Solutions to Steven Kay's Statistical Estimation book

Satish Bysany
Aalto University School of Electrical Engineering

March 1, 2011

[section]

1 Introduction

This is as set of notes describing solutions to Steven Kay's book *Fundamentals of Statistical Signal Processing: Estimation Theory.* A brief review of notation is in order.

1.1 Notation

- $\mathbf{I}$ is identity matrix.
- $\mathbf{0}$ represents a matrix or vector of all zeros.
- $\mathbf{e}$ is a column vector of all ones.
- $\mathbf{J}$ is exchange matrix, with 1s on the anti-diagonal and 0s elsewhere.
- $\mathbf{e_j}$ is a column vector whose j^{th} element is 1, rest all 0.
- $\mathbf{a} \cdot \mathbf{b} \doteq \mathbf{a}^H \mathbf{b}$ is the dot product of $\mathbf{a}$ and $\mathbf{b}$
- $\frac{\partial}{\partial \mathbf{t}} f(\mathbf{t})$ is the derivative of a scalar function $f(\mathbf{t})$ depending on $M \times 1$ real vector parameter $\mathbf{t}$, is defined by

$$\frac{\partial}{\partial \mathbf{t}} f(\mathbf{t}) = \begin{bmatrix} \frac{\partial}{\partial t_1} f(\mathbf{t}) \\ \frac{\partial}{\partial t_2} f(\mathbf{t}) \\ \vdots \\ \frac{\partial}{\partial t_M} f(\mathbf{t}) \end{bmatrix}$$

- $\frac{\partial}{\partial t}\mathbf{h}(t)$ is the derivative of a $M \times 1$ real vector function $\mathbf{h}(t)$ depending upon a scalar value t.

$$\frac{\partial}{\partial t}\mathbf{f}(t) = \begin{bmatrix} \frac{\partial}{\partial t} f_1(t) \\ \frac{\partial}{\partial t} f_2(t) \\ \vdots \\ \frac{\partial}{\partial t} f_M(t) \end{bmatrix}$$

2 Chapter 2

Solutions to Problems in Chapter 2

2.1 Problem 2.1

The data $\mathbf{x} = \{x[0], x[1], \ldots, x[N-1]\}$ are observed where the $x[n]$'s are i.i.d. as $\mathcal{N}(0, \sigma^2)$. We wish to estimate the variance σ^2 as

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=0}^{N-1} x^2[n] \tag{1}$$

Solution From the problem definition, it follows that, $\forall\ n$,

$$\mu = E\left(x[n]\right) = 0$$
$$\sigma^2 = E\left((x[n]-\mu)^2\right) = E\left(x^2[n]\right)$$

Now take the $E(\cdot)$ operator on both sides of Eq(1) and using the fact that, for *any* two random variables X and Y,

$$E(X+Y) = E(X) + E(Y)$$

$$E\left(\hat{\sigma}^2\right) = \frac{1}{N}\sum_{n=0}^{N-1} E\left(x^2[n]\right) = \frac{1}{N}\sum_{n=0}^{N-1} \sigma^2 = \frac{N\sigma^2}{N} = \sigma^2 \tag{2}$$

Hence the estimator 1 is unbiased. Note that, this result holds even if the $x[n]$'s are *not* independent!

Next, applying the variance operator $\mathrm{var}(\cdot)$ on both sides of Eq(1) and using the fact that, for *independent* random variables X and Y,

$$\mathrm{var}(aX + bY) = a^2\ \mathrm{var}(X) + b^2\ \mathrm{var}(Y)$$

$$\operatorname{var}\left(\hat{\sigma}^2\right) = \frac{1}{N^2}\sum_{n=0}^{N-1} \operatorname{var}\left(x^2[n]\right) \tag{3}$$

Let $X \sim \mathcal{N}(0,1)$ be normal distribution with zero-mean and unit variance. Then, by definition, $Y = X^2 \sim \chi_1^2$ is chi-square distributed with 1 degree of freedom. We know that $\operatorname{mean}\left(\chi_n^2\right) = n, \operatorname{var}\left(\chi_n^2\right) = 2n$, so, $\operatorname{var}(Y) = \operatorname{var}(X^2) = 2 \cdot 1 = 2$.

