§4.3. Nonrandom Parameter Estimation : General Structure

In the previous section, we considered the problem of estimating a random parameter indexing a class of distributions on the observation space. In this section, we consider the problem in which we have a parameter (indexing the class of observation statistics) that is not modeled as a random variable but is unknown. There is not enough prior information about the parameter to assign a prior probability distribution to it.

Suppose 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 that the parameter θ is real-valued. We donot know anything about the true value of θ other than the fact that it lies in Λ.

The problem is : Given the observation $Y = y$, what is the best estimate of θ?

Similar to before, we seek an estimate $\hat{\theta}(y)$ that minimizes some average performance criterion. In the remaining of this chapter, we exclusively consider the squared-error cost, although some results here apply straightforwardly to other cost assignments as well.

In the absence of a prior on Λ, the only averagings of cost that can be done is with respect to the distribution of Y given θ, i.e., the conditional risk function:

$$R_\theta(\hat{\theta}) \triangleq E_\theta\{(\hat{\theta}(Y)-\theta)^2\}, \quad \theta \in \Lambda.$$

We can not minimize $R_\theta(\hat{\theta})$ uniformly for $\theta \in \Lambda$ since there is only one true θ, say θ_0. The conditional mean-squared error can be mode zero by choosing $\hat{\theta}(y)=\theta_0$ for all observations $y \in \Gamma$. But such an estimato would perform poorly if θ_0 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 paramoter unless the class of estimators is somehow restricted to contain only reasonable estimators [e.g., to exclude estimators such as $\hat{\theta}(y)=\theta_0$ for a nontrue θ_0].

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

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

Such an estimator is termed <u>unbiased</u>.

With this restriction, the conditional mean-squared error becomes the variance of the estimate under

P_θ.

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 problem with a goal of characterizing MVUEs.

* Definition 4.3.1: Sufficiency

Suppose that Δ is an arbitrary set and $\mathcal{D}$ is an event class on Δ. A function $T: (\Gamma, \mathcal{G}) \rightarrow (\Delta, \mathcal{D})$ 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).

We can only learn about θ by viewing the statistical behavior of Y. 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 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 observation as much as possible.

* Definition 4.3.2: Minimal Sufficiency

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

Minimal sufficient statistics do not exist for many estimation problems, and they are often difficult to identify when they do exist.

* Proposition 4.3.1: 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

$$p_\theta(y) = g_\theta[T(y)]\,h(y) \qquad (*)$$

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

Proof: We prove this result only for the case in which Γ is discrete, whose general idea applies to the general case. A proof of the general case can be found in Lehmann L(1986).

Suppose that Γ is discrete and $\{P_\theta ; \theta \in \Lambda\}$ satisfies (*) 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

$$p_\theta(y|t) \triangleq P_\theta(Y=y \,/\, T(Y)=t)$$

$$= \frac{P_\theta(T(Y)=t \,/\, Y=y)\, P_\theta(Y=y)}{P_\theta(T(Y)=t)}.$$

Since $P_\theta(T(Y)=t \,|\, Y=y) = \begin{cases} 1, & \text{if } T(y)=t ; \\ 0, & \text{if } T(y) \neq t, \end{cases}$

and $P_\theta(Y=y) = p_\theta(y)$,

we have

$$p_\theta(y|t) = \begin{cases} p_\theta(y)/P_\theta(T(Y)=t) & \text{if } T(y)=t, \\ 0 & \text{if } T(y) \neq t, \end{cases} \qquad (**)$$

Note

$$P_\theta(T(Y)=t) = \sum_{\{y|T(y)=t\}} p_\theta(y).$$

Thus, from (*),

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

$$= g_\theta(t) \sum_{\{y|T(y)=t\}} h(y)$$

and also, $p_\theta(y) = g_\theta[T(y)]\, h(y) = g_\theta(t)\, h(y).$

Thus, from (**)

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

which does not depend on θ. This proves that T is a sufficient statistic for $\{P_\theta;\ \theta\in\Lambda\}$.

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

$$\begin{aligned} P_\theta(y) &= P_\theta[y/T(y)]\, P_\theta[T(Y)=T(y)] \\ &= P_\theta(Y=y,\ T(Y)=T(y)] \\ &\quad + \underbrace{P_\theta(Y=y,\ T(Y)\neq T(y)]}_{0} \end{aligned}$$

Since T is sufficient for θ, $P_\theta[y/T(y)]$ depends only on y and not on θ. Let

$$h(y) \hat{=} P_\theta(y/T(y))$$

and

$$g_\theta[T(y)] \hat{=} P_\theta[T(y) = T(y)]$$

is a function only of $T(y)$ and θ.

Then we have (*). Thus the proposition is proved.

$\square$

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

Thus, the likelihood ratio $L(y)$ is a sufficient statistic for the binary hypothesis-testing problems. This statistic is always one dimensional regadless of the nature of Γ.

* Proposition 4.3.2: The Rao-Blackwell Theorem

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

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

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

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

with equality if and only if

$$P_\theta(\hat{g}(Y) = \tilde{g}[T(Y)]) = 1.$$

Proof: First of all, $\tilde{g}$ does not depend on θ from 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].

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

where we have used the fact that $E[E\{\xi|z\}] = E(\xi)$ Thus $\tilde{g}$ is unbiased.

To prove the variance inequality, 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

$$E_\theta\{[\tilde{g}[T(Y)]]^2\} \le E_\theta\{[\hat{g}(Y)]^2\}.$$

$$E_\theta\{[\tilde{g}[T(Y)]]^2\} = E_\theta\{[E_\theta\{\hat{g}(Y)|T(Y)\}]^2\}$$

$$\leq E_\theta\{E_\theta\{[\hat{g}(Y)]^2 | T(Y)\}\}$$
$$= E_\theta\{[\hat{g}(Y)]^2\},$$

we the inequality follows from applying Jensen's inequality to

$$[E_\theta\{\hat{g}(Y)|T(Y)\}]^2 \leq E_\theta\{[\hat{g}(Y)]^2 | T(Y)\}.$$

Also, the equality in the Jenson's inequality holds if and only if

$$P_\theta[\hat{g}(Y) = E_\theta\{\hat{g}(Y)|T(Y)\} \mid T(Y)] = 1.$$

Since $\tilde{g}[T(Y)] \triangleq E_\theta\{\hat{g}(Y)|T(Y)\}$, the condition is equivalent to

$$P_\theta[\hat{g}(Y) = \tilde{g}[T(Y)] = 1.$$

This completes the proof. □

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 $T(Y)$.

