] Also, when the solution to the likelihood equation does not satisfy the CRLB, there may be other estimators with the same bias that have smaller variance then $\hat{\theta}_{ML}$.

From the above discussion we see that the solution to the likelihood equation can sometimes be an MVUE. For the case in which the observation space is $\mathbb{R}^n$ with Y consisting of i.i.d. components, it happens that within regularity the solution to the likelihood equation is unbiased and achieves the CRLB asymptotically as $n \to \infty$. Before studying these asymptotic properties we give the following two examples to illustrate the maximum-likelihood approach.

Example IV.D.1: Maximum-Likelihood Estimation of the Parameter of the Exponential Distribution

Suppose that $\Gamma = \mathbb{R}^n, \Lambda = (0,\infty)$, and $Y_1,\ldots,Y_n$ are i.i.d. exponential random variables with parameter θ, i.e., $p_\theta(\underline{y}) = \Pi_{k=1}^n f_\theta(y_k)$ with

$$f_\theta(y_k) = \begin{cases} \theta e^{-\theta y_k} & \text{if } y_k \geq 0 \\ 0 & \text{if } y_k < 0. \end{cases} \tag{IV.D.6}$$

We have $p_\theta(y) = \theta^n \exp\{-\theta n\bar{y}\}$ with $\bar{y}\stackrel{\Delta}{=}(1/n)\sum_{k=1}^n y_k$, so the likelihood equation is

$$\frac{\partial}{\partial\theta}\log p_\theta(\underline{y})\Big|_{\theta=\hat{\theta}_{ML}(y)} = \frac{n}{\theta} - n\bar{y}\Big|_{\theta=\hat{\theta}_{ML}(\underline{y})} = 0, \tag{IV.D.7}$$

which has the unique solution $\hat{\theta}_{ML}(\underline{y}) = 1/\bar{y}$. Since $\partial^2\log p_\theta(\underline{y})/\partial\theta^2 = -n/\theta^2 < 0$, this solution gives the unique maximum of $p_\theta(\underline{y})$. Note that $E_\theta\{Y_k\} = 1/\theta$, so that $E_\theta\{\bar{Y}\} = 1/\theta$ and thus it makes sense to estimate θ as $1/\bar{y}$. In fact, the weak law of large numbers implies that $\bar{Y} \to 1/\theta$ in probability under P_θ, which in turn implies that $1/\bar{Y} \to \theta$ in probability under P_θ; i.e., the MLE converges in probability to the true parameter value, a property known as *consistency*. This property of MLEs is not specific to this example but rather is true in a very general context as we shall see below.

Fisher's information for this case can be computed via

$$I_\theta = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log p_\theta(\underline{Y})\right] = -E_\theta\{-n/\theta^2\} = n/\theta^2,$$

and so the CRLB is θ^2/n. Since $\partial \log p_\theta(\underline{y})/\partial\theta$ is not of the form $k(\theta)[\hat{\theta}_{ML}(\underline{y}) - f(\theta)]$, we know that the information inequality is not achieved in this problem. However, we can compute the mean and variance of $\hat{\theta}_{ML}$ directly. In particular, by using characteristic functions it is straightforward to show that the sample mean $\overline{Y}$ has pdf

$$p_{\overline{Y}}(\overline{y}) = \begin{cases} \frac{(n\theta)^n}{n!}\overline{y}^{n-1}e^{-n\theta\overline{y}} & \text{if } \overline{y} \geq 0 \\ 0 & \text{if } \overline{y} < 0, \end{cases}$$

from which we can compute (for $n > 1$)

$$E_\theta\{\hat{\theta}_{ML}(\underline{Y})\} = E_\theta\left\{\frac{1}{\overline{Y}}\right\} = \frac{n\theta}{n-1} \tag{IV.D.8}$$

and (for $n > 2$)

$$\mathrm{Var}_\theta[\hat{\theta}_{ML}(\underline{Y})] = \frac{\theta^2 n^2}{(n-1)^2(n-2)}. \tag{IV.D.9}$$

We see from (IV.D.8) that although $\hat{\theta}_{ML}(\underline{Y})$ is biased, it does have the property that $\lim_{n\to\infty} E_\theta\{\hat{\theta}_{ML}(\underline{Y})\} = \theta$; that is, it is *asymptotically unbiased.* Also, we note that

$$\mathrm{Var}_\theta[\hat{\theta}_{ML}(\underline{Y})]I_\theta = \frac{n^3}{(n-1)^2(n-2)} \to 1$$

as $n \to \infty$; and thus $\hat{\theta}_{ML}$ has variance asymptotically equal to the CRLB, a property known as *asymptotic efficiency.* As we shall see, these two properties of asymptotic unbiasedness and efficiency are characteristic of MLEs under general conditions for i.i.d. observations.

As a final comment on this example, we note from Proposition IV.C.3 that $\overline{Y}$ is a complete sufficient statistic for θ in this model. Also, from (IV.D.8) we see that

$$\frac{n-1}{n}\hat{\theta}_{ML}(\underline{y}) \equiv \left(\frac{1}{n-1}\sum_{k=1}^{n} y_k\right)^{-1}$$

is an unbiased estimator of θ depending on $\overline{Y}$. Thus

$$\frac{n-1}{n}\hat{\theta}_{ML}(\underline{Y}) \triangleq \hat{\theta}_{MV}(\underline{y})$$

is an MVUE of θ in this problem. From (IV.D.9), its variance is seen to be given by

$$\mathrm{Var}_\theta[\hat{\theta}_{MV}(\underline{Y})] = \frac{\theta^2}{n-2}, \tag{IV.D.10}$$

a quantity that is larger than the CRLB (as it must be since we know the CRLB cannot be achieved here), but that approaches the CRLB as n becomes large.

The variance of (IV.D.10) equals the MSE of $\hat{\theta}_{MV}$ since it is unbiased. For the MLE, the MSE is

$$E_\theta\{[\hat{\theta}_{ML}(\underline{Y}) - \theta]^2\} = \mathrm{Var}_\theta[\hat{\theta}_{ML}(\underline{Y})] + b^2(\theta), \tag{IV.D.11}$$

where $b(\theta) \triangleq E_\theta\{\hat{\theta}_{ML}(\underline{Y})\} - \theta$ is the *bias* of $\hat{\theta}_{ML}$. Using (IV.D.8) and (IV.D.9), we have

$$E_\theta\{[\hat{\theta}_{ML}(\underline{Y}) - \theta]^2\} = \frac{\theta^2(n+2)}{(n-1)(n-2)},$$

a quantity that is strictly greater than $\theta^2/(n-2)$, the MSE of $\hat{\theta}_{MV}$. Thus in this case the MVUE is preferable to the MLE, although they are asymptotically equivalent.

Example IV.D.2: Maximum-Likelihood Estimation of Signal Amplitude

Consider the model treated in Example IV.C.3:

$$Y_k = N_k + \mu s_k, \quad k = 1, \ldots, n$$

with $N_1, \ldots, N_n$ i.i.d. $\mathcal{N}(0, \sigma^2)$ and $\underline{s} = (s_1, \ldots, s_n)^T$ known. The likelihood equation for estimating μ with σ^2 known is given by

$$\begin{aligned} 0 &= \frac{\partial}{\partial\mu}\left(\frac{1}{2}\sum_{k=1}^{n}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{k=1}^{n}(y_k - \mu s_k)^2\right)\Big|_{\mu=\hat{\mu}_{ML}(\underline{y})} \\ &= \frac{1}{\sigma^2}\sum_{k=1}^{n} s_k[y_k - \hat{\mu}_{ML}(\underline{y})s_k] = 0, \end{aligned} \tag{IV.D.12}$$

which implies that

$$\hat{\mu}_{ML}(\underline{y}) = \frac{1}{n}\sum_{k=1}^{n} s_k y_k / \overline{s^2}, \tag{IV.D.13}$$

where, as before, $\overline{s^2} \triangleq (1/n)\sum_{k=1}^{n} s_k^2$. Since

$$-\frac{\partial^2}{\partial\mu^2}\sum_{k=1}^{n}\left(\frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(y_k - \mu s_k)^2\right) = -n\overline{s^2}/\sigma^2 < 0, \tag{IV.D.14}$$

we see that $\log p_\theta(\underline{y})$ is concave in μ, so the solution to the likelihood equation does give a global maximum here.

Note that $\hat{\mu}_{ML}$ is the same as the MVUE of μ (see Example IV.C.3), so that $E_\theta\{\hat{\mu}_{ML}(\underline{Y})\} = \mu$ and $\mathrm{Var}_\theta[\hat{\mu}_{ML}(\underline{Y})] = \sigma^2/n\overline{s^2}$. From (IV.D.14), we see that $I_\theta = n\overline{s^2}/\sigma^2$, so

$$\mathrm{CRLB} = \frac{\sigma^2}{n\overline{s^2}} = \mathrm{Var}_\theta[\hat{\mu}_{ML}(\underline{Y})]. \tag{IV.D.15}$$

Note that with $\theta = \mu$, we can write

$$\frac{\partial}{\partial\theta}\log p_\theta(\underline{y}) = k(\theta)[\hat{\theta}_{ML}(\underline{y}) - \theta]$$

with $k(\theta) = I_\theta = n\overline{s^2}/\sigma^2$, as is required for achievement of the CRLB.