Introducing $Z = \sigma X$, implies that $\operatorname{var}(Z) = \sigma^2 \operatorname{var}(X) = \sigma^2$. Since $E(Z) = \sigma E(X) = 0$, we conclude $Z \sim \mathcal{N}(0, \sigma^2)$.

Now consider $\operatorname{var}(Z^2) = \operatorname{var}(\sigma^2 X^2) = \sigma^4 \operatorname{var}(X^2) = 2\sigma^4$. Since each of $x[n] \sim \mathcal{N}(0, \sigma^2)$, we have,

$$\operatorname{var}(x^2[0]) = \operatorname{var}(x^2[1]) = \cdots = \operatorname{var}(x^2[N-1]) = 2\sigma^4$$

Hence, Eq(3) simplifies to

$$\operatorname{var}\left(\hat{\sigma}^2\right) = \frac{1}{N^2}\sum_{n=0}^{N-1}(2\sigma^4) = \frac{2\sigma^4 N}{N^2} = \frac{2\sigma^4}{N} \tag{4}$$

As $N \to \infty, \operatorname{var}\left(\hat{\sigma}^2\right) \to 0$.

2.2 Problem 2.5

Two samples $\{x[0], x[1]\}$ are independently observed from $\mathcal{N}\left(0, \sigma^2\right)$ distribution. The estimator

$$\hat{\sigma}^2 = \frac{1}{2}\left(x^2[0] + x^2[1]\right) \tag{5}$$

is unbiased. Find the PDF of $\hat{\sigma}^2$ to determine if it is symmetric about σ^2

Solution Consider two standard normal random variables X_0 and X_1, that is, $X_i \sim \mathcal{N}(0,1), i = 0, 1$. Then, by definition, $X = X_0^2 + X_1^2$ is $\chi^2(n)$-distributed with $n = 2$ degrees of freedom. It's PDF is

$$f_X(x) = \frac{1}{2}e^{-x/2} \qquad x > 0$$

Let $x[0] = \sigma X_0$ and $x[1] = \sigma X_1$. Then

$$\begin{aligned} x^2[0] + x^2[1] &= \sigma^2(X_0^2 + X_1^2) = \sigma^2 X \\ \implies \hat{\sigma}^2 &= \frac{\sigma^2}{2}X \qquad \text{from Eq(5)} \end{aligned}$$

We know that, for two *continuous* random varaibles X and Y related as $Y = aX + b$,

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

Taking $a = \frac{\sigma^2}{2}, b = 0, \theta = \sigma^2$, the PDF of $\hat{\sigma}^2$ is

$$f_{\hat{\sigma}^2}(y;\theta) = \frac{1}{a} f_X\left(\frac{y}{a}\right) = \frac{2}{\sigma^2}\left(\frac{1}{2}e^{\frac{-y}{2a}}\right) = \frac{1}{\sigma^2}e^{-y/\sigma^2} = \frac{1}{\theta}e^{-y/\theta} \qquad y > 0$$

It's obvious that $f_{\hat{\sigma}^2}(y;\theta) \neq f_{\hat{\sigma}^2}(y;-\theta)$, so the PDF is not symmetric about $\theta = \sigma^2$. Note carefully that the PDF is symmetric about σ, not σ^2.

3 Chapter 3: CRLB

3.1 Formulas

Let a random variable X depend on some parameter t. We write the PDF of X as $f_X(x;t)$ – it represents a family of PDFs, each one with a different value of t. When the PDF is viewed as a function of t for a *given, fixed* value of x, it is termed as likelihood function. We define, the log-likelihood function as

$$L(t) \doteq L_X(t|x) \doteq \ln f_X(x;t) \tag{6}$$

Note that t is a deterministic, but unknown parameter. We simply write it as $L(t)$ when the random variable X is known from context. For the sake of notation, we define

$$\dot{L} = \frac{\partial}{\partial t}L(t) = \frac{\partial}{\partial t}\ln f_X(x;t) = \frac{1}{f_X(x;t)}\frac{\partial}{\partial t}f_X(x;t) \tag{7}$$

$$\ddot{L} = \frac{\partial^2}{\partial t^2}L(t) = \frac{\partial^2}{\partial t^2}\ln f_X(x;t) \tag{8}$$