This theorem also 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 $g^*[T(y)]$ is the only function of $T(y)$ for which $E_\theta\{g^*[T(Y)]\} = g(\theta)$. Let $\hat{g}(y)$

be any unbiased estimator of $g(\theta)$. Then, by the Rao-Blackwell theorem,

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

is unbiased for $g(\theta)$ and it is a function of $T(y)$. so, by the uniqueness of g^*, we must have $g^* = \tilde{g}$. The Rao-Blackwell theorem also asserts that

$$Var_\theta[\tilde{g}(T(Y))] \leq Var_\theta[\hat{g}(\theta)].$$

Since $\hat{g}$ is arbitrary, we have that

$$Var_\theta(g^*[T(Y)]) \leq Var_\theta(\hat{g}(Y))$$

for any unbiased estimate of $g(\theta)$; i.e., $g^*[T(y)]$ is an MVUE of $g(\theta)$.

$\Rightarrow$ An MVUE of $g(\theta)$ can be constructed if we can find a sufficient statistic T with such a unique estimate $g^*[T(y)]$.

* Definition 4.3.3. 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 similar to that of a set of vectors in $\mathbb{R}^n$." To see this, let us only consider the case when 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

$$\underline{f} = [f(\gamma_1), \ldots, f(\gamma_n)]^T$$

$$\underline{p}_\theta = [\, p_\theta(\gamma_1), \ldots, p_\theta(\gamma_n)\,]^T.$$

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 zeroes. That is, $\{p_\theta ; \theta \in \Lambda\}$ is complete if $\underline{0}$ is the only vector that is orthogonal to all the vectors $\{\underline{p}_\theta ; \theta \in \Lambda\}$, which is the completeness of vectors $\{\underline{p}_\theta ; \theta \in \Lambda\}$.

* Example 2: Completeness of the Binomial Distribution

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

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

For any function f on Γ we have

$$E_\theta\{f(Y)\} = \sum_{y=0}^{n} \frac{n!}{y!\,(n-y)!} f(y)\, \theta^y (1-\theta)^{n-y}$$

$$= (1-\theta)^n \sum_{y=0}^{n} a_y\, x^y$$

where $a_y \triangleq \dfrac{n!}{y!\,(n-y)!} f(y)$, for $y = 0, 1, \ldots, n$,

and $x \triangleq \dfrac{\theta}{1-\theta}$.

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

$$\sum_{y=0}^{n} a_y x^y = 0 \quad \text{for all } x>0$$

$\Rightarrow$ $a_y = 0$ for $y=0, 1, \cdots, n$.

$\Rightarrow$ $f(y)=0$ for $y=0, 1, \cdots, n$.

$\Rightarrow$ $\{P_\theta ; \theta\in\Lambda\}$ is complete.

Note that this completeness is retained for any Λ containing at least $(n+1)$ nonzero parameter values.

* The notions of completeness and sufficiency are closely related.

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

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

Since T is sufficient, $f(y)$ does not depend on θ. For each $\theta\in\Lambda$, we have

$$E_\theta\{f(Y)\} = E_\theta\{Y\} - E_\theta\{E_\theta\{Y/T(Y)\}\}$$

$$= E_\theta(Y) - E_\theta(Y) = 0$$

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

Since $E_\theta\{Y | T(Y) = T(y)\}$ is a function of $T(y)$, the later condition implies that y itself is a function of $T(y)$. But, $T(y)$ is a function of y. So, $T(y)$ must be a one-to-one function of y; that is, T is a trivial sufficient statistic. This implies that if $\{P_\theta; \theta \in \Lambda\}$ is complete, then there is no nontrivial sufficient statistic for θ, i.e., the observation Y can not be reduced without destroying information about θ.

* Completeness is a useful concept in characterizing MVUEs.

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 statistics.

Suppose that T is complete and let $\tilde{g}[T(y)]$ and $g^*[T(y)]$ be any functions of $T(y)$ that are unbiased estimates of $g(\theta)$. We have

$$E_\theta\{\tilde{g}[T(Y)] - g^*[T(Y)]\}$$

$$= E_\theta\{\tilde{g}[T(Y)]\} - E_\theta\{g^*[T(Y)]\}$$

$$= g(\theta) - g(\theta) = 0, \quad \text{for all } \theta \in \Lambda.$$

Thus, by the completeness of T, we have

$P_\theta(\tilde{g}[T(Y)] = g^*[T(Y)]) = 1$ for all $\theta \in \Lambda$,

i.e., $\tilde{g}(T(Y))$ and $g^*[T(Y)]$ are the same estimator.

⇒ Any unbiased estimator that is a function of a complete sufficient statistic is unique and thus is an MVUE.

⇒ 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, $\tilde{g}[T(y)] \triangleq E_\theta\{\hat{g}(Y) \mid T(Y) = T(y)\}$ is an MVUE of $g(\theta)$.

* Definition 4.3.4: Exponontial Families

A Class of Distributions $\{P_\theta;\ \theta \in \Lambda\}$ is said to be an exponential family if there are real-valued functions C, $Q_1, \cdots, Q_m$, $T_1, \cdots, T_m$, and h such that P_θ has density

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

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

Many distributions can be put into the form of exponential families, such as, Gaussian, Poisson, Laplacian, binomial, geometric, and certain multivariate forms of these. Exponential families play an important role in the theory of MVUE by the following result.

* Proposition 4.3.3: 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

$$p_\theta(y) = C(\theta)\exp\left\{\sum_{\ell=1}^{m} \theta_\ell T_\ell(y)\right\} h(y), \qquad (*)$$

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

Outline of Proof: A complete proof can be found in Lehmann (1986).

We first note that T is sufficient for θ by the factorization theorem, so we only need to show the completeness of T.