Suppose that now that μ is known but we wish to estimate σ^2. The likelihood equation becomes

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\log p_\theta(\underline{y})\Big|_{\sigma^2=\hat{\sigma}^2_{ML}(\underline{y})} \\ = \frac{1}{2\hat{\sigma}^2_{ML}(\underline{y})} - \frac{1}{2[\hat{\sigma}^2_{ML}(\underline{y})]^2}\sum_{k=1}^{n}(y_k - \mu s_k)^2 = 0,\end{aligned} \tag{IV.D.16}$$

which has the unique solution

$$\hat{\sigma}^2_{ML}(\underline{y}) = \frac{1}{n}\sum_{k=1}^{n}(y_k - \mu s_k)^2. \tag{IV.D.17}$$

Since

$$\frac{\partial}{\partial\sigma^2}\log p_\theta(\underline{y}) = \frac{n}{2\sigma^4}\left[\hat{\sigma}^2_{ML}(\underline{y}) - \sigma^2\right], \tag{IV.D.18}$$

we see that $\log p_\theta(\underline{y})$ is increasing in σ^2 for $\sigma^2 < \hat{\sigma}^2_{ML}(\underline{y})$ and decreasing in σ^2 for $\sigma^2 > \hat{\sigma}^2_{ML}(\underline{y})$. Thus $\log p_\theta(\underline{y})$ achieves its absolute maximum at $\hat{\sigma}^2_{ML}(\underline{y})$. We also see from (IV.D.18) that with $\theta = \sigma^2$,

$$\frac{\partial}{\partial\theta}\log p_\theta(\underline{y}) = \frac{n}{2\theta^2}[\hat{\theta}_{ML}(\underline{y}) - \theta], \tag{IV.D.19}$$

which from Example IV.C.4 implies that $\hat{\sigma}^2_{ML}(\underline{y})$ is unbiased and achieves the CRLB, and thus that $\hat{\sigma}^2_{ML}$ is an MVUE of σ^2. By inspection of (IV.D.19) we have $I_\theta = n/2\theta^2 \equiv n/2\sigma^4$, so

$$\mathrm{CRLB} = \frac{2\sigma^4}{n} = \mathrm{Var}_\theta[\hat{\sigma}^2_{ML}(\underline{Y})]. \tag{IV.D.20}$$

Now suppose that both μ and σ^2 are unknown. Putting $\theta = (\mu, \sigma^2)$, the MLE of θ is found by maximizing $p_\theta(\underline{y})$ over μ and σ^2. Since the maximum $\hat{\mu}_{ML}(\underline{y})$ from (IV.D.13) does not depend on σ^2, we have that

$$\begin{aligned}\max_{(\mu,\sigma^2)}\log p_\theta(\underline{y}) &= \max_{\sigma^2}\{\max_{\mu}\log p_\theta(\underline{y})\} \\ &= \max_{\sigma^2}\left\{-\frac{1}{2}\sum_{k=1}^{n}\log(2\pi\sigma^2)\right. \\ &\qquad \left. - \frac{1}{2\sigma^2}\sum_{k=1}^{n}[y_k - \hat{\mu}_{ML}(\underline{y})s_k]^2\right\}.\end{aligned}$$

But the right-hand side of this equation is the same maximization problem as for estimating σ^2 with known μ, with μ set equal to $\hat{\mu}_{ML}(\underline{y})$. Thus the maximum is achieved by (IV.D.17) with $\hat{\mu}_{ML}$ substituted for μ and the MLE for $\theta = (\mu, \sigma^2)$ is $\hat{\theta}_{ML}(\underline{y}) = [\hat{\mu}_{ML}(\underline{y}), \hat{\sigma}^2_{ML}(\underline{y})]$, where

$$\hat{\mu}_{ML}(\underline{y}) = \frac{1}{n}\sum_{k=1}^{n}s_k y_k/\overline{s^2} \tag{IV.D.21a}$$

and

$$\hat{\sigma}^2_{ML}(\underline{y}) = \frac{1}{n}\sum_{k=1}^{n}[y_k - \hat{\mu}_{ML}(\underline{y})s_k]^2. \tag{IV.D.21b}$$

The estimate $\hat{\mu}_{ML}(\underline{y})$ is still an MVUE of μ in this case. However, from (IV.C.19) we see that $\hat{\sigma}^2_{ML}(\underline{y})$ is $[(n-1)/n]\hat{\sigma}^2_{MV}(\underline{y})$. Thus

$$E_\theta\{\hat{\sigma}^2_{ML}(\underline{Y})\} = \frac{n-1}{n}\sigma^2,$$

and the MLE of σ^2 is biased here (although it is asymptotically unbiased). Note that

$$\mathrm{Var}_\theta[\hat{\sigma}^2_{ML}(\underline{Y})] = [(n-1)^2/n^2]\mathrm{Var}_\theta[\hat{\sigma}^2_{MV}(\underline{Y})],$$

so that $\hat{\sigma}^2_{ML}(\underline{y})$ has lower variance than the MVUE. It can be shown that (see Exercise 14.)[6]

$$\mathrm{Var}_\theta(\hat{\sigma}^2_{MV}(\underline{Y})) = \frac{2\sigma^4}{n-1}. \tag{IV.D.22}$$

[6]It is interesting to note that the MVUE of σ^2 with μ known has variance $2\sigma^4/n$ [from (IV.D.20)] and the MVUE with μ unknown has variance $2\sigma^4/(n-1)$ [from (IV.D.22)]. Thus for unbiased estimation of σ^2, there is a "penalty" of one observation when μ is unknown.

Thus for the MVUE of σ^2 the MSE, $E_\theta\{[\hat{\sigma}^2_{MV}(\underline{Y}) - \sigma^2]^2\}$, is $2\sigma^4/(n-1)$. Alternatively, for the MLE of σ^2, the MSE is given by

$$\begin{aligned} E_\theta\{[\sigma^2_{ML}(\underline{Y}) - \sigma^2]^2\} &= \mathrm{Var}_\theta[\sigma^2_{ML}(\underline{Y})] + [E_\theta\{\hat{\sigma}^2_{ML}(\underline{Y})\} - \sigma^2]^2 \\ &= \frac{(n-1)^2}{n^2}\frac{2\sigma^4}{n-1} + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 \\ &= \sigma^4\frac{(2n-1)}{n^2}. \end{aligned} \tag{IV.D.23}$$

The ratio of these two quantities is

$$\frac{E_\theta\{[\hat{\sigma}^2_{MV}(\underline{Y}) - \sigma^2]^2\}}{E_\theta\{[\hat{\sigma}^2_{ML}(\underline{Y}) - \sigma^2]^2\}} = \left(\frac{n}{n-1}\right)\left(\frac{2n}{2n-1}\right) > 1. \tag{IV.D.24}$$

We see from (IV.D.24) that *the MLE in this case has a uniformly lower MSE than the MVUE.* This is because the increase in MSE due to the bias of the MLE is more than offset by the increase in variance of the MVUE needed to achieve unbiasedness. Thus, achieving the goal of minimum-variance unbiased estimation does not always lead to an optimum estimate in terms of mean-squared error.

One of the principal motivations for using maximum-likelihood estimation is illustrated by the two examples above; namely, estimates based on independent samples have good asymptotic properties as the number of samples increases without bound. The reason for this asymptotic behavior can be seen from the arguments in the following paragraphs.

Suppose that we have a sequence of i.i.d. observations $Y_1, Y_2, \ldots, Y_n$, each with marginal density f_θ coming from the family $\{f_\theta; \theta \in \Lambda\}$. Let $\hat{\theta}_n$ denote a solution to the likelihood equation for sample size n, i.e.,

$$\frac{\partial}{\partial\theta}\log p_\theta(\underline{y})\Big|_{\theta=\hat{\theta}_n(\underline{y})} = \sum_{k=1}^{n}\psi[y_k; \hat{\theta}_n(\underline{y})] = 0,$$

where $\psi(y_k;\theta) \triangleq \partial \log f_\theta(y_k)/\partial\theta$. Equivalently, we can write

$$\frac{1}{n}\sum_{k=1}^{n}\psi[y_k; \hat{\theta}_n(\underline{y})] = 0. \tag{IV.D.25}$$

For a fixed parameter value $\theta' \in \Lambda$, consider the quantity $\sum_{k=1}^{n}\psi(Y_k;\theta')/n$. Assuming that θ is the true parameter value (i.e., $Y_k \sim f_\theta$), the weak law of large numbers implies that

$$\frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta') \xrightarrow{i.p.} E_\theta\{\psi(Y_1;\theta')\}.$$

We have

$$\begin{aligned} E_\theta\{\psi(Y_1;\theta')\} &= \int_{\mathbf{R}}\frac{\partial}{\partial\theta}\log f_\theta(y_1)\,|_{\theta=\theta'}\, f_\theta(y_1)\mu(dy_1) \\ &\triangleq J(\theta;\theta'). \end{aligned} \tag{IV.D.26}$$

Assuming that the order of integration and differentiation can be interchanged in (IV.D.26), $J(\theta;\theta)$ can be written as

$$\begin{aligned} J(\theta;\theta) &= \int\left[\frac{\partial}{\partial\theta}\log f_\theta(y_1)\right] f_\theta(y_1)\mu(dy_1) \\ &= \int\frac{\partial}{\partial\theta}f_\theta(y_1)\mu(dy_1) \\ &= \frac{\partial}{\partial\theta}\int f_\theta(y_1)\mu(dy_1) = \frac{\partial}{\partial\theta}(1) = 0. \end{aligned}$$

Thus the equation $J(\theta;\theta') = 0$ has a solution $\theta' = \theta$. Suppose that this is the unique root of $J(\theta;\theta')$, and suppose that $J(\theta;\theta')$ and $\sum_{k=1}^{n}\psi(Y_k;\theta')/n$ are both smooth functions of θ'. Then, since $\sum_{k=1}^{n}\psi(Y_k;\theta')/n$ is close to $J(\theta;\theta')$ for large n, we would expect the roots of these two functions to be close when n is large. That is, $\hat{\theta}_n(\underline{Y})$ should be close to the true parameter value θ when n is large. And as $n \to \infty$, we would expect that $\hat{\theta}_n(\underline{Y}) \to \theta$ in some statistical sense. In fact, within the appropriate smoothness and uniqueness conditions, the solutions to the likelihood equation are *consistent* ; that is, they converge in probability to the true parameter value:

$$\lim_{n\to\infty} P_\theta(|\hat{\theta}_n(\underline{Y}) - \theta| > \epsilon) = 0 \text{ for all } \epsilon > 0.$$

One set of conditions under which solutions to the likelihood equation are consistent is summarized in the following.

Proposition IV.D.1: Consistency of MLEs

Suppose that $\{Y_k\}_{k=1}^{\infty}$ is an i.i.d. sequence of random variables each with density f_θ, and assume that J and ψ are well defined as above. Suppose further that the following conditions hold:

(1) $J(\theta;\theta')$ is a continuous function of θ' and has a unique root at $\theta' = \theta$, at which point it changes sign.

(2) $\psi(Y_k;\theta')$ is a continuous function of θ' (with probability 1).

(3) For each n, $\sum_{k=1}^{n}\psi(Y_k;\theta')/n$ has a unique root $\hat{\theta}_n$ (with probability 1).