Taking the expectation w.r.t X, if the **regularity condition**

$$E(\dot{L}) = 0 \tag{9}$$

is satisfied, then there exists a lower bound on the variance of an *unbiased* estimator $\hat{t}$,

$$\mathrm{var}(\hat{t}\,) \geq \frac{1}{-E(\ddot{L})} \tag{10}$$

Furthermore, for the equality sign, and for all t,

$$\operatorname{var}(\hat{t}\,) = \frac{1}{-E(\ddot{L})} \iff \dot{L} = g(t)(h(x) - t) \iff \hat{t} = h(x) \tag{11}$$

where $g(\cdot)$ and $h(\cdot)$ are some functions. Note that the above applies only for *unbiased* estimates, so $E(\hat{t}\,) = t = E[h(x)]$. The minimum variance is also given by,

$$\operatorname{var}(\hat{t}\,) = \frac{1}{-E(\ddot{L})} = \frac{1}{g(t)} \implies g(t) = -E(\ddot{L}) \tag{12}$$

Note: $\hat{t}$ is an estimate of t. Hence, $\hat{t}$ cannot depend on t itself (if it does, such an estimate is useless!). So the result $\hat{t} = h(x)$ intuitively makes sense, because $\hat{t}$ depends only on the observed, given data x and not at all on t. **But** the mean and variance of $\hat{t}$ generally *do* depend on t and that is ok ! For the MVUE case, mean $E(\hat{t}\,) = t$ and variance $\operatorname{var}(\hat{t}) = g(t)$ – both are purely functions of t alone.

Replacing the scalar random variable X by a vector of random variables $\mathbf{x}$, the results still hold.

Facts

- Identity, if the regularity condition is satisfied, then

$$E\left(\dot{L}^2\right) = -E\left(\ddot{L}\right)$$

- Fisher information $I(t)$ for data $\mathbf{x}$ is defined by

$$I(t) = -E(\ddot{L})$$

  So, the minimum variance is the reciprocal of Fisher information. The "more the information", the lower is the CRLB.

- For a deterministic signal $s[n;t]$ with an unknown parameter t in zero-mean AWGN $w[n] \sim \mathcal{N}\left(0, \sigma^2\right)$,

$$x[n] = s[n;t] + w[n] \qquad n = 1, 2, \ldots, N$$

  the minimum variance (the CRLB, if it exists) is given by

$$\operatorname{var}\left(\hat{t}\,\right) \geq \frac{\sigma^2}{\sum_{n=0}^{N-1} \left(\frac{\partial}{\partial t} s[n;t]\right)^2} = \frac{\sigma^2}{\|\frac{\partial}{\partial t}\mathbf{s}\|^2}$$

- For an estimate $\hat{t}$ of t, if the CRLB is known, then for any transformation $\tau = g(t)$ for some function $g(\cdot)$ has the new CRLB

$$\text{CRLB}_\tau = \text{CRLB}_t \left(\frac{\partial}{\partial t} g(t) \right)^2$$

- The CRLB always increases as we estimate more parameters for same given data.

Let $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_M]^T$ be a vector parameter. Assume that an estimator $\hat{\boldsymbol{\theta}} = \left[\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_M\right]^T$ is unbiased, that is,

$$E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta} \iff E(\hat{\theta}_i) = \theta_i$$

The $M \times M$ *Fisher information matrix* $\mathbf{I}(\boldsymbol{\theta})$ is a matrix, whose $(i,j)^{th}$ element is given by

$$[\mathbf{I}(\boldsymbol{\theta})]_{i,j} = -E\left[\frac{\partial^2 \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right]$$

Note that $p(\mathbf{x}; \boldsymbol{\theta})$ is a scalar function, depending on vector parameters $\mathbf{x}$ and $\boldsymbol{\theta}$. For example, if $w[n]$ is i.i.d $\mathcal{N}\left(0, \sigma^2\right)$ and $x[n] = \theta_1 + n\theta_2 + w[n]$, then

$$p(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x[n] - \theta_1 - n\theta_2)^2 \right\}$$

Say $\mathbf{x} = [1,2,5,3], \boldsymbol{\theta} = [1,2], \sigma = 2$ implies $p(\mathbf{x}; \boldsymbol{\theta}) = 1.89 \times 10^{-3}$.