With Y distributed according to $(*)$, $T(Y)$ has a density (on $\mathbb{R}^m$) of the form

$$g_\theta(t) = C(\theta)\exp\left\{\sum_{\ell=1}^{m} \theta_\ell t_\ell\right\} h_T(t)$$

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

$$E_\theta\{f[T(Y)]\} = C(\theta)\int_{\mathbb{R}^m} f(t)\exp\{\sum_{l=1}^{m}\theta_l t_l\}\, h_T(t)\,\mu(dt) \qquad (**)$$

Suppose that Λ contains an m-dimensional rectangle $J=\{\theta \mid a_1\le\theta_1\le b_1,\ a_2\le\theta_2\le b_2,\ \dots,\ a_m\le\theta_m\le b_m\}$. J can be changed to

$$J'=\{\theta \mid -1\le\theta_1\le 1,\ -1\le\theta_2\le 1,\ \dots,\ -1\le\theta_m\le 1\}.$$

Consider $(**)$ as a function of a complex variable by replacing θ_l with $\theta_l + i u_l$, $l=1,\cdots,m$. It canbe shown that this function is analytic in the region $C=\{\theta+iu \mid -1\le\theta_l\le 1,\ -\infty\le u_l<\infty,\ l=1,\cdots,m\}$,

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

$$C'=\{\theta+iu \mid \theta_l=0,\ -\infty<u_l<\infty,\ l=1,\cdots,m\},$$

i.e.,

$$C(\theta)\int_{\mathbb{R}^m} f(t)\exp\{i\sum_{l=1}^{m}u_l t_l\}\, h_T(t)\,\mu(dt)=0$$

for all $u\in\mathbb{R}^m$, which is a multidimensional Fourier tramform. This implies $P_\theta(f(Y)=0)=1$ for all $\theta\in\Lambda$. $\square$.

* Example 4.3.3. Minimum-Variance Unbiased Estimation (MVUE) of Signal Amplitude

Consider the model $Y_k = N_k + \mu S_k$, $k=1,2,\cdots,n$, where $N_1,\cdots,N_n$ 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 signed 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

$$\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(y_k-\mu s_k)^2\}$$
$$= c(\theta_1)\exp\{\theta_1 T_1(\underline{y})\}\, h(\underline{y}) \qquad (**)$$

where $\theta_1 = \mu/\sigma^2$

$$T_1(\underline{y}) = \sum_{k=1}^{n} s_k y_k$$
$$c(\theta_1) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{-\frac{\theta_1^2\sigma^2}{2}\sum_{k=1}^{n} s_k^2\}$$

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

Assuming that μ is an arbitrary real number, the parameter set is $\Lambda = \{\theta_1 \mid -\infty<\theta_1<\infty\} = \mathbb{R}$. A one-dimensional rectangle is an interval, and Λ contains an interval, so from Proposition 4.3.3 and (**), we see that $T_1(y)$ 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 $s_1 \neq 0$, the estimate $\hat{g}(\underline{y}) = y_1/s_1$ is an unbiased estimator of $g(\theta)$. Since T_1 is complete, the estimate

$$\tilde{g}[T_1(\underline{y})] = E_\theta\{\hat{g}(\underline{Y}) / T_1(\underline{Y}) = T_1(y)\}$$

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

$$E_\theta\{\hat{g}(\underline{Y})\} = \mu,$$

$$E_\theta\{T_1(\underline{Y})\} = n\mu\overline{s^2}$$

$$Var_\theta\{\hat{g}(\underline{Y})\} = \sigma^2/s_1^2$$

$$Var_\theta\{T_1(Y)\} = n\sigma^2\overline{s^2}$$

and $Cov_\theta[\hat{g}(\underline{Y}), T_1(\underline{Y})] = \sigma^2$
where $\overline{s^2} = \frac{1}{n}\sum_{k=1}^{n} s_k^2$.

So, applying the results of Section $\overset{(IV.B)}{4.2}$ [

$$\hat{\underline{\mu}}(\underline{y}) = \underline{\mu}_\theta + \Sigma_{\theta y}\Sigma_y^{-1}(\underline{y} - \underline{\mu}_y)]$$

we have

$$\tilde{g}[T_1(\underline{y})] = E_\theta\{\hat{g}(\underline{Y})\} + Cov_\theta[\hat{g}(\underline{Y}), T_1(\underline{Y})]$$

$$\times [Var_\theta[T_1(\underline{Y})]]^{-1}[T_1(\underline{y}) - E_\theta\{T_1(\underline{Y})\}]$$

$$= \mu + \sigma^2 (n\sigma^2 - \overline{s^2})^{-1} [T_1(\underline{y}) - n\mu \overline{s^2}]$$

$$= T_1(\underline{y}) / (n\overline{s^2}) = \frac{\sum_{k=1}^{n} s_k y_k}{n\overline{s^2}} \qquad (\checkmark)$$

Thus, we have constructed an <u>MVUE</u> for the signal amplitude μ. The variance of this estimator is $\mathrm{Var}_\theta(\tilde{g}[T_1(\underline{Y})]) = \dfrac{\sigma^2}{n\overline{s^2}}$.

When both μ and σ^2 are unknown, with μ ranging over $\mathbb{R}$ and σ^2 ranging over $(0, \infty)$, we also want to estimate both of these parameters.

Note that the previous $h(\underline{y})$ is a function of σ^2, so (***) is not a correct exponential family if σ^2 is not known. But it can be written as

$$\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(y_k - \mu s_k)^2\right\}$$

$$= C(\theta)\exp\{\theta_1 T_1(\underline{y}) + \theta_2 T_2(\underline{y})\}\, h(\underline{y})$$

where θ_1 and T_1 are as in (***), but we define

$$\theta = (\theta_1, \theta_2),$$

$$\theta_2 = -1/(2\sigma^2)$$

$$T_2(\underline{y}) = \sum_{k=1}^{n} y_k^2$$

and $C(\theta) = \left(-\dfrac{\theta_2}{\pi}\right)^{n/2} \exp\left\{\dfrac{\theta_1^2}{4\theta_2}\sum_{k=1}^{n} s_k^2\right\}$

and $h(\underline{y}) = 1$.

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

We wish to estimate $\mu = g_1(\theta) \stackrel{\Delta}{=} -\theta_1/(2\theta_2)$
and $\sigma^2 = g_2(\theta) \stackrel{\Delta}{=} -\frac{1}{2\theta_2}$

Note that the estimate in (v) is computed without knowledge of σ^2, it is unbiased and is a function of $T_1(\underline{y})$ hence a function 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 look for an unbiased estimator of σ^2 and then condition it on $T(\underline{y})$.