Then $\hat{\theta}_n \to \theta$ (i.p.).

Proof: Choose $\epsilon > 0$. By condition (1), $J(\theta;\theta+\epsilon)$ and $J(\theta;\theta-\epsilon)$ must have opposite signs. Define $\delta = \min\{|J(\theta;\theta+\epsilon)|, |J(\theta;\theta-\epsilon)|\}$ and for each

n, define the events

$$\begin{aligned} A_n^+ &= \{|J(\theta;\theta+\epsilon) - \frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta+\epsilon)| \le \delta\}, \\ A_n^- &= \{|J(\theta;\theta-\epsilon) - \frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta-\epsilon)| \le \delta\}, \end{aligned} \quad \text{(IV.D.27)}$$

and $A_n = A_n^+ \cap A_n^-$.

Now, on A_n^+, $\sum_{k=1}^n \psi(Y_k, \theta+\epsilon)/n$ must have the same sign as $J(\theta;\theta+\epsilon)$, and on A_n^-, $\sum_{k=1}^n (Y_k;\theta-\epsilon)/n$ must have the same sign as $J(\theta;\theta-\epsilon)$. Thus on A_n, $\sum_{k=1}^n \psi(Y_k;\theta+\epsilon)/n$ and $\sum_{k=1}^n \psi(Y_k;\theta-\epsilon)/n$ have opposite signs. By the continuity assumption (2), $\sum_{k=1}^n \psi(Y_k;\theta')/n$ can change sign only by passing through zero. Thus on A_n, the root $\hat{\theta}_n$ is between $\theta-\epsilon$ and $\theta+\epsilon$. This implies that A_n is a subset of $\{|\hat{\theta}_n - \theta| \le \epsilon\}$, so that $P(|\hat{\theta}_n - \theta| \le \epsilon) \ge P(A_n)$.

By the weak law of large numbers,

$$\frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta+\epsilon) \to J(\theta;\theta+\epsilon)(i.p.)$$

and

$$\frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta-\epsilon) \to J(\theta;\theta-\epsilon)(i.p.). \quad \text{(IV.D.28)}$$

Thus $P(A_n^+) \to 1$ and $P(A_n^-) \to 1$ as $n \to \infty$. We have

$$\begin{aligned} 1 &\ge P(|\hat{\theta}_n - \theta| \le \epsilon) \\ &\ge P(A_n) = P(A_n^+) + P(A_n^-) - P(A_n^+ \cup A_n^-) \\ &\ge P(A_n^+) + P(A_n^-) - 1 \to 1. \end{aligned} \quad \text{(IV.D.29)}$$

Thus $P(\hat{\theta}_n - \theta| \le \epsilon) \to 1$, and since ϵ was chosen arbitrarily we have the desired result. □

Remarks: The conditions on this proposition can be relaxed in various ways. First, the continuity of the functions $J(\theta;\theta')$ and $\psi(Y_k;\theta')$ can be relaxed to continuity in a neighborhood of $\theta' = \theta$. Also, it is not necessary to assume the existence of the roots $\hat{\theta}_n$, since the development above shows that there must be a root to the likelihood equation on A_n, which has probability tending to 1. In fact, with only the assumption of local continuity, the proof above can be used to show that *with probability tending to 1, there is a sequence of roots to the likelihood equation converging to any isolated root of* $J(\theta;\theta')$. Thus if $J(\theta;\theta')$ has multiple roots, inconsistent sequences can arise by solving the likelihood equation.

In addition to consistency, we saw in the examples above that the solutions to the likelihood equation may also be asymptotically unbiased and efficient. We know that under the conditions of Proposition IV.D.1, $\hat{\theta}_n$ converges to θ in probability. Thus if we would write

$$\lim_{n\to\infty} E_\theta\{\hat{\theta}_n\} = E_\theta\{\lim_{n\to\infty}\hat{\theta}_n\} \quad \text{(IV.D.30)}$$

for this type of convergence, then asymptotic unbiasedness would follow. The interchange of limits and expectations in (IV.D.30) is not always valid for convergence in probability. However, under various conditions on ψ, this interchange can be shown to be valid. (A sufficient condition for the validity of this interchange is the existence of a random variable X such that $|\hat{\theta}_n| \le X$ for each n and $E_\theta\{X\} < \infty$. This is known as the *dominated convergence theorem.*) Thus asymptotic unbiasedness is not an unreasonable property to expect in view of the consistency of $\hat{\theta}_n$.

It is less clear why MLEs might be asymptotically efficient. To see why this might be so, we consider the related question of finding the asymptotic distribution of the error, $\hat{\theta}_n - \theta$. In particular, we prove the following proposition.

Proposition IV.D.2: Asymptotic Normality of MLEs

Suppose that $\{Y_k\}_{k=1}^\infty$ is a sequence of i.i.d. random variables each with density f_θ, and that $\{\hat{\theta}_n\}_{n=1}^\infty$ is a consistent sequence of roots of the likelihood equation. Suppose further that ψ satisfies the following regularity conditions.

(1) $0 < i_\theta \triangleq E_\theta\{[\psi(Y_1;\theta)]^2\} < \infty$.

(2) The derivatives $\psi'(Y_1;\theta') \triangleq \partial\psi(Y_1;\theta')/\partial\theta'$ and $\psi''(Y_k;\theta') \triangleq \partial^2\psi(Y_k;\theta')/(\partial\theta')^2$ exist (with probability 1).

(3) There is a function $M(Y_1)$ such that $|\psi''(Y_1;\theta')| \le M(Y_1)$ for all $\theta' \in \Lambda$ and $E_\theta\{M(Y_1)\} < \infty$.

(4) $J(\theta;\theta) = 0$, where $J(\theta;\theta')$ is defined as in (IV.D.26).

(5) Condition (5) of Proposition IV.C.4 holds.

Then

$$P_\theta(\sqrt{ni_\theta}(\hat{\theta}_n - \theta) \le x) \to \Phi(x) \text{ for all } x \in \mathbb{R},$$

where Φ is the standard Gaussian distribution function. That is, $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution to a $\mathcal{N}(0, 1/i_\theta)$ random variable.

Proof: Using Taylor's theorem, we can expand the left-hand side of the likelihood equation, $(1/n)\sum_{k=1}^{n}\psi(Y_k;\hat{\theta}_n)=0$, about θ to yield

$$\frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta) + (\hat{\theta}_n-\theta)\frac{1}{n}\sum_{k=1}^{n}\psi'(Y_k;\theta) + \frac{1}{2}(\hat{\theta}_n-\theta)^2\frac{1}{n}\sum_{k=1}^{n}\psi''(Y_k;\bar{\theta}_n)=0, \quad \text{(IV.D.31)}$$

where $\bar{\theta}_n$ is between θ and $\hat{\theta}_n$. Rearranging (IV.D.31) gives an expression for the quantity $\sqrt{n}(\hat{\theta}_n-\theta)$:

$$\sqrt{n}(\hat{\theta}_n-\theta)=\frac{-\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\psi(Y_k;\theta)}{\frac{1}{n}\sum_{k=1}^{n}\psi'(Y_k;\theta)+(\hat{\theta}_n-\theta)\frac{1}{2n}\sum_{k=1}^{n}\psi''(Y_k;\bar{\theta}_n)}. \quad \text{(IV.D.32)}$$

Consider the denominator on the right-hand side of (IV.D.32). By the weak law of large numbers, the first term, $\sum_{k=1}^{n}\psi'(Y_k;\theta)/n$, converges to $E_\theta\{\psi'(Y_1;\theta)\}$ in probability. By condition (3), the second term $(\hat{\theta}_n-\theta)\sum_{k=1}^{n}\psi''(Y_k;\bar{\theta}_n)/2n$ is bounded as

$$\left|\tfrac{1}{2}(\hat{\theta}_n-\theta)\tfrac{1}{n}\textstyle\sum_{k=1}^{n}\psi''(Y_k;\hat{\theta}_n)\right| \le \tfrac{1}{2}|\hat{\theta}_n-\theta|\tfrac{1}{n}\textstyle\sum_{k=1}^{n}M(Y_k). \quad \text{(IV.D.33)}$$

Now, $|\bar{\theta}_n-\theta|\to 0$ (i.p.) and the weak law of large numbers implies that $(1/n)\sum_{k=1}^{n}M(Y_k)\to E_\theta\{M(Y_1)\}<\infty$. Thus the second term converges in probability to zero and the denominator then converges in probability to $E_\theta\{\psi'(Y_1;\theta\}$.

The numerator sum $\sum_{k=1}^{n}\psi(Y_k;\theta)$ in (IV.D.32) is the sum of n i.i.d. random variables, each with mean $E_\theta\{\psi(Y_1;\theta)\}=J(\theta;\theta)=0$ and variance $E_\theta\{\psi^2(Y_1;\theta)\}=i_\theta<\infty$. Thus by the central limit theorem, $-(1/\sqrt{n})\sum_{k=1}^{n}\psi(Y_k;\theta)$ converges in distribution to a $\mathcal{N}(0,i_\theta)$ random variable.

The two results above imply that $\sqrt{n}(\hat{\theta}_n-\theta)$ converges in distribution to a $\mathcal{N}(0,v^2)$ random variable with

$$v^2=i_\theta/E_\theta^2\{\psi'(Y_1;\theta)\}. \quad \text{(IV.D.34)}$$

But using the argument used in deriving (IV.C.22), $E_\theta\{\psi'(Y_k;\theta)\}=-E_\theta\{\psi^2(Y_1;\theta)\}=-i_\theta$, so $v^2=1/i_\theta$. This completes the proof. □

Remarks: It is easy to see that Fisher's information is given by $I_\theta=ni_\theta$ for this i.i.d. case. Heuristically, we can think of the conclusion of this proposition as the condition that $\hat{\theta}_n$ is asymptotically $\mathcal{N}(\theta,\,1/ni_\theta)$; that is, asymptotically $\hat{\theta}_n$ has mean θ and variance equal to $1/ni_\theta$, the Cramér-Rao lower bound. Actually, what we have proved is that the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n-\theta)$ has zero mean and variance $1/i_\theta$, which is not the same as $E_\theta\{\sqrt{n}(\hat{\theta}_n-\theta)\}\to 0$ and $\mathrm{Var}_\theta[\sqrt{n}(\hat{\theta}_n-\theta)]\to 1/i_\theta$. The latter two conditions (the second of which is asymptotic efficiency) may, in fact, hold; however, additional conditions are required to assume this. [These properties can be examined via (IV.D.32).] Nevertheless, the conclusion of Proposition IV.D.2 is sufficient practical justification for considering the MLE to be an asymptotically optimum (MVUE) estimator. And, in fact, asymptotic unbiasedness and efficiency are often alternatively defined in terms of the mean and variance of the asymptotic error distribution.