Note: The Fisher matrix is symmetric, because the partial derivatives do not depend on order of evaluation. If the *regularity condition*

$$E\left[\frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\mathbf{x}; \boldsymbol{\theta}) \right] = \mathbf{0} \qquad \forall\, \boldsymbol{\theta}$$

is satisfied (where the expectation is taken w.r.t $p(\mathbf{x}; \boldsymbol{\theta})$) then the covariance matrix of any unbiased estimator $\hat{\boldsymbol{\theta}}$ satisfies

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} - \mathbf{I}^{-1}(\boldsymbol{\theta}) \geq \mathbf{0} \iff \text{var}(\theta_i) \geq [\mathbf{I}^{-1}(\boldsymbol{\theta})]_{i,i}$$

Note: $[\mathbf{I}^{-1}(\boldsymbol{\theta})]_{i,i}$ means first you calculate the whole matrix inverse and then take the $(i,i)^{th}$ element. The covariance matrix of any vector $\mathbf{y}$ is given by

$$\begin{aligned}\boldsymbol{\mu}_{\mathbf{y}} &= E(\mathbf{y}) \\ \mathbf{C}_{\mathbf{y}} &= E\left[(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}})(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}})^T\right]\end{aligned}$$

Furthermore, an estimator that attains the lower bound,

$$\mathbf{C}_{\hat{\boldsymbol{\theta}}} = \mathbf{I}^{-1}(\boldsymbol{\theta}) \iff \frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\mathbf{x};\boldsymbol{\theta}) = \mathbf{I}(\boldsymbol{\theta})(\mathbf{g}(\mathbf{x})-\boldsymbol{\theta})$$

for some M-dimensional function $\mathbf{g}$ and some $M \times M$ matrix I. That estimator, which is the MVUE, is $\hat{\boldsymbol{\theta}} = \mathbf{g}(\mathbf{x})$, and its covariance matrix is $\mathbf{I}^{-1}(\boldsymbol{\theta})$.

3.2 Problem 3.1

If $x[n]$ for $n = 0, 1, \ldots, N-1$ are i.i.d. according to $\mathcal{U}(0, \theta)$, show that the regularity condition does not hold. That is,

$$E\left[\frac{\partial}{\partial \theta} \ln p(\mathbf{x};\theta)\right] \neq 0 \quad \forall\ \theta > 0$$

Solution By definition of the expectation operator,

$$E\left[\frac{\partial}{\partial \theta} \ln p(\mathbf{x};\theta)\right] = \int \left(\frac{\partial}{\partial \theta} \ln p(\mathbf{x};\theta)\right) p(\mathbf{x};\theta)\ d\mathbf{x} = \int \frac{\partial}{\partial \theta} p(\mathbf{x};\theta)\ d\mathbf{x} \tag{13}$$

follows from Eq(7). Denote the N random variables as $x_i = x[i-1]$ for $i = 1, 2, \ldots, N$. It is given in the problem that their PDFs are identical:

$$p(x_i\ ;\theta) = \begin{cases} 1/\theta & 0 < x_i \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

and

$$\int_0^\theta p(x_i\ ;\theta)\ dx_i = 1$$

The multiple integral in Eq(13) simplifies to product of integrals

$$\int \frac{\partial}{\partial \theta} p(\mathbf{x};\theta)\ d\mathbf{x} = \left(\int_0^\theta \frac{\partial}{\partial \theta} p(x_1\ ;\theta)\ dx_1\right) \cdots \left(\int_0^\theta \frac{\partial}{\partial \theta} p(x_N\ ;\theta)\ dx_N\right)$$

because the x_i's are independent. Note that the limits of the integral depend on θ, so we cannot interchange the order of differentiation and integration,

$$\int_0^\theta \frac{\partial}{\partial\theta} p(x_i\,;\theta)\,dx_i \neq \frac{\partial}{\partial\theta}\int_0^\theta p(x_i\,;\theta)\,dx_i$$

Hence, the regularity condition fails to hold. In fact, LHS= $1/\theta$, but RHS=0!