Since $T_1(\underline{Y}) \sim \mathcal{N}(n\mu\overline{s^2}, n\sigma^2\overline{s^2})$, we have

$$E_\theta\{T_1^2(\underline{Y})\} = Var_\theta[T_1(Y)] + (E_\theta\{T_1(\underline{Y})\})^2$$

$$= n\sigma^2\overline{s^2} + n^2\mu^2(\overline{s^2})^2.$$

Also, $E_\theta\{T_2(\underline{Y})\} = \sum_{k=1}^{n} E_\theta(Y_k^2)$

$$= \sum_{k=1}^{n} (\sigma^2 + \mu^2 s_k^2)$$

$$= n\sigma^2 + n\mu^2\overline{s^2}.$$

From these two results we see that the quality

$$T_2(\underline{Y}) - T_1^2(\underline{Y})/(n\overline{s^2})$$

has mean

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

Thus, $\tilde{g}_2[T(\underline{y})] = [T_2(\underline{y}) - T_1^2(\underline{y})/(n\overline{s^2})]/(n-1)$

is an unbiased estimator of σ^2. By the completeness of T, it is an MVUE. We rewrite $\tilde{g}_2$ as

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

where $\hat{\mu}$ is the MVUE of μ from (V).
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 $\frac{1}{n-1}\sum_{k=1}^{n}(\hat{n}_k)^2$.

Note that a more natural estimator for the second moment would be

$$\frac{1}{n}\sum_{k=1}^{n}(\hat{n}_k)^2;$$

but as we see from the analysis above, the latter estimate is biased.

The above theory provides a method to find minimum-variance unbiased estimator. However for many models of interest, it may be hard to apply the theory.

Then, the question is how we can evaluate an estimator, if it is not the optimal one.

* Proposition 4.3.4: 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, that is the set $\{y \mid p_\theta(y)>0\}$ is the same for all $\theta\in\Lambda$,
3) $\dfrac{\partial p_\theta(y)}{\partial\theta}$ exists and is finite for all $\theta\in\Lambda$ and all y in the support of p_θ
4) $\dfrac{\partial}{\partial\theta}\int_\Gamma h(y)\,p_\theta(y)\,\mu(dy)$ exists and equals to $\int_\Gamma h(y)\,\dfrac{\partial p_\theta(y)}{\partial\theta}\,\mu(dy)$, for all $\theta\in\Lambda$, for $h(y)=\hat{\theta}(y)$ and $h(y)=1$.

Then,
$$\mathrm{Var}_\theta[\hat{\theta}(Y)] \geq \frac{\left[\frac{\partial}{\partial\theta}E_\theta\{\hat{\theta}(Y)\}\right]^2}{I_\theta} \qquad (*)$$

where $I_\theta \triangleq E_\theta\{(\frac{\partial}{\partial\theta}\log p_\theta(Y))^2\}$.

Furthermore, if the following condition also holds:

5) $\frac{\partial^2 p_\theta(y)}{\partial\theta^2}$ 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

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

Proof: $E_\theta\{\hat\theta(Y)\} = \int_\Gamma \hat\theta(y)\,p_\theta(y)\mu(dy).$

Using 4), $\frac{\partial}{\partial\theta}E_\theta\{\hat\theta(Y)\} = \int_\Gamma \hat\theta(y)\frac{\partial}{\partial\theta}p_\theta(y)\mu(dy).$

4) also implies

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

$$= \frac{\partial}{\partial\theta}(1) = 0.$$

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

$$= \int_\Gamma (\hat\theta(y) - E_\theta\{\hat\theta(Y)\})\left[\frac{\partial}{\partial\theta}\log p_\theta(y)\right]p_\theta(y)\,\mu(dy)$$

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

Applying the Schwarz inequality, we have

$$\left(\frac{\partial}{\partial\theta}E_\theta\{\hat\theta(Y)\}\right)^2 \le \underbrace{E_\theta\{[\hat\theta(Y)-E_\theta\{\hat\theta(Y)\}]^2\}}_{\parallel\ \mathrm{Var}_\theta[\hat\theta(Y)]}\, I_\theta$$

This leads to the first result.

Note that $\frac{\partial^2}{\partial\theta^2}\log p_\theta(Y) = \frac{\partial^2}{\partial\theta^2}p_\theta(Y)\Big/p_\theta(Y)$

$$- \left(\frac{\partial}{\partial\theta}\log p_\theta(Y)\right)^2.$$

Taking $E_\theta\{\cdot\}$ on both sides, 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 5), $\int_\Gamma \frac{\partial^2}{\partial\theta^2}p_\theta(y)\,\mu(dy) = \frac{\partial^2}{\partial\theta^2}\int_\Gamma p_\theta(y)\,\mu(dy)$

$$= 0.$$

This proves the second result.

□

The quantity I_θ defined above is known as Fisher's information for estimating θ from Y, $(*)$ is called the information inequality.

The existence of an estimate that achieves equality in the information inequality is possible only under special circumstance.

For the particular case in which $\hat{\theta}$ is unbiased, $E_\theta\{\hat{\theta}(Y)\} = \theta$, the information inequality reduces to

$$\mathrm{Var}_\theta[\hat{\theta}(Y)] \geq \frac{1}{I_\theta}$$

which is known as the Cramér-Rao lower bound (CRLB).

* Example 4.3.4: The Information Inequality for Exponential Families

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

$$p_\theta(y) = c(\theta)e^{g(\theta)T(y)}h(y) \qquad (*)$$

where c, g, T, and h are real-valued functions and $g(\theta)$ has derivative $g'(\theta)$. Assume $E_\theta\{|T(Y)|\} < \infty$ and

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

and Conditions 1)-4) hold. Since $p_\theta(y)$ is a density function, we may write

$$c(\theta) = \left[\int_\Gamma e^{g(\theta)T(y)}h(y)\mu(dy)\right]^{-1}.$$

To compute I_θ for this family of densities, we write

$$\log p_\theta(y) = g(\theta)T(y) + \log h(y) - \log\left[\int_\Gamma e^{g(\theta)T(y)}h(y)\mu(dy)\right],$$

Then, $\frac{\partial}{\partial\theta}\log p_\theta(y) = g'(\theta)T(y)$

$$- \frac{g'(\theta)\int_\Gamma T(y)e^{g(\theta)T(y)}h(y)\mu(dy)}{\int_\Gamma e^{g(\theta)T(y)}h(y)\mu(dy)}$$

$$= g'(\theta)[T(y) - E_\theta\{T(Y)\}].$$

Thus, $I_\theta \triangleq E_\theta\{(\frac{\partial}{\partial\theta}\log p_\theta(Y))^2\}$

$$= [g'(\theta)]^2 E_\theta\{[T(Y) - E_\theta\{T(Y)\}]^2\}$$
$$= [g'(\theta)]^2 \mathrm{Var}_\theta[T(Y)],$$

and the information inequality in this case is

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

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

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

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

and the lower bound in the information inequality equals

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

This means that $T(Y)$ achieves the information lower bound, so it has the minimum variance among all estimators $\hat{\theta}$ satisfying

$$\frac{\partial}{\partial\theta} E_\theta\{\hat{\theta}(Y)\} = \frac{\partial E_\theta\{T(Y)\}}{\partial\theta}.$$

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.