IV.E Further Aspects and Extensions of Maximum-Likelihood Estimation

IV.E.1 Estimation of Vector Parameters

It should be noted that all of the analysis of the preceding section can be generalized to the case in which the parameter is a vector, say of dimension m. In this case, the likelihood equation is a vector equation

$$\begin{aligned}\frac{\partial}{\partial\theta_1}\log p_{\underline{\theta}}(y)\Big|_{\underline{\theta}=\underline{\hat{\theta}}} &= 0\\ &\vdots\\ \frac{\partial}{\partial\theta_m}\log p_{\underline{\theta}}(y)\Big|_{\underline{\theta}=\underline{\hat{\theta}}} &= 0,\end{aligned} \quad \text{(IV.E.1)}$$

which for i.i.d. models becomes

$$\begin{aligned}\sum_{k=1}^{n}\psi_1(y_k;\underline{\hat{\theta}}_n) &= 0\\ &\vdots\\ \sum_{k=1}^{n}\psi_m(y_k;\underline{\hat{\theta}}_n) &= 0,\end{aligned} \quad \text{(IV.E.2)}$$

where $\psi_j(y_k;\underline{\theta})=\partial\log f_{\underline{\theta}}(y_k)/\partial\theta_j$ and where $f_{\underline{\theta}}$ is the marginal density of Y_k.

The information inequality (Proposition IV.C.4) can, within regularity, e extended to the vector case. For example, the Cramér-Rao lower bound ı the variance of unbiased estimates becomes

$$\mathrm{Cov}_{\underline{\theta}}(\hat{\underline{\theta}}) \geq \mathbf{I}_{\underline{\theta}}^{-1}, \tag{IV.E.3}$$

/here $\mathrm{Cov}_{\underline{\theta}}(\hat{\underline{\theta}}) \triangleq E_{\underline{\theta}}\{(\hat{\underline{\theta}}-\underline{\theta})(\hat{\underline{\theta}}-\underline{\theta})^T\}$, and $\mathbf{I}_{\underline{\theta}}$ is the $m \times m$ *Fisher information natrix* with $j-l$th element

$$(\mathbf{I}_{\underline{\theta}})_{j,l} = E_{\underline{\theta}}\left(\left[\frac{\partial}{\partial\theta_j}\log p_{\underline{\theta}}(Y)\right]\left[\frac{\partial}{\partial\theta_l}\log p_{\underline{\theta}}(Y)\right]\right). \tag{IV.E.4}$$

Note that $\mathbf{I}_{\underline{\theta}}$ is the covariance matrix of the zero-mean vector

$$\left(\frac{\partial}{\partial\theta_1}\log p_{\underline{\theta}}(Y), \frac{\partial}{\partial\theta_2}\log p_{\underline{\theta}}(Y), \ldots, \frac{\partial}{\partial\theta_m}\log p_{\underline{\theta}}(Y)\right)^T,$$

and so it is at least nonnegative definite. Equation (IV.E.3) assumes that it is positive definite. The inequality $\mathbf{A} \geq \mathbf{B}$ for matrices means that $(\mathbf{A}-\mathbf{B})$ is nonnegative definite. For the i.i.d. case, (IV.E.3) becomes

$$\mathrm{Cov}_{\underline{\theta}}(\hat{\underline{\theta}}) \geq \frac{1}{n}\mathbf{i}_{\underline{\theta}}^{-1} \tag{IV.E.5}$$

where

$$(\mathbf{i}_{\underline{\theta}})_{j,l} = E_{\underline{\theta}}\{\psi_j(Y_1;\underline{\theta})\psi_l(Y_1;\underline{\theta})\}. \tag{IV.E.6}$$

Within conditions similar to those of Proposition IV.D.1, solutions to the likelihood equation are consistent, i.e.,

$$\| \hat{\underline{\theta}}_n - \underline{\theta} \| \triangleq \left[\frac{1}{m}\sum_{j=1}^{m}[(\hat{\underline{\theta}}_n)_j - \theta_j)]^2\right]^{1/2} \to 0 \quad (i.p.); \tag{IV.E.7}$$

and within conditions similar to those of Proposition IV.D.2,

$$\sqrt{n}(\hat{\underline{\theta}}_n - \underline{\theta}) \to \mathcal{N}(\underline{0}, \mathbf{i}_{\underline{\theta}}^{-1}) \tag{IV.E.8}$$

n distribution. Thus the vector parameter case is very similar to the scalar ne.

Details of this and other aspects of the behavior of MLEs for i.i.d. models an be found in the book by Lehmann (1983).

V.E.2 Estimation of Signal Parameters

The asymptotic properties of MLEs can also be extended to some time arying problems. Of particular interest is the situation in which we have eal-valued observations of the form

$$Y_k = s_k(\theta) + N_k, \quad k = 1, \ldots, n, \tag{IV.E.9}$$

where $\{s_k(\theta)\}_{k=1}^n$ is a signal sequence that is a known function of the unknown parameter θ, and where $\{N_k\}_{k=1}^n$ is an i.i.d. noise sequence with marginal probability density f. We assume for simplicity that θ is a scalar parameter lying in an interval Λ.

The maximum-likelihood estimate of θ in (IV.E.9) solves the equation

$$\hat{\theta}_n = \arg\max_{\theta\in\Lambda}\left[\sum_{k=1}^{n}\log f[Y_k - s_k(\theta)]\right],$$

or equivalently,

$$\hat{\theta}_n = \arg\min_{\theta\in\Lambda}\left[-\sum_{k=1}^{n}\log f[Y_k - s_k(\theta)]\right], \tag{IV.E.10}$$

and the likelihood equation is thus

$$\sum_{k=1}^{n} s_k'(\hat{\theta}_n)\psi[Y_k - s_k(\hat{\theta}_n)] = 0, \tag{IV.E.11}$$

where $\psi \triangleq = -f'/f, f'(x) \triangleq df(x)/dx$, and $s_k'(\theta) \triangleq \partial s_k(\theta)/\partial\theta$. For example, when f is a $\mathcal{N}(0,\sigma^2)$ density, (IV.E.10) and (IV.E.11) are equivalent to

$$\hat{\theta}_n = \arg\left[\min_{\theta\in\Lambda}\sum_{k=1}^{n}[Y_k - s_k(\theta)]^2\right] \tag{IV.E.12}$$

and

$$\sum_{k=1}^{n} s_k'(\hat{\theta}_n)[Y_k - s_k(\hat{\theta}_n)] = 0, \tag{IV.E.13}$$

respectively. The particular estimator (IV.E.12) is sometimes known as the *least-squares estimate* of θ, since it chooses that value of θ for which $\{s_k(\theta)\}_{k=1}^n$ is the least-squares fit to the data. That is, it chooses θ to minimize the sum of the squared errors between the data and the signal that arises from that choice of θ. Least squares is a classical estimation technique and is used frequently in models such as (IV.E.9) even when the errors cannot be assumed to be Gaussian.

Solutions to the likelihood equation (IV.E.11) can have asymptotic properties similar to those for MLEs in i.i.d. models. However, the time variation of the signal adds different considerations to the asymptotic analysis. For example, if the signal becomes identically zero (or otherwise independent of θ) after some finite number of samples, it would be unrealistic to expect consistency in this model. To illustrate the types of conditions needed on the signal for the solutions to the likelihood equation (IV.E.11) to enjoy the properties of their i.i.d. counterparts, we will analyze the particular case of

the least squares estimate (IV.E.13). Similar results will hold for the general case (IV.E.11) within sufficient regularity on ψ.

The equation (IV.E.13) satisfied by the least-squares estimate can be written using the observation model (IV.E.9) as

$$\sum_{k=1}^{n} s_k'(\hat{\theta}_n)N_k + \sum_{k=1}^{n} s_k'(\hat{\theta}_n)[s_k(\theta) - s_k(\hat{\theta}_n)] = 0. \tag{IV.E.14}$$

To analyze the behavior of $\hat{\theta}_n$, let us consider for each $\theta' \in \Lambda$ the sequence of random variables

$$J_n(\theta;\theta') \triangleq \sum_{k=1}^{n} s_k'(\theta')N_k + \sum_{k=1}^{n} s_k'(\theta')[s_k(\theta) - s_k(\theta')]. \tag{IV.E.15}$$

Note that in the absence of noise ($N_k \equiv 0$), $\hat{\theta}_n = \theta$ is a solution to the likelihood equation (IV.E.14). However, unless $\theta' = \theta$ is the only root of

$$K_n(\theta;\theta') \triangleq \sum_{k=1}^{n} s_k'(\theta')[s_k(\theta) - s_k(\theta')], \tag{IV.E.16}$$

Equation (IV.E.14) may not lead to a perfect estimate even in the noiseless case. Thus for consistency in (IV.E.14), we would expect that we need the noise term, $\sum_{k=1}^{n} s_k'(\theta')N_k$, in (IV.E.15) to be asymptotically negligible relative to the term, $K_n(\theta;\theta')$, and for the latter term to have a unique root asymptotically. Since the solution to (IV.E.14) is unchanged if we divide each side by some $d_n > 0$, we can modify the statements above to apply to the corresponding terms in $J_n(\theta;\theta')/d_n$; i.e., if we can find a sequence $\{d_n\}_{n=1}^{\infty}$ such that $\sum_{k=1}^{n} s_k'(\theta')N_k/d_n$ is asymptotically negligible and $K_n(\theta;\theta')/d_n$ has a unique root asymptotically, then we can expect the roots of (IV.E.14) to be consistent by analogy with what happens in the i.i.d. case.