3.3 Problem 3.3

The data $x[n] = Ar^n + w[n]$ for $n = 0, 1, \cdots, N-1$ are observed, where $w[n]$ is WGN with variance σ^2 and $r > 0$ is known. Find the CRLB of A. Show that an efficient estimator exists and find its variance.

Solution Assuming that $x[i]$'s are statistically independent, the joint PDF is

$$\begin{aligned}
p(\mathbf{x}; A) &= \prod_{i=0}^{N-1} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x[n] - Ar^n)^2\right) \\
&= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - Ar^n)^2\right) \\
\Longrightarrow \ln p(\mathbf{x}; A) &= -\ln\left(2\pi\sigma^2\right)^{N/2} - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n] - Ar^n)^2 \\
\Longrightarrow \frac{\partial}{\partial A}\ln p(\mathbf{x}; A) &= \frac{1}{\sigma^2}\sum_{n=0}^{N-1} r^n\,(x[n] - Ar^n)
\end{aligned}$$

Since the sum

$$S = \sum_{n=0}^{N-1} r^{2n} = \begin{cases} \frac{r^{2N}-1}{r^2-1} & r \neq 1 \\ N & r = 1 \end{cases}$$

is deterministic and known (because both r and N are known), the above equation simplifies to

$$\frac{\partial}{\partial A}\ln p(\mathbf{x}; A) = \frac{1}{\sigma^2}\left(\sum_{n=0}^{N-1} r^n x[n] - AS\right) \tag{14}$$

$$\dot{L} = \frac{S}{\sigma^2}\left(\sum_{n=0}^{N-1}\frac{r^n}{S}x[n] - A\right) \tag{15}$$

$$= g(A)(h(\mathbf{x}) - A) \tag{16}$$

where $g(A) = S/\sigma^2$ is a constant (doesn't even depend on A!) and

$$h(\mathbf{x}) = \sum_{n=0}^{N-1} \frac{r^n}{S} x[n]$$

is depends on $\mathbf{x}$ but not on A. Hence, from Theorem 3.1, the MVUE estimate $\hat{A}$ is

$$\hat{A} = h(\mathbf{x}) = \frac{1}{S} \sum_{n=0}^{N-1} r^n x[n]$$

and the variance of $\hat{A}$ satisfies

$$\mathrm{var}(\hat{A}) \geq \frac{\sigma^2}{S} \qquad \text{and} \qquad \mathrm{CRLB} = \frac{1}{g(A)} = \frac{\sigma^2}{S}$$

We can also find the second derivative, from Eq(14),

$$\ddot{L} = \frac{\partial^2}{\partial A^2} \ln p(\mathbf{x}; A) = \frac{S}{\sigma^2}(0 - 1)$$

and, as required, CRLB $= -1/E[\ddot{L}]$ and, in our case, $E[\ddot{L}] = \ddot{L}$ because it is constant (does not depend on $\mathbf{x}$ or A).

3.4 Problem 3.5

If $x[n] = A + w[n]$ for $n = 1, 2, \ldots, N$ are observed, where $\mathbf{w} = [w[1], w[2], \ldots, w[N]]T \sim \mathcal{N}(0, \mathbf{C})$, find the CRLB for A. Does an efficient estimator exist and if so, what is its variance ?

Solution The joint p.d.f. of $\mathbf{x}$ is given by

$$\begin{aligned}
p(\mathbf{x}; A) &= \frac{1}{\sqrt{\det(2\pi\mathbf{C})}} \exp\left\{-\frac{1}{2}(\mathbf{x} - A\mathbf{e})^T \mathbf{C}^{-1} (\mathbf{x} - A\mathbf{e})\right\} \\
\Longrightarrow \ln p(\mathbf{x}; A) &= \ln \frac{1}{\sqrt{\det(2\pi\mathbf{C})}} - \frac{1}{2}(\mathbf{x} - A\mathbf{e})^T \mathbf{C}^{-1} (\mathbf{x} - A\mathbf{e}) \\
\Longrightarrow \frac{\partial}{\partial A} \ln p(\mathbf{x}; A) &= -\frac{1}{2} \frac{\partial}{\partial A} \left[(\mathbf{x} - A\mathbf{e})^T \mathbf{C}^{-1} (\mathbf{x} - A\mathbf{e})\right]
\end{aligned}$$