From the above analysis, we see that the exponential form $(*)$ is sufficient for the variance of T to achieve the information lower bound within the regularity assummed above. It turns out that this form is also necessary for achieving the lower bound for all $\theta \in \Lambda$, again within the regularity conditions.

necessity:

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 before. It happens if and only if

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

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

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

where $h(y)$ does not depend on θ. It has the exponential form (*) with h as given,

$$C(\theta) = \exp\left\{-\int_a^\theta k(\sigma)f(\sigma)\,d\sigma\right\},$$

$$g(\theta) = \int_a^\theta k(\sigma)\,d\sigma,$$

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

$\square$

We conclude within regularity, the information lower bound is achieved by $\hat\theta$ if and only if $\hat\theta(y) = T(y)$ in a one-dimensional exponential family.

§4.4. Maximum-Likelihood Estimation

In many cases, it is not possible to find MVUEs. Then, we need to find good estimators. One of good estimators is maximum-likelihood estimator.

To motivate maximum-likelihood estimation, let us first consider MAP estimation:

$$\hat\theta_{MAP}(y) = \arg\left\{\max_{\theta\in\Lambda} p_\theta(y)w(\theta)\right\}.$$

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, this estimate is called the maximum likelihood estimate (MLE)

$$\hat{\theta}_{ML}: \quad \hat{\theta}_{ML}(y) = \arg\left\{\max_{\theta\in\Lambda} p_\theta(y)\right\}.$$

There are two things wrong with the above argument.

a), it is not always possible to construct a uniform distribution on Λ, since Λ may not be a bounded set.

b), more importantly, 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.

Despite these, MLE is very useful one in many applications, such as communications.

Maximizing $p_\theta(y)$ is equivalent ~~to~~ maximizing $\log p_\theta(y)$, and assumming sufficient smoothness of this function, a necessary condition for the MLE is

$$\frac{\partial}{\partial\theta}\log p_\theta(y)\Big|_{\theta=\hat{\theta}_{MLE}(y)} = 0,$$

which is known as the <u>likelihood equation</u>, and we will see that its solutions have useful propeties even when they are not maxima of $p_\theta(y)$.

For example, suppose we have equality in the Cramer-Rao Lower Bound, i.e., suppose that $\hat{\theta}$ is an unbiased estimate of θ with

$$Var_\theta[\hat{\theta}(Y)] = \frac{1}{I_\theta}$$ [Note that such a $\hat{\theta}$ is an MVUE of θ]

Then, the equality in Schwarz inequality holds and

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

where we used the facts $f(\theta) = \theta$, $k(\theta) = \frac{I_\theta}{f'(\theta)}$.

Then, the likelihood equation becomes

$$\frac{\partial}{\partial\theta}\log p_\theta(y)\Big|_{\theta=\hat{\theta}_{ML}(y)} = I_\theta[\hat{\theta}(y) - \theta]\Big|_{\theta=\hat{\theta}_{ML}(y)}$$

$$= 0$$

$$\Rightarrow \hat{\theta}_{ML}(y) = \hat{\theta}(y).$$

⇒ If $\hat{\theta}$ achieves the CRLB, it is the solution to the likelihood equation.

⇒ Only solutions to the likelihood equation can achieve the CRLB.

Unfortunately, it is not always true that solutions to the likelihood equation can achieve the CRLB or even that are unbiased.

* Example 4.4.1: Maximum-Likelihood Estimation of the Parameter of the Exponential Distribution

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

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

Thus,

$$p_\theta(\underline{y}) = \theta^n \exp\{-\theta n \bar{y}\} \text{ with}$$

$\bar{y} \triangleq \frac{1}{n} \sum_{k=1}^{n} y_k$, and the likelihood equation is

$$\frac{\partial}{\partial\theta} \log p_\theta(\underline{y}) \Big|_{\theta=\hat{\theta}_{ML}(\underline{y})} = \frac{n}{\theta} - n\bar{y} \Big|_{\theta=\hat{\theta}_{ML}(\underline{y})}$$

$$= 0$$

$$\Rightarrow \quad \hat{\theta}_{ML}(\underline{y}) = \frac{1}{\bar{y}}.$$

Since $\frac{\partial^2}{\partial \theta^2} \log p_\theta(\underline{y}) = -\frac{n}{\theta^2} < 0$, the above solution $\hat{\theta}_{ML}(\underline{y})$ is the unique maximum of $p_\theta(y)$.

Note that $E_\theta\{Y_k\} = \frac{1}{\theta}$, so $E_\theta(\bar{Y}) = \frac{1}{\theta}$

and thus, it makes sense to estimate θ as $\frac{1}{\bar{y}}$.

The weak law of large numbers implies that $\bar{Y} \rightarrow \frac{1}{\theta}$ in probability under P_θ, i.e.,

$$\frac{1}{\bar{Y}} \rightarrow \theta \text{ in probability under } P_\theta,$$

or the MLE converges in probability to the true parameter value, a property known as <u>consistency</u>. 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 is

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

Thus, the CRLB is $\frac{\theta^2}{n}$.

Since $\frac{\partial \log p_\theta(y)}{\partial \theta}$ 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. It is easy to see

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

thus, when $n > 1$,

$$E_\theta\{\hat{\theta}_{ML}(\underline{Y})\} = E_\theta\{\frac{1}{\bar{y}}\} = \frac{n\theta}{n-1},$$

and for $n > 2$,

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

From the mean, although $\hat{\theta}_{ML}(\underline{Y})$ is biased, it does have the property $\lim_{n\to\infty} E_\theta\{\hat{\theta}_{ML}(\underline{Y})\} = \theta$, i.e., asymptotically unbiased. Also,

$$Var_\theta[\hat{\theta}_{ML}(\underline{Y})]\, I_\theta = \frac{n^3}{(n-1)^2(n-2)} \to 1 \text{ as } n \to \infty.$$

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

As a final comment on this example, we note from Proposition 4.3.3 that $\overline{Y}$ is a complete sufficient statistic for θ in this model. Also from the above mean calculation,