Note that, on assuming $\mathcal{N}(0,\sigma^2)$ noise, we have

$$\frac{1}{d_n}J_n(\theta;\theta') \sim N\left(\frac{1}{d_n}K_n(\theta;\theta'), \frac{\sigma^2}{d_n^2}\sum_{k=1}^{n}[s_k'(\theta')]^2\right). \tag{IV.E.17}$$

It is easily seen from this that for given $\theta, \theta' \in \Lambda$, $J_n(\theta;\theta')/d_n$ converges in probability to a constant if and only if

$$\lim_{n\to\infty} \frac{1}{d_n^2}\sum_{k=1}^{n}[s_k'(\theta')]^2 = 0 \tag{IV.E.18}$$

and

$$\lim_{n\to\infty} \frac{1}{d_n}K_n(\theta;\theta') \text{ exists.} \tag{IV.E.19}$$

From this result we can prove the following proposition, which is analogous to Proposition IV.D.1.

Proposition IV.E.1: Consistency of Least Squares

Suppose that we have the model of (IV.E.9) with $\mathcal{N}(0,\sigma^2)$ noise and that there exists a sequence of scalars $\{d_n\}_{n=1}^{\infty}$ such that (IV.E.18) and (IV.E.19) hold for all $\theta' \in \Lambda$. Suppose further that $s_k(\theta')$, $s_k'(\theta')$, and

$$J(\theta;\theta') \triangleq \lim_{n\to\infty} \frac{1}{d_n}K_n(\theta;\theta') \tag{IV.E.20}$$

are all continuous functions of θ', and that $J(\theta;\theta')$ has a unique root at $\theta' = \theta$. Then, with probability tending to 1, the likelihood equation (IV.E.13) has a sequence of roots converging in probability to θ. In particular, if (IV.E.13) has a unique root $\hat{\theta}_n$ for each n, then $\hat{\theta}_n \to \theta$ (i.p.).

The proof of this result is virtually identical to that of Proposition IV.D.1, and is left as an exercise. As an example, consider the problem of signal-amplitude estimation (see Example IV.D.2), in which

$$s_k(\theta) = \theta s_k, \quad k = 1, 2, \ldots, n, \tag{IV.E.21}$$

for a known sequence $\{s_k\}_{k=1}^{\infty}$. In this case, we have $s_k'(\theta) = s_k$, so that $\sum_{k=1}^{n}[s_k'(\theta)]^2 = \sum_{k=1}^{n} s_k^2$ and $K_n(\theta;\theta') = (\theta-\theta')\sum_{k=1}^{n} s_k^2$. Thus a sufficient condition for consistency following from the proposition is the existence of a divergent sequence $\{d_n\}_{n=1}^{\infty}$ such that

$$0 < \lim_{n\to\infty} \frac{1}{d_n}\sum_{k=1}^{n} s_k^2 < \infty. \tag{IV.E.22}$$

Asymptotic normality can also be assured for the least-squares estimate in (IV.E.9) under regularity conditions on the signal sequence. Note that if $s_k(\theta)$ has third derivatives, the likelihood equation can be expanded in a Taylor series about θ, to give

$$\begin{aligned}\sum_{k=1}^{n} & s_k'(\theta)[Y_k - s_k(\theta)] \\ &+ (\hat{\theta}_n - \theta)\sum_{k=1}^{n}[s_k''(\theta)[Y_k - s_k(\theta)] - [s_k'(\theta)]^2] \\ &+ \frac{1}{2}(\hat{\theta}_n - \theta)^2\sum_{k=1}^{n}[s_k'''(\bar{\theta}_n)[Y_k - s_k(\bar{\theta}_k)] - 3s_k''(\bar{\theta}_n)s_k'(\bar{\theta}_n)] = 0.\end{aligned} \tag{IV.E.23}$$

with $\bar{\theta}_n$ between θ and $\hat{\theta}_n$. On rearranging we have

$$\hat{\theta}_n - \theta = \frac{-\sum_{k=1}^{n} s_k'(\theta) N_k}{\sum_{k=1}^{n} s_k''(\theta) N_k - \sum_{k=1}^{n} [s_k'(\theta)]^2 + \frac{1}{2}(\hat{\theta}_n - \theta) \sum_{k=1}^{n} Z_k(\bar{\theta}_n)} \tag{IV.E.24}$$

where

$$Z_k(\theta') \triangleq [s_k'''(\theta')[N_k + s_k(\theta) - s_k(\theta')] - 3 s_k''(\theta') s_k'(\theta')].$$

From this expression for the error, the following result can be proven.

Proposition IV.E.2: Asymptotic Normality of Least Squares

Suppose that we have the model of (IV.E.9) with $\mathcal{N}(0, \sigma^2)$ noise, and $\{\hat{\theta}_n\}_{n=1}^{\infty}$ is a consistent sequence of least-squares estimates of θ. Suppose further that the following regularity conditions hold:

(1) There exists a function M such that $|Z_k(\theta')| \leq M(N_k)$ uniformly in θ', and $E_\theta\{M(N_k)\} < \infty$. [The existence of the relevant derivatives of $s_k(\theta)$ is also assumed.]

(2) $\lim_{n\to\infty} (1/n) \sum_{k=1}^{n} [s_k'(\theta)]^2 > 0$.

(3) $\lim_{n\to\infty} \sum_{k=1}^{n} [s_k''(\theta)]^2 / [\sum_{k=1}^{n} [s_k'(\theta)]^2]^2 = 0$.

Then,

$$\left(\sum_{k=1}^{n} [s_k'(\theta)]^2 \right)^{1/2} (\hat{\theta}_n - \theta) \to \mathcal{N}(0, \sigma^2) \tag{IV.E.25}$$

in distribution.

The proof of this result is similar to that for the analogous i.i.d. case and is left as an exercise. Note that Fisher's information is given here by

$$I_\theta = \sum_{k=1}^{n} [s_k'(\theta)]^2 / \sigma^2. \tag{IV.E.26}$$

Thus in the same sense as in the i.i.d. case, the least-squares estimate is asymptotically efficient for (IV.E.9) with $\mathcal{N}(0, \sigma^2)$ errors.

The signal-amplitude estimation problem, $s_k(\theta) = \theta s_k$, again provides a straightforward example. In this case, the differentiability conditions are trivial, $Z_k(\theta') \equiv 0$, and $s_k''(\theta) = 0$; thus the only condition needed for asymptotic normality is that $\lim_{n\infty} \sum_{k=1}^{n} s_k^2 / n > 0$. Recall, however, that

the desirable properties of the MLE in this particular case follow by direct analysis (even for finite n), as was seen in Example IV.D.2.

A less obvious example is given by the following.

Example IV.E.1: Identification of a First-Order Linear System

An important class of applications of parameter estimation problems falls within the context of *system identification*, in which we wish to infer the structure of some input/output system by putting in an input and observing the output. One of the simplest possible identification problems is that of identifying a stable first-order time-invariant linear system. This type of system can be described by the signal model

$$s_k(\theta) = \theta s_{k-1}(\theta) + u_k, \quad k = 1, 2, \ldots, n, \tag{IV.E.27}$$

where $|\theta| < 1$ and $\{u_k\}_{k=1}^{n}$ is the known input sequence. Note that θ here is the coefficient of the homogeneous equation $s_k(\theta) = \theta s_{k-1}(\theta)$, and thus this parameter completely determines the system once we have made the assumptions of linearity, time invariance, and unit order. The observation of the system output is usually corrupted by measurement noise, so assuming that this noise is i.i.d., the estimation of θ is a problem in the form of IV.E.9. We consider the case of $\mathcal{N}(0, \sigma^2)$ errors and the least-squares estimate of θ.

Assume that the system (IV.E.27) is initially at rest $[s_0(\theta) = 0]$, in which case the solution to (IV.E.27) is given by

$$s_k(\theta) = \sum_{l=1}^{k} \theta^{k-1} u_l. \tag{IV.E.28}$$

Whether or not θ can be identified (as $n \to \infty$) depends on the input sequence $\{u_k\}_{k=1}^{n}$. Consider, for example, a constant input signal $u_k = 1$ for all $k \geq 1$. The output is then

$$s_k(\theta) = \sum_{l=1}^{k} \theta^{i-l} = \sum_{m=0}^{k-1} \theta^m = \frac{1 - \theta^k}{1 - \theta},$$

and

$$s_k'(\theta) = \frac{(1 - \theta^k) - k\theta^{k-1}(1 - \theta)}{(1 - \theta)^2}.$$

This implies that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} [s_k'(\theta)]^2 = \frac{(2 - \theta)^2}{(1 - \theta)^4} \tag{IV.E.29}$$

and

$$\lim_{n\to\infty} \left[\frac{1}{n} \sum_{k=1}^{n} [s_k'(\theta')[s_k(\theta) - s_k(\theta')] \right] = \frac{(2 - \theta')(\theta - \theta')}{(1 - \theta')^2 (1 - \theta)}. \tag{IV.E.30}$$

Since (IV.E.30) has a unique root at $\theta' = \theta$ and the relevant quantities are continuous for $|\theta'| < 1$, (IV.E.29) and (IV.E.30) imply that the hypothesis of Proposition IV.E.1 is satisfied with $d_n = n$. Thus we have a consistent sequence of roots to the likelihood equation. [In fact, since $J(\theta; \theta')$ is bounded away from zero off a neighborhood of $\theta' = \theta$, it can be shown that any sequence of roots is consistent.]

It is not difficult to see why the consistent estimation of θ is possible in this case. Note that the asymptotic value of $s_k(\theta)$ is $1/(1-\theta)$. Thus the system achieves a unique steady-state value for each value of parameter θ. From this we would expect to be able to determine the parameter value perfectly by observing the noisy output for $k = 1, 2, \ldots, \infty$, since the noise can be averaged out in infinite time. On the other hand, suppose that we use an input with only finite duration. Then, since the system is stable, the steady-state output of the system is zero for every parameter value. It is easy to see that the hypothesis of Proposition IV.E.1 fails to hold in this case. If the measurement noise were not present, it might be possible to determine the parameter perfectly in this case from the transient behavior; however, the presence of the noise makes it necessary that the parameter be identifiable in the steady state as well. The quality of an input that produces this effect is sometimes known as *persistence of excitation.* (A related quality that is sometimes required of an input in linear-system identification problems is *sufficient richness.* Basically, this property means that the frequency content of the input signal is sufficiently rich to excite all oscillatory modes of the system.)