Using the result that

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{m}^T \mathbf{Q} \mathbf{m} = 2 \left(\frac{\partial}{\partial \boldsymbol{\theta}} \mathbf{m}^T\right) \mathbf{Q} \mathbf{m}$$

Setting $\mathbf{Q} = \mathbf{C}^{-1}$ and $\mathbf{m} = (\mathbf{x} - A\mathbf{e})$,

$$\frac{\partial}{\partial A}\mathbf{m}^T = \frac{\partial}{\partial A}(\mathbf{x} - A\mathbf{e})^T = (0 - \frac{\partial}{\partial A}A\mathbf{e}^T) = -\mathbf{e}^T$$

So

$$\frac{\partial}{\partial A}\ln p(\mathbf{x}; A) = \mathbf{e}^T\mathbf{C}^{-1}(\mathbf{x} - A\mathbf{e}) = (\mathbf{e}^T\mathbf{C}^{-1}\mathbf{x} - A\mathbf{e}^T\mathbf{C}^{-1}\mathbf{e})$$

The scalar $\mathbf{e}^T\mathbf{Q}\mathbf{e}$ is nothing but sum of all the elements of $\mathbf{Q}$ for any $\mathbf{Q}$. Consider, for example,

$$[1, 1, 1]\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \tag{17}$$

$$= [a + d + g, b + e + h, c + f + i]\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \tag{18}$$

$$= a + d + g + b + e + h + c + f + i \tag{19}$$

So, denoting $\alpha = \mathbf{e}^T\mathbf{C}^{-1}\mathbf{e}$,

$$\frac{\partial}{\partial A}\ln p(\mathbf{x}; A) = (\mathbf{e}^T\mathbf{C}^{-1}\mathbf{x} - A\mathbf{e}^T\mathbf{C}^{-1}\mathbf{e}) = \alpha\left(\frac{\mathbf{e}^T\mathbf{C}^{-1}\mathbf{x}}{\alpha} - A\right)$$

The above expression is clearly of the form

$$\frac{\partial}{\partial A}\ln p(\mathbf{x}; A) = g(A)(h(\mathbf{x}) - A)$$

Hence, there exists a MVUE (the efficient estimator) given by

$$\text{MVUE} = \hat{A} = h(\mathbf{x}) = \frac{\mathbf{e}^T\mathbf{C}^{-1}\mathbf{x}}{\alpha} = \frac{\mathbf{e}^T\mathbf{C}^{-1}\mathbf{x}}{\mathbf{e}^T\mathbf{C}^{-1}\mathbf{e}}$$

and its variance is

$$\text{var}(\hat{A}) = \frac{1}{\alpha} = \frac{1}{\sum_{i=1}^{N}\sum_{j=1}^{N}(\mathbf{C}^{-1})_{i,j}}$$

3.5 Problem 3.9

We observe two samples of a DC level in *correlated* Gaussian noise

$$x[0] = A + w[0]$$
$$x[1] = A + w[1]$$

where $\mathbf{w} = [w[0], w[1]]^T$ is zero mean with covariance matrix

$$\mathbf{C} = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

The parameter ρ is the cross-correlation coefficient between $w[0]$ and $w[1]$. Compute the CRLB of A and compare it to the case when $\rho = 0$ (WGN). Also explain what happens when $\rho = \pm 1$.

Solution: This is a special case of Problem 3.5 (see above) for $N = 2$. Since

$$\mathbf{C}^{-1} = \frac{1}{\sigma^2(\rho^2 - 1)} \begin{bmatrix} -1 & \rho \\ \rho & -1 \end{bmatrix}$$

the CRLB is

$$\operatorname{var} \hat{A} = \frac{1}{\mathbf{e}^T \mathbf{C}^{-1} \mathbf{e}} = \frac{\sigma^2(\rho^2 - 1)}{2(\rho - 1)}$$

When $\rho = 0$, $\operatorname{var} \hat{A} = \sigma^2/2$, as expected. But when $\rho \to \pm 1$, the matrix $\mathbf{C}$ becomes singular, hence its inverse does not exist; it means that the samples $w[0]$ and $w[1]$ are almost perfectly correlated and hence do not carried any additional information.