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

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 and

$$\mathrm{Var}_\theta\{\hat{\theta}_{MV}(\underline{Y})] = \frac{\theta^2}{n-2}, \qquad (*)$$

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 (*) equals the MSE of $\hat{\theta}_{MV}$ since it is unbiased. For the MLE, the MSE is

$$E_\theta\{[\hat{\theta}_{ML}(\underline{Y}) - \theta]^2\} = \mathrm{Var}_\theta[\hat{\theta}_{ML}(\underline{Y})] + b^2(\theta)$$

where $b(\theta) \triangleq E_\theta\{\hat{\theta}_{ML}(\underline{Y})\} - \theta$ is the bias of $\hat{\theta}_{ML}$. Thus,

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

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

* Example 4.4.2: Maximum-Likelihood Estimation of Signal Amplitude

Consider the model in Example 4.3.3:

$$Y_k = N_k + \mu s_k, \quad k=1,2,\cdots,n,$$

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

$$-\frac{\partial}{\partial\mu}\left(\frac{1}{2}\sum_{k=1}^{n}\log(2\pi\sigma^2)+\frac{1}{2\sigma^2}\sum_{k=1}^{n}(y_k-\mu s_k)^2\right)\Bigg|_{\mu=\hat{\mu}_{ML}(\underline{y})}$$

$$=\frac{1}{\sigma^2}\sum_{k=1}^{n}s_k\left[y_k-\hat{\mu}_{ML}(\underline{y})\,s_k\right]=0$$

$$\Rightarrow \quad \hat{\mu}_{ML}(\underline{y})=\frac{1}{n}\,\frac{\sum_{k=1}^{n}s_k y_k}{\overline{s^2}}$$

where $\overline{s^2}\triangleq\frac{1}{n}\sum_{k=1}^{n}s_k^2$.

Since

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

$\log p_\theta(y)$ is concave in μ and the solution to the likelihood equation does give a global maximum.

Note that $\hat{\mu}_{ML}$ is the same as the MVUE of μ

in (v) in Example 4.3.3, so that

$$E_\theta\{\hat{\mu}_{ML}(\underline{Y})\} = \mu, \quad Var_\theta[\hat{\mu}_{ML}(\underline{Y})] = \frac{\sigma^2}{n\overline{s^2}}.$$

And it is not hard to have $I_\theta = \frac{n\overline{s^2}}{\sigma^2}$.

So, $CRLB = \frac{\sigma^2}{n\overline{s^2}} = Var_\theta[\hat{\mu}_{ML}(\underline{Y})]$.

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 $k(\theta) = I_\theta = n\frac{\overline{s^2}}{\sigma^2}$.

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

$$\frac{\partial}{\partial\sigma^2}\log p_\theta(\underline{y})\Big|_{\sigma^2 = \hat{\sigma}^2_{ML}(\underline{y})}$$

$$= \frac{1}{2\hat{\sigma}^2_{ML}(\underline{y})} - \frac{1}{2[\hat{\sigma}^2_{ML}(\underline{y})]^2}\sum_{k=1}^{n}(y_k - \mu s_k)^2 = 0$$

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

Since $\frac{\partial}{\partial\sigma^2}\log p_\theta(\underline{y}) = \frac{n}{2\sigma^4}[\hat{\sigma}^2_{ML}(\underline{y}) - \sigma^2]$,

$\log p_\theta(\underline{y})$ increases in σ^2 when $\sigma^2 < \hat{\sigma}^2_{ML}(\underline{y})$

decreases in σ^2 - - - -> - - - .

$\Rightarrow$ log $p_\theta(\underline{y})$ achieves its absolute maximum at $\hat{\sigma}^2_{ML}(\underline{y})$.

Also, if let $\theta = \sigma^2$,

$$\frac{\partial}{\partial\theta}\log_{\theta}(\underline{y}) = \frac{n}{2\theta^2}\left[\hat{\theta}_{ML}(\underline{y}) - \theta\right]. \quad (*)$$

which from Example 4.3.4 implies that $\hat{\sigma}^2_{ML}(\underline{y})$ is unbiased and achieves the CRLB, and thus that $\hat{\sigma}^2_{ML}(\underline{y})$ is an MVUE of σ^2. From (*), we have

$$I_\theta = \frac{n}{2\theta^2} \equiv \frac{n}{2\sigma^4} \quad \text{and}$$

$$\text{CRLB} = \frac{2\sigma^4}{n} = \text{Var}_\theta\left[\hat{\sigma}^2_{ML}(\underline{Y})\right].$$

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

$$\max_{(\mu,\sigma^2)} \log p_\theta(\underline{y}) = \max_{\sigma^2}\left\{\max_{\mu} \log p_\theta(\underline{y})\right\}$$

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

The right hand side of the above equation is the same maximimization problem as for estimating

σ^2 with known μ, if μ is set to $\hat{\mu}_{ML}(\underline{y})$.
Thus,

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

and $$\hat{\sigma}^2_{ML}(\underline{y}) = \frac{1}{n}\sum_{k=1}^{n}\left[y_k - \hat{\mu}_{ML}(\underline{y}) s_k\right]^2.$$

The estimate $\hat{\mu}_{ML}(\underline{y})$ is still an MVUE of μ here. However, from Example 4.3.3,

$$\hat{\sigma}^2_{ML}(\underline{y}) = \frac{n-1}{n}\hat{\sigma}^2_{MV}(\underline{y})$$

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

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

So that $\hat{\sigma}^2_{ML}(\underline{y})$ has lower variance than the MVUE :

$$Var_\theta(\hat{\sigma}^2_{MV}(\underline{Y})] = \frac{2\sigma^4}{n-1}.$$

The MSE of σ^2 for the MVUE and ML are

$$E_\theta\{[\hat{\sigma}^2_{MV}(\underline{Y}) - \sigma^2]^2\} = \frac{2\sigma^4}{n-1}$$

$$E_\theta\{[\hat{\sigma}^2_{ML}(\underline{Y}) - \sigma^2]^2\} = Var_\theta[\hat{\sigma}^2_{ML}(\underline{Y})] + [E_\theta\{\hat{\sigma}^2_{ML}(\underline{Y})\} - \sigma^2]^2$$

$$= \frac{(n-1)^2}{n^2}\frac{2\sigma^4}{n-1} + \left(\frac{n-1}{n}\sigma^2 - \sigma^2\right)^2 = \sigma^4\frac{2n-1}{n^2}.$$

The ratio of these two quantities is

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

In this case, the MLE 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.