For the constant input signal, Proposition IV.E.2 cannot be applied directly to this model with $\Lambda = (-1, 1)$ because $Z_k(\theta')$ cannot be uniformly bounded on this set. However, if we assume that θ is bounded away from unity [i.e., if we take $\Lambda = (-1, \theta_u)$ with $\theta_u < 1$], then the regularity conditions of Proposition IV.D.4 do hold, and asymptotic normality and efficiency of the consistent roots of the likelihood equation follow. Note that the asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \theta)$ in this case is $\sigma^2(1-\theta)^4/(2-\theta)^2$.

Some additional aspects of maximum-likelihood and least-squares estimates of signal parameters are discussed below and in Chapter VII. However, before leaving this subject for now, we note that the properties of least squares summarized in Propositions IV.E.1 and IV.E.2 hold more generally. In particular, we have the following.

Proposition IV.E.3: Consistency and Asymptotic Normality of Least-Squares with Non-Gaussian Noise

Propositions IV.E.1 and IV.E.2 remain valid if the assumption $N_k \sim \mathcal{N}(0, \sigma^2)$ is replaced by the assumption $E\{N_k\} = 0$ and $E\{N_k^2\} = \sigma^2 < \infty$

Note, however, that this result does not imply that least squares is asymptotically efficient when the noise is not Gaussian, since Fisher's information is no longer given by (IV.E.26) in the non-Gaussian case.

IV.E.3 Robust Estimation of Signal Parameters

Consider again the model of (IV.E.9), in which we have noted that MLEs are asymptotically optimum in the sense of minimum asymptotic variance.

As we discussed in Section III.E, statistical models such as this are only approximately valid in practice, and an important question arising in such situations is whether or not procedures designed for a particular model are *robust*; i.e., whether their performance is insensitive to small changes in the model.

Consider, for example, a nominal model in which the noise samples have the $\mathcal{N}(0,1)$ distribution. Then, within regularity, and assuming that $e_\theta \triangleq \lim_{n\to\infty} \sum_{k=1}^n [s_k'(\theta)]^2/n$ exists and is positive, the least-squares estimate is asymptotically $\mathcal{N}(\theta, 1/ne_\theta)$. Suppose, however, that the actual statistical behavior of the noise is described by a pdf that is only approximately $\mathcal{N}(0,1)$. For example, suppose that the noise density f is of the form

$$f(x) = (1-\epsilon)\frac{1}{\sqrt{2\pi}} e^{-x^2/2} + \epsilon h(x), \quad x \in \mathbb{R}, \tag{IV.E.31}$$

where $h(x)$ is an arbitrary density, symmetric about zero, and with variance

$$\sigma_h^2 \triangleq \int_{-\infty}^{\infty} x^2 h(x)\,dx$$

finite but not bounded. Then, by Proposition IV.E.3, the least-squares estimate will have asymptotic variance

$$v_h^2 \sim \frac{(1-\epsilon) + \epsilon\sigma_h^2}{ne_\theta}. \tag{IV.E.32}$$

Note that v_h^2 can be arbitrarily large for any $\epsilon > 0$ since σ_h^2 is not bounded. In particular, the worst-case asymptotic variance over the class of densities (IV.E.31) is

$$\sup_h [(1-\epsilon) + \epsilon\sigma_h^2] = \infty \tag{IV.E.33}$$

for any $\epsilon > 0$.

This points to a lack of robustness of the least-squares estimate for situations in which a small fraction of the noise samples may come from a high variance distribution. (This may happen, for example, in radar measurements, in which very high-variance impulsive interference may be present in a small fraction ϵ of the measurements. Observations that are improbably large for a given nominal model are sometimes termed *outliers.*) As

in the signal detection problems treated in Section III.E, an alternative to asymptotic variance at a nominal model is needed as a design criterion for such situations.

Suppose that the noise density f in (IV.E.9) is an even symmetric function. Consider estimates of θ of the form

$$\sum_{k=1}^{n} s_k'(\hat{\theta}_n)\psi[Y_k - s_k(\hat{\theta}_n)] = 0, \tag{IV.E.34}$$

where ψ is a general odd-symmetric function. With $\psi(x) = x$, (IV.E.34) gives the least-squares estimate, and with $\psi(x) = -f'(x)/f(x)$, (IV.E.34) gives the MLE. Estimates of this form are known as *M-estimates*. Assuming that $0 < e_\theta < \infty$ and within regularity on ψ, f, and $\{s_k(\theta)\}_{k=1}^{\infty}$, it can be shown, using the techniques developed above, that M-estimates are consistent and asymptotically $N[\theta, V(\psi, f)/ne_\theta]$, where

$$V(\psi, f) \triangleq \frac{\int \psi^2 f}{(\int \psi' f)^2} \tag{IV.E.35}$$

with $\psi'(x) = d\psi(x)/dx$.

In view of these properties, one possible way of designing a robust estimator for an uncertainty class $\mathcal{F}$ of noise densities is to seek a function ψ that minimizes the worst case M-estimate variance, $\sup_{f\in\mathcal{F}} V(\psi, f)$. That is, one possible design method is to restrict attention to M-estimates and solve

$$\min_{\psi} \sup_{f\in\mathcal{F}} V(\psi; f). \tag{IV.E.36}$$

The problem (IV.E.36) has been studied by Huber (1981) for general sets $\mathcal{F}$. Within appropriate conditions, its solution is basically as follows.

Consider the functional

$$I(f) \triangleq \int (f')^2/f, \tag{IV.E.37}$$

and let f_L be a density in $\mathcal{F}$ that minimizes $I(f)$ over $\mathcal{F}$; i.e.,

$$I(f_L) = \min_{f\in\mathcal{F}} I(f). \tag{IV.E.38}$$

Then the M-estimate with ψ-function $\psi_R(x) = -f_L'(x)/f_L(x)$ solves (IV.E.36). Note that for any f,

$$V(\psi, f)\,|_{\psi=-f'/f} = 1/I(f), \tag{IV.E.39}$$

so that $[ne_\theta I(f)]^{-1}$ is the asymptotic variance of the MLE in our model with given f. [Fisher's information here is $ne_\theta I(f)$.] Thus f_L is the member of $\mathcal{F}$ whose corresponding optimum estimate (the MLE) has the worst optimum performance. For this reason f_L can be considered a *least-favorable density*, and the robust M-estimate is the best estimate for this least-favorable model.

The problem $\min_{f\in\mathcal{F}} I(f)$ has been solved for a number of uncertainty models $\mathcal{F}$ [see Huber (1981)]. For example, for the ϵ-contaminated $\mathcal{N}(0,1)$ model of (IV.E.31), the least favorable density is given by

$$f_L(x) = \begin{cases} (1-\epsilon)\frac{1}{\sqrt{2\pi}} e^{-x^2/2} & \text{if } |x| \le k' \\ (1-\epsilon) e^{-k'(|x|-k')} \frac{1}{\sqrt{2\pi}} e^{(-k')^2/2} & \text{if } |x| > k', \end{cases} \tag{IV.E.40}$$

where k' is a constant given by the solution to

$$(1-\epsilon)^{-1} = 2\Phi(k') - 1 + \frac{1}{k'}\left(\frac{2}{\pi}\right)^{1/2} e^{-(k')^2/2}. \tag{IV.E.41}$$

The corresponding robust ψ function is

$$\psi_k(x) = \begin{cases} x & \text{if } |x| \le k' \\ k' \text{ sgn } (x) & \text{if } |x| > k'. \end{cases} \tag{IV.E.42}$$

Thus, as in the analogous hypothesis testing problem, robustness is brought about by limiting the effects of outliers.

For further discussion of this and other approaches to robust estimation, the reader is referred to the survey article by Kassam and Poor (1985) and the books by Huber (1981) and Hampel, *et al.* (1986).

IV.E.4 Recursive Parameter Estimation

We see from the preceding discussions that maximum-likelihood estimates often have nice properties, particularly when the sample size is large. However, they sometimes have the disadvantages of being cumbersome to compute. For example, with n i.i.d. samples drawn from the density f_θ, computation of the MLE requires the maximization of the function $\sum_{k=1}^{n} \log f_\theta(y_k)$. Unless the maximizing θ can be found as a closed-form function of $\underline{y}$, an iterative technique must be used to find $\hat{\theta}_{ML}(\underline{y})$. This requires the storage and simultaneous manipulation of all n samples (unless a lower-dimensional sufficient statistic is available), a task that is undesirable if n is very large. It is thus sometimes desirable to consider alternatives to maximum likelihood that can be implemented in a recursive or sequential manner so that the contribution of each sample to the estimate is computed as the sample is taken.

One such estimation technique is suggested by the MLE. In particular, consider a consistent sequence $\{\hat{\theta}_n\}_{n=1}^{\infty}$ solving the likelihood equation

$$\sum_{k=1}^{n} \psi(Y_k; \hat{\theta}_n) = 0 \tag{IV.E.43}$$

with $\psi(Y_k;\theta) = \partial \log f_\theta(Y_k)/\partial\theta$, as before. Since $\{\hat{\theta}_n\}_{n=1}^{\infty}$ is consistent, the difference, $\hat{\theta}_n - \hat{\theta}_{n-1}$, converges to zero as $n \to \infty$. Thus (IV.E.43) can be approximated by expanding about $\hat{\theta}_{n-1}$ to give

$$\sum_{k=1}^{n} \psi(Y_k;\hat{\theta}_{n-1}) + (\hat{\theta}_n - \hat{\theta}_{n-1}) \sum_{k=1}^{n} \psi'(Y_k;\hat{\theta}_{n-1}) \sim 0, \qquad \text{(IV.E.44)}$$

with $\psi'(Y_k;\theta) = \partial\psi(Y_k;\theta)/\partial\theta$. Rearranging (IV.E.44) gives

$$\hat{\theta}_n \sim \hat{\theta}_{n-1} - \frac{\sum_{k=1}^{n} \psi(Y_k;\hat{\theta}_{n-1})}{\sum_{k=1}^{n} \psi'(Y_k;\hat{\theta}_{n-1})}. \qquad \text{(IV.E.45)}$$

Since $\hat{\theta}_{n-1}$ solves $\sum_{k=1}^{n-1} \psi(Y_k;\hat{\theta}_{n-1}) = 0$, the numerator sum on the right side of (IV.E.45) has only one term, $\psi(Y_n;\hat{\theta}_{n-1})$. Let us write the denominator sum as

$$n\left[\frac{1}{n}\sum_{k=1}^{n} \psi'(Y_k;\hat{\theta}_{n-1})\right]. \qquad \text{(IV.E.46)}$$

Now, the weak law of large numbers implies that

$$-\frac{1}{n}\sum_{k=1}^{n} \psi'(Y_k;\theta) \to i_\theta \quad (i.p.),$$

where $i_\theta = -E_\theta\{\psi'(Y_k;\theta)\} = E\{\psi^2(Y_k;\theta)\}$ is Fisher's information per sample. Since $\hat{\theta}_{n-1} \to \theta$, we can approximate

$$\frac{1}{n}\sum_{k=1}^{n} \psi'(Y_k;\hat{\theta}_{n-1}) \sim i_{\hat{\theta}_{n-1}}. \qquad \text{(IV.E.47)}$$

On combining (IV.E.45) and (IV.E.47) we have that, asymptotically, a consistent sequence of solutions to the likelihood equation will satisfy

$$\hat{\theta}_n \sim \hat{\theta}_{n-1} + \frac{\psi(Y_n;\hat{\theta}_{n-1})}{n i_{\hat{\theta}_{n-1}}}. \qquad \text{(IV.E.48)}$$

This is an asymptotic recursive equation for $\hat{\theta}_n$, since $\hat{\theta}_n$ is computed from $\hat{\theta}_{n-1}$ and Y_n only.

It turns out that the (nonasymptotic) recursion

$$\hat{\theta}_n = \hat{\theta}_{n-1} + \frac{\psi(Y_n;\hat{\theta}_{n-1})}{n i_{\hat{\theta}_{n-1}}}, \quad n = 1, \ldots, \qquad \text{(IV.E.49)}$$

(with $\hat{\theta}_0$ arbitrary) suggested by (IV.E.48) has the same desirable asymptotic properties (i.e., consistency and efficiency) as the MLE within regularity on the model. This recursion is an example of a more general class

of recursive parameter estimation algorithm known as *stochastic approximation* algorithms. Because of their recursive nature, such algorithms are of considerable interest in applications in which on-line or real-time parameter estimation is necessary. In modified form they are also useful in real-time tracking of slowly varying parameters. The reader interested in further aspects of such algorithms is referred to the book by Nevel'son and Has'minskii (1973). Similar recursive modifications of the MLE and least-squares estimates for time-varying problems such as (IV.E.9) have also been developed. The reader is referred to Ljung and Soderstrom (1982) and Goodwin and Sin (1984) for the development of these ideas.

IV.F Exercises

1. Suppose Θ is a random parameter and that, given $\Theta = \theta$, the real observation Y has density

$$p_\theta(y) = (\theta/2)e^{-\theta|y|}, \quad y \in \mathbb{R}.$$

Suppose further that Θ has prior density

$$w(\theta) = \begin{cases} 1/\theta, & 1 \le \theta \le e \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the MAP estimate of Θ based on Y.

(b) Find the MMSE estimate of Θ based on Y.

2. Suppose we have a real observation Y given by

$$Y = N + \Theta S$$

where $N \sim \mathcal{N}(0,1)$, $P(S=1) = P(S=-1) = 1/2$, and Θ has pdf

$$w(\theta) = \begin{cases} Ke^{\theta^2/2}, & 0 \le \theta \le 1 \\ 0, & \text{otherwise} \end{cases}$$

where $K = [\int_0^1 e^{\theta^2/2} d\theta]^{-1}$. Assume that N, Θ, and S are independent.

(a) Find the MMSE estimate of Θ given $Y = y$.

(b) Find the MAP estimate of Θ given $Y = y$.

3. Suppose Θ is a random parameter with prior density

$$w(\theta) = \begin{cases} \alpha e^{-\alpha\theta}, & \theta \ge 0 \\ 0, & \theta < 0 \end{cases}$$

where $\alpha > 0$ is known. Suppose our observation Y is a Poisson random variable with rate Θ; i.e., that

$$p_\theta(y) \equiv P(Y = y|\Theta = \theta) = \frac{\theta^y e^{-\theta}}{y!}, \quad y = 0, 1, 2, \ldots.$$

Find the MMSE and MAP estimates of Θ based on Y. How would you find the MMAE estimate?

4. Suppose we have a single observation y of a random variable Y given by

$$Y = N + \Theta$$

where N is a Gaussian random variable with mean zero and variance σ^2. The parameter Θ is a random variable, independent of N, with probability mass function

$$w(\theta) = P(\Theta = \theta) = \begin{cases} \frac{1}{2}, & \theta = -1 \\ \frac{1}{2}, & \theta = +1. \end{cases}$$

   (a) Find $\hat{\theta}_{MMSE}$ and $\hat{\theta}_{MAP}$. (You may consider the parameter set Λ to be $\mathbb{R}$.)

   (b) Under what conditions are the two estimates in (a) approximately equal?

5. Suppose Θ is a random parameter with prior density

$$w(\theta) = \begin{cases} e^{-\theta}, & \theta \geq 0 \\ 0, & \theta < 0, \end{cases}$$

and that Y has conditional density

$$p_\theta(y) = \tfrac{1}{2} e^{-|y-\theta|}, \quad -\infty < y < \infty.$$

Find $\hat{\theta}_{MMSE}$ and $\hat{\theta}_{MAP}$.

6. Suppose that N_1 and N_2 are two jointly Gaussian random variables with zero means, unit variances, and correlation coefficient ρ ($|\rho| < 1$). Suppose further that we observe Y_1 and Y_2 given by

$$Y_k = \frac{N_k}{\sqrt{\Theta}}, \quad k = 1, 2,$$

where Θ is a random parameter, independent of N_1 and N_2, with prior density

$$w(\theta) = \begin{cases} 1/\alpha, & \theta \in [0, \alpha] \\ 0, & \theta \notin [0, \alpha] \end{cases}$$

where $\alpha > 0$ is known.

   (a) Find the minimum-mean-squared-error estimate of Θ.

   (b) Find the MAP estimate of Θ.

   (c) Find the minimum-mean-absolute-error estimate of Θ.

7. Suppose Θ is uniformly distributed on the interval (0, 1) and that we observe $Y = N + \Theta$ where N is a random variable, independent of Θ, with density

$$p_N(n) = \begin{cases} e^{-n}, & n \geq 0 \\ 0, & n < 0. \end{cases}$$

Find $\hat{\theta}_{MMSE}, \hat{\theta}_{ABS}$, and $\hat{\theta}_{MAP}$.

8. (a) Consider the observation model of Exercise 7 but with the prior of Exercises 5 (i.e., N and Θ both have the unit exponential distribution). Find the MMSE and MMAE estimates of Θ based on Y.

   (b) Find the minimum mean-squared error for (a).

   (c) Consider now the observation model

$$Y_k = N_k + \Theta, \quad k = 1, \ldots, n,$$

   where $N_1, N_2, \ldots, N_n$, and Θ are i.i.d random variables with the unit exponential distribution. Find the MAP estimate of Θ based on $Y_1, Y_2, \ldots, Y_n$.

9. Repeat Exercise 1 for the situation in which we have a sequence of observations $Y_1, Y_2, \ldots, Y_n$, that are conditionally i.i.d. with the given pdf p_θ given $\Theta = \theta$.

10. Derive Eq. (IV.B.47).

11. Suppose that we observe a sequence

$$Y_k = X_k + N_k, \quad k = 1, \ldots, n$$

where $N_1, \ldots, N_n$ is a sequence of independent Gaussian random variables, each with zero mean and variance σ^2, and $X_1, \ldots, X_n$ are defined by the equations

$$X_0 = \Theta$$

$$X_k = \alpha X_{k-1}, \quad k = 1, \ldots, n$$

where α is known and Θ is a Gaussian random parameter with zero mean and variance q^2.

   (a) Assuming that Θ and N are independent, find the MMSE estimate of Θ based on $Y_1, \ldots, Y_n$.

(b) For each $n = 1, 2, \ldots$, let $\hat{\theta}_n$ denote the MMSE estimate of Θ based on $Y_1, \ldots, Y_n$. Show that $\hat{\theta}_n$ can be computed recursively by

$$\hat{\theta}_n = K_n^{-1}[K_{n-1}\hat{\theta}_{n-1} + \alpha^n y_n], \quad n = 1, 2, \ldots,$$

where $\hat{\theta}_0 = 0$ and the coefficients K_n are defined by

$$K_0 = \sigma^2/q^2 \quad \text{and} \quad K_n = K_{n-1} + \alpha^{2n}, \quad n = 1, 2, \ldots.$$

Draw a block diagram of this implementation.

(c) Find an expression for the mean-squared error

$$e_n = E\{(\hat{\theta}_n - \Theta)^2\}, \quad n = 1, 2, \ldots.$$

What happens when $n \to \infty; q^2 \to \infty; \sigma^2 \to 0; \alpha < 1; \alpha = 1; \alpha > 1$?

12. Suppose θ is a nonrandom parameter satisfying $\theta > 1$. Suppose further that, given $\theta, Y_1, Y_2, \ldots, Y_n$ are i.i.d. observations with each density

$$f_\theta(y) = \begin{cases} (\theta - 1)y^{-\theta}, & y \geq 1 \\ 0, & y \leq 1. \end{cases}$$

Find a sufficient statistic for θ that has a complete family of distributions. Justify your answer.

13. Suppose we toss a coin n independent times and define an observation sequence

$$Y_k = \begin{cases} 1 & \text{if the } k\text{th outcome is heads} \\ 0 & \text{if the } k\text{th outcome is tails} \end{cases}$$

$k = 1, 2, \ldots, n$. Let $\theta = P(Y_k = 1), k = 1, \ldots, n$.

(a) Find an MVUE of θ.

(b) Find the ML estimate of θ. Find its bias and variance.

(c) Compute the Cramér-Rao lower bound and compare with results from (a) and (b).