$\Rightarrow$ achieving the goal of minimum-variance unbiased estimation does not always lead to an optimum estimate in terms of mean-squared error.

* Suppose that we have a sequence of i.i.d. observations $Y_1, \cdots, Y_n$, 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 $\psi(y_k;\theta)\triangleq\partial f_\theta(y_k)/\partial\theta$. Or equivalently,

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

For a fixed parameter value $\theta'\in\Lambda$, consider

$$\frac{1}{n}\sum_{k=1}^{n} \psi(Y_k;\theta').$$

Assume that θ is the true parameter value, i.e., $Y_k \sim f_\theta$, the weak law of large numbers implies

$$\frac{1}{n}\sum_{k=1}^{n} \psi(Y_k;\theta') \xrightarrow{\text{in probability}} E_\theta\{\psi(Y_1;\theta')\},$$

$$E_\theta\{\psi(Y_1;\theta')\} = \int_{\mathbb{R}} \frac{\partial}{\partial\theta}\log f_\theta(y_1)\Big/_{\theta=\theta'} f_\theta(y_1)\,\mu(dy_1)$$

$$\triangleq J(\theta;\theta').$$

Assume that the order of integration and differentiation Can be intorchanged. Then

$$J(\theta;\theta) = \int\Big[\frac{\partial}{\partial\theta} f_\theta(y_1)\Big] f_\theta(y_1)\,\mu(dy_1)$$

$$= \int \frac{\partial}{\partial\theta} f_\theta(y_1)\,\mu(dy_1)$$

$$= \frac{\partial}{\partial\theta}\int f_\theta(y_1)\,\mu(dy_1) = \frac{\partial}{\partial\theta}(1) = 0.$$

Thus, the equation $J(\theta;\theta')=0$ has a solution $\theta'=\theta$. Suppose that this is the unique root of $J(\theta;\theta')$ and $J(\theta;\theta')$ and $\frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta')$ are both smooth functions of θ'. Since $\frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta')$ 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. This implies that $\hat{\theta}_n(\underline{Y})$ should be close to the true parameter value θ when n

is large, i.e., $\hat{\theta}_n(Y) \to \theta$ as $n \to \infty$ in some statistical sense.

Within the appropriate smoothness and uniqueness conditions, solutions to the likelihood equations are consistent; i.e., they converge in probability to the true parameter value:

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

More precise conditions are as follows.

* Proposition 4.4.1: 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, $\frac{1}{n}\sum_{k=1}^{n} \psi(Y_k;\theta')$ has a unique root $\hat{\theta}_n$ (with probability 1).

Then, $\hat{\theta}_n \to \theta$ in probability.

Proof: Choose $\varepsilon > 0$, small enough.
By Condition 1), $J(\theta;\theta+\varepsilon)$ and $J(\theta;\theta-\varepsilon)$

must have opposite signs. Define

$$\delta = \min\{|J(\theta;\theta+\varepsilon)|, |J(\theta;\theta-\varepsilon)|\}.$$

For each n, define the events

$$A_n^+ = \{|J(\theta;\theta+\varepsilon) - \frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta+\varepsilon)| \le \delta\}$$

$$A_n^- = \{|J(\theta;\theta-\varepsilon) - \frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta-\varepsilon)| \le \delta\}$$

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

On A_n^+, $\frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta+\varepsilon)$ must have the same sign as $J(\theta;\theta+\varepsilon)$.

On A_n^-, $\frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta-\varepsilon)$ must have the same sign as $J(\theta;\theta-\varepsilon)$.

Thus, on A_n, $\frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta+\varepsilon)$ and $\frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta-\varepsilon)$ have opposite signs. By the continuity condition 2), $\frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta')$ can change sign only by passing through zero. Thus, on A_n, the root $\hat{\theta}_n$ is between $\theta-\varepsilon$ and $\theta+\varepsilon$.

This implies that $A_n \subset \{|\hat{\theta}_n - \theta| \le \varepsilon\}$, i.e.,

$$P(|\hat{\theta}_n - \theta| \le \varepsilon) \ge P(A_n).$$

By the weak law of large numbers,

$\frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta+\varepsilon) \to J(\theta;\theta+\varepsilon)$ in probability

and $\frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta-\varepsilon) \to J(\theta;\theta-\varepsilon)$ in probability.

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

$$1 \ge P(|\hat{\theta}_n-\theta|\le\varepsilon) \ge P(A_n) = P(A_n^+)+P(A_n^-) - P(A_n^+\cup A_n^-)$$

$$\ge P(A_n^+)+P(A_n^-) - 1 \to 1 \text{ as } n\to\infty.$$

$$\Rightarrow \quad P(|\hat{\theta}_n-\theta|\le\varepsilon)\to 1. \qquad \Box.$$

Remarks: The continuities can be relaxed to the local continuities.

* $\hat{\theta}_n \to \theta$ in probability may not imply

$$\lim_{n\to\infty} E_\theta\{\hat{\theta}_n\} = E_\theta\{\lim_{n\to\infty}\hat{\theta}_n\} \qquad (*)$$

In other words, the interchangeablity of the integration sign E_θ and the limit sign $n\to\infty$ may not be ensured by the convergence in probability. When does the interchangeability hold? It is the "real-analysis" about! It is called dominated convergence theorem: There exists a random variable Z such that $|\hat{\theta}_n|\le Z$ and $E_\theta(Z)<\infty$.

If (*) holds, the asymptotic unbiasedness will follow.

* Proposition 4.4.2: Asymptotic Normality of MLEs

Suppose that $\{Y_k\}_{k=1}^{\infty}$ is a sequence of i.i.d. random variable 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 before.
5) Condition 5) of Proposition 4.3.4 holds.