14. Derive Eq. (IV.D.22).

15. Suppose Y is Poisson. Find the ML estimate of its rate. Compute the bias, variance, and Cramér-Rao lower bound.

16. Suppose θ is a positive (nonrandom) parameter. Suppose further that we have a sequence of observations $Y_1, \ldots, Y_n$ where, given $\theta, Y_1, \ldots, Y_n$ are i.i.d. each with pdf

$$f_\theta(y) = \begin{cases} \frac{(y)^M e^{-y/2\theta}}{(2\theta)^{M+1} M!}, & y \geq 0 \\ 0, & y < 0 \end{cases}$$

where M is a known positive integer.

(a) Find the ML estimate of θ.

(b) Compute the bias and variance of the estimate from part (a).

(c) Compute the Cramér-Rao lower bound on the variance of unbiased estimates of θ.

(d) Is the ML estimate consistent? Is it efficient?

17. Suppose we observe two jointly Gaussian random variables Y_1 and Y_2, each of which has zero mean and unit variance. We want to estimate the correlation coefficient $\rho = E\{Y_1 Y_2\}$.

(a) Find the equation for the maximum-likelihood estimate of ρ based on observation of (Y_1, Y_2).

(b) Compute the Cramér-Rao lower bound for unbiased estimates of ρ.

18. Suppose we observe a sequence $Y_1, Y_2, \ldots, Y_n$ given by

$$Y_k = N_k + \theta s_k, \quad k = 1, \ldots, n$$

where $\underline{N} = (N_1, \ldots, N_n)^T$ is a zero-mean Gaussian random vector with covariance matrix $\mathbf{\Sigma} > 0; s_1, s_2, \ldots, s_n$ is a known signal sequence; and θ is a (real) nonrandom parameter.

(a) Find the maximum-likelihood estimate of the parameter θ.

(b) Compute the bias and variance of your estimate.

(c) Compute the Cramér-Rao lower bound for unbiased estimates of θ and compare with your result from (b).

(d) What can be said about the consistency of $\hat{\theta}_{ML}$ as $n \to \infty$? Suppose, for example, that there are positive constants a and b such that

$$\frac{1}{n}\sum_{k=1}^{n} s_k^2 > a \text{ for all } n$$

and

$$\lambda_{\min}(\mathbf{\Sigma}^{-1}) > b \text{ for all } n$$

where $\lambda_{\min}(\mathbf{\Sigma}^{-1})$ denotes the minimum eigenvalue of the matrix $\mathbf{\Sigma}^{-1}$.

19. Suppose θ is a positive nonrandom parameter and that we have a sequence $Y_1, \ldots, Y_n$ of observations given by

$$Y_k = \theta^{1/2} N_k, \quad k = 1, 2, \ldots, n$$

where $\underline{N} = (N_1, \ldots, N_n)^T$ is a Gaussian random vector with zero mean and covariance matrix $\mathbf{\Sigma}$. Assume that $\mathbf{\Sigma}$ is positive definite.

(a) Find the maximum-likelihood estimate of θ based on $Y_1, \ldots, Y_n$.

(b) Show that the maximum-likelihood estimate is unbiased.

(c) Compute the Cramér-Rao lower bound on the variance of unbiased estimates of θ.

(d) Compute the variance of the maximum-likelihood estimate of θ and compare to the Cramér-Rao lower bound.

20. Consider the observation model

$$Y_k = \theta^{1/2} s_k R_k + N_k, \quad k = 1, 2, \ldots, n$$

where $s_1, s_2, \ldots, s_n$ is a known signal, $N_1, N_2, \ldots, N_n, R_1, R_2, \ldots, R_n$ are i.i.d. $\mathcal{N}(0,1)$ random variables, and $\theta \geq 0$ is an unknown parameter.

(a) Find the likelihood equation for estimating θ from $Y_1, Y_2, \ldots, Y_n$.

(b) Find the Cramér-Rao lower bound on the variance of unbiased estimates of θ.

(c) Suppose $s_1, s_2, \ldots, s_n$ is a sequence of $+1$'s and -1's. Find the MLE of θ explicitly.

(d) Compute the bias and variance of your estimate from (c), and compare the latter with the Cramér-Rao lower bound.

21. Suppose Y_1 and Y_2 are independent Poisson random variables each with parameter λ. Define the parameter θ by

$$\theta = e^{-\lambda}.$$

(a) Show that $Y_1 + Y_2$ is a complete sufficient statistic for θ. [Assume λ ranges over $(0, \infty)$.]

(b) Define an estimate $\hat{\theta}$ by

$$\hat{\theta}(y) = \tfrac{1}{2}[f(y_1) + f(y_2)]$$

where f is defined by

$$f(y) = \begin{cases} 1 & \text{if } y = 0 \\ 0 & \text{if } y \neq 0 \end{cases}$$

Show that $\hat{\theta}$ is an unbiased estimate of θ.

(c) Find an MVUE of θ. (*Hint*: $Y_1 + Y_2$ is Poisson with parameter 2λ.)

(d) Find the maximum-likelihood estimate of θ. Is the MLE unbiased; if so, why; if not, why not?

(e) Compute the Cramér-Rao bound on the variance of unbiased estimates of θ.

22. Suppose $\theta > 0$ is a parameter of interest and that given θ, $Y_1, \ldots, Y_n$ is a set of i.i.d. observations with marginal distribution function

$$F_\theta(y) = [F(y)]^{1/\theta}, \quad -\infty < y < \infty,$$

where F is a known distribution function with pdf f.

(a) Show that

$$\hat{\theta}_{MV}(\underline{y}) = -\frac{1}{n} \sum_{k=1}^{n} \log F(y_k)$$

is an MVUE of θ.

(b) Suppose now that θ is replaced by a random variable Θ drawn at random using the prior density

$$w(\theta) = c^m \exp(-c/\theta)/(\Gamma(m)\theta^{m+1}), \quad \theta > 0,$$

where $c > 0$ and $m > 1$ are constants. Use the fact that $E\{\Theta\} = c/(m-1)$ to show that the MMSE estimator of Θ from $Y_1, \ldots, Y_n$ is

$$\hat{\theta}_{\text{MMSE}}(\underline{y}) = \left(c - \sum_{k=1}^{n} \log F(y_k) \right) / \, (m + n - 1).$$

(c) Compare $\hat{\theta}_{\text{MV}}$ and $\hat{\theta}_{\text{MMSE}}$ with regard to the role of the prior information.

23. Suppose we observe

$$Y_k = A \sin\left(\frac{k\pi}{2} + \Phi \right) + N_k, \quad k = 1, \ldots, n$$

where $\underline{N} \sim \mathcal{N}(\underline{0}, \sigma^2 \mathbf{I})$ and n is even.

(a) Suppose A and Φ are nonrandom with $A \geq 0$ and $\Phi \in [-\pi, \pi]$. Find their ML estimates.

(b) Suppose A and Φ are random and independent with priors

$$w_\Phi(\phi) = \begin{cases} \frac{1}{\pi}, & -\pi \leq \phi \leq \pi \\ 0, & \text{otherwise} \end{cases}$$

$$w_A(a) = \begin{cases} (a/\beta^2) e^{-a^2/2\beta^2} & \text{if } a \geq 0 \\ 0 & \text{if } a < 0 \end{cases}$$

where β is known. Assuming A and Φ are independent of $\underline{N}$, find the MAP estimates of A and Φ.

(c) Under what conditions are the estimates from (a) and (b) approximately equal?

. Suppose that, given $\Theta = \theta, Y_1, \ldots, Y_n$ are i.i.d. real observations with marginal densities

$$f_\theta(y) = \begin{cases} \theta^{-1} e^{-y/\theta}, & y \geq 0 \\ 0, & y < 0. \end{cases}$$

(a) Find the maximum-likelihood estimate of θ based on $Y_1, \ldots, Y_n$. Compute its mean and variance.

(b) Compute the Cramér-Rao lower bound for the variance of unbiased estimates of θ.

(c) Suppose Θ is uniformly distributed on (0, 1]. Find the MAP estimate of Θ

(d) For $n = 3$, find the MMSE estimate of Θ. Assume the same prior as in part (c).

(e) For $n = 2$, find the MMAE estimate of Θ. Assume the same prior as in part (c).

25. Suppose that, given $\Theta = \theta$, the real observation Y has pdf

$$p_\theta(y) = \begin{cases} \frac{6(y^2+\theta y)}{2+3\theta}, & 0 \leq y \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Suppose Θ is uniformly distributed on $[0, 1]$. Find the MMSE estimate and corresponding minimum Bayes risk.

(b) With Θ as in (a), find the MAP estimate and the MMAE estimate of Θ.

(c) Find the maximum-likelihood estimate of θ and compute its bias.

(d) Compute the Cramér-Rao lower bound on the variance of unbiased estimates of θ.

V

Elements of Signal Estimation

V.A Introduction

In Chapter IV we discussed methods for designing estimators for static parameters, that is, for parameters that are not changing with time. In many applications we are interested in the related problem of estimating dynamic or time-varying parameters. In the traditional terminology, a dynamic parameter is usually called a *signal*, so the latter problem is known as *signal estimation* or *tracking*.

Such problems arise in many applications. For example, one function of many radar systems is to track targets as they move through the radar's scanning area. This means that the radar must estimate the position of the target (and perhaps its velocity) at successive times. Since the targets of interest are usually moving and the position measurements are noisy, this is a signal estimation problem. Another application is that of analog communications, in which analog information (e.g., audio or video) is transmitted by modulating the amplitude, frequency, or phase of a sinusoidal carrier. The receiver's function in this situation is to determine the transmitted information with as high a fidelity as possible on the basis of a noisy observation of the received waveform. Again, since the transmitted information is time varying, this problem is one of signal estimation.

The dynamic nature of the parameter in signal estimation problems adds a new dimension to the statistical modeling of these problems. In particular, the dynamic properties of the signal (i.e., how fast and in what manner it can change) must be modeled at least statistically in order to obtain meaningful signal estimation procedures. Also, performance expectations for estimators of dynamic parameters should be different from those for static parameters. In particular, unlike the static case, we cannot expect an estimator of a signal to be perfect as the number of observations becomes infinite because of the time variation in the signal.

In this chapter we discuss the basic ideas behind some of the signal estimation techniques used most often in practice. In Section V.B we discuss *Kalman-Bucy filtering*, which provides a very useful algorithm for estimating signals that are generated by finite-dimensional linear dynamical models. In Section V.C the general problem of estimating signals as lin-