Then,

$$P_\theta\left(\sqrt{n i_\theta}\,(\hat{\theta}_n - \theta) \le x\right) \to \Phi(x) \quad \text{for all } x \in \mathbb{R},$$

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

$$\frac{1}{n}\sum_{k=1}^{n}\psi(Y_k; \hat{\theta}_n)$$

about θ to yield

$$\frac{1}{n}\sum_{k=1}^{n}\psi(Y_k;\theta)+(\hat{\theta}_n-\theta)\frac{1}{n}\sum_{k=1}^{n}\psi'(Y_k;\theta)$$

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

where $\hat{\theta}_n \in (\theta, \hat{\theta}_n)$. Rearranging it gives

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

By the weak law of large numbers,

$$\frac{1}{n}\sum_{k=1}^{n}\psi'(Y_k;\theta)\longrightarrow E_\theta\{\psi'(Y_1;\theta)\} \text{ in probability.}$$

By Condition 3),

$$\left|\frac{1}{2}(\hat{\theta}_n-\theta)\frac{1}{n}\sum_{k=1}^{n}\psi''(Y_k;\bar{\theta}_n)\right|\le\frac{1}{2}|\hat{\theta}_n-\theta|\left|\frac{\sum_{k=1}^{n}M(Y_k)}{n}\right|.$$

It is known $|\hat{\theta}_n-\theta|\to 0$ in probability.

The weak law of large numbers implies that

$$\frac{1}{n}\sum_{k=1}^{n}M(Y_k)\longrightarrow E_\theta\{M(Y_1)\}<\infty.$$

Thus, the denominator of the right hand side of (*) converges to $E_\theta\{\psi'(Y_1;\theta)\}$ in probability.

The numerator sum $\sum_{k=1}^{n}\psi(Y_k;\theta)$ in (*) 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

$$-\frac{1}{\sqrt{n}}\sum_{k=1}^{n}\psi(Y_k;\theta) \to \mathcal{N}(0, i_\theta) \quad \text{in distribution.}$$

Thus, $\sqrt{n}(\hat{\theta}_n-\theta)$ converges in distribution to a $\mathcal{N}(0, v^2)$ random variable with

$$v^2 = i_\theta / E_\theta^2\{\psi'(Y_1;\theta)\}.$$

Similar to before (Prop. 4.3.4),

$$E_\theta\{\psi'(Y_k;\theta)\} = -E_\theta\{\psi^2(Y_1;\theta)\} = -i_\theta$$

$$\Rightarrow \quad v^2 = \frac{1}{i_\theta} \qquad \square$$

Remark: For i.i.d., the Fisher's information is $I_\theta = n\, i_\theta$. The proposition implies that θ_n^2 is asymptotically $\mathcal{N}(\theta, \frac{1}{n i_\theta})$, i.e., asymptotically $\hat{\theta}_n$ has mean θ and variance $1/(n i_\theta)$ that is the CRLB. What we have proved in the proposition is that the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n-\theta)$ has zero mean and variance $\frac{1}{i_\theta}$, which is not the same as $E_\theta\{\sqrt{n}(\hat{\theta}_n-\theta)\} \to 0$ and $\mathrm{Var}_\theta(\sqrt{n}(\hat{\theta}_n-\theta)) \to 1/i_\theta$.

asymptotic unbiasedness

asymptotic efficiency.

Although this is the case, these two may hold with some additional conditions.

MLE to be an asymptotically optimum (MVUE) estimator.

§4.5: Further Aspects and Extensions of Maximum-Likelihood Estimation

§4.5.1. Estimation of Vector Parameters

The previous study can be generalized to the case of parameter vectors, say dimension m. In this case, the likelihood equation is a vector equation:

$$\frac{\partial}{\partial \theta_1} \log p_{\underline{\theta}}(y) \Big|_{\underline{\theta}=\hat{\underline{\theta}}} = 0$$

$$\vdots$$

$$\frac{\partial}{\partial \theta_m} \log p_{\underline{\theta}}(y) \Big|_{\underline{\theta}=\hat{\underline{\theta}}} = 0$$

which for i.i.d. models becomes

$$\sum_{k=1}^{n} \psi_1(y_k; \hat{\underline{\theta}}_n) = 0,$$

$$\vdots$$

$$\sum_{k=1}^{n} \psi_m(y_k; \hat{\underline{\theta}}_n) = 0,$$

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

* The information inequality (Prop. 4.3.4) can, within regularity, be extended to the vector case.

The Cramér-Rao Lower bound in the variance of unbiased estimates becomes

$$\mathrm{Cov}_{\underline{\theta}}(\hat{\underline{\theta}}) \geq I_{\underline{\theta}}^{-1} \qquad (*)$$

where $\mathrm{Cov}_{\underline{\theta}}(\hat{\underline{\theta}}) \triangleq E_{\underline{\theta}}\{(\hat{\underline{\theta}}-\underline{\theta})(\hat{\underline{\theta}}-\underline{\theta})^T\}$, and $I_{\underline{\theta}}$ is the $m \times m$ Fisher information matrix with j-lth element

$$(I_{\underline{\theta}})_{j,l} = E_{\underline{\theta}}\left(\left[\frac{\partial}{\partial\theta_j}\log p_{\underline{\theta}}(Y)\right]\left[\frac{\partial}{\partial\theta_l}\log p_{\underline{\theta}}(Y)\right]\right),$$

$I_{\underline{\theta}}$ is the covariance matrix of the zero-mean vector

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

$I_{\underline{\theta}}$ is non-negative definite. In the CRLB (*), $I_{\underline{\theta}}$ is assumed positive definite, and the inequality $A \geq B$ for matrices means that $A-B$ is nonnegative definite.

For the i.i.d. case, (*) becomes

$$\mathrm{Cov}_{\underline{\theta}}(\hat{\underline{\theta}}) \geq \frac{1}{n}\, i_{\underline{\theta}}^{-1}$$

where $(i_{\underline{\theta}})_{j,l} = E_{\underline{\theta}}\{\psi_j(Y_1;\underline{\theta})\,\psi_l(Y_1;\underline{\theta})\}$.

* Within conditions similar to those of Prop.4.4.1, solutions to the likelihood equations are consistent, i.e.,

$$\|\underline{\hat{\theta}}_n - \theta\| \triangleq \left[\frac{1}{m}\sum_{j=1}^{m}\left[(\hat{\underline{\theta}}_n)_j - \theta_j\right]^2\right]^{\frac{1}{2}} \to 0$$

as $n \to \infty$ in probability, and within conditions similar to those of Prop.4.4.2,

$$\sqrt{n}\,(\underline{\hat{\theta}}_n - \underline{\theta}) \to N(\underline{0}, i_{\underline{\theta}}^{-1}) \text{ as } n \to \infty$$

in distribution.

§ 4.5.2. Estimation of Signal Parameters

The asymptotic properties of MLEs can also be extended to some time-varying problems. In particular, we have the following real-valued observations:

$$Y_k = s_k(\theta) + N_k, \quad k = 1, 2, \cdots, n,$$

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

The maximum-likelihood estimate of θ solves the equation

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