3.6 Problem 3.13

Consider polynomial curve fitting

$$x[n] = \sum_{k=0}^{p-1} A_k n^k + w[n]$$

for $n = 0, 1, \ldots, N-1$. $w[n]$ is i.i.d. WGN with variance σ^2. It is desired to estimate $\{A_0, A_1, \ldots, A_{p-1}\}$. Find the Fisher information matrix for this problem.

Solution: The joint p.d.f. is

$$\begin{aligned}
p(\mathbf{x};\mathbf{A}) &= \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left[x[n]-\sum_{k=0}^{p-1} A_k n^k\right]^2\right\} \\
&= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\left[x[n]-\sum_{k=0}^{p-1} A_k n^k\right]^2\right\} \\
\implies \ln p(\mathbf{x};\mathbf{A}) &= \ln\frac{1}{(2\pi\sigma^2)^{N/2}} - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\left[x[n]-\sum_{k=0}^{p-1} A_k n^k\right]^2 \\
\implies \frac{\partial}{\partial A_i}\ln p(\mathbf{x};\mathbf{A}) &= 0 - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}\left[2\left\{x[n]-\sum_{k=0}^{p-1} A_k n^k\right\}\left(0-n^i\right)\right]
\end{aligned}$$

Because

$$\begin{aligned}
\frac{\partial}{\partial A_i}\sum_{k=0}^{p-1} A_k n^k &= \frac{\partial}{\partial A_i}\left(A_1 n^1 + A_2 n^2 + \ldots + A_i n^i + \ldots + A_N n^N\right) \\
&= \left(0 + 0 + \ldots + \frac{\partial}{\partial A_i} A_i n^i + 0\right) \\
&= n^i
\end{aligned}$$

Hence, the simplification:

$$\begin{aligned}
\frac{\partial}{\partial A_i}\ln p(\mathbf{x};\mathbf{A}) &= \frac{1}{\sigma^2}\sum_{n=0}^{N-1} n^i \left\{x[n]-\sum_{k=0}^{p-1} A_k n^k\right\} \\
\implies \frac{\partial^2}{\partial A_j \partial A_i}\ln p(\mathbf{x};\mathbf{A}) &= \frac{1}{\sigma^2}\sum_{n=0}^{N-1} n^i \left(0-n^j\right) \\
&= -\frac{1}{\sigma^2}\sum_{n=0}^{N-1} n^{i+j}
\end{aligned}$$

Hence, by definition, $(i,j)^{\text{th}}$ entry of the the $p \times p$ Fisher information matrix $\mathbf{I(A)}$ is given by

$$[\mathbf{I(A)}]_{i,j} = -E\left[\frac{\partial^2}{\partial A_i \partial A_j}\ln p(\mathbf{x};\mathbf{A})\right] = \frac{1}{\sigma^2}\sum_{n=0}^{N-1} n^{i+j}$$

for $i,j = 0,1,\ldots,p-1$. Note that the Fisher information matrix is symmetric, so the order of evaluation of partial derivatives can be interchanged. See pg. 42, Eq (3.22) in the textbook for a special case of the above for $p = 2$. Note that for the $(0,0)^{th}$ entry of the matrix, the above expression gives

$$\sum_{n=0}^{N-1} n^{i+j} = \sum_{n=0}^{N-1} n^{0+0} = (0^0 + 1^0 + \ldots + (N-1)^0)$$

where 0^0 must be taken as 1 (even though some authors disagree).

4 Chapter 5

Neyman-Fisher Factorization Theorem If we can factor the p.d.f $p(\mathbf{x};\ th)$ as

$$p(\mathbf{x};\ th) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$$

where $g(\cdot)$ is a function depending on $\mathbf{x}$ only through $T(\mathbf{x})$ and $h(\cdot)$ is a function depending only on x, then $T(\mathbf{x})$ is a sufficient statistic for θ. The converse is also true.

4.1 Problem 5.2

The IID observations x_n for $n = 1, 2, \ldots, N$ have exponential p.d.f

$$p(x_n; \sigma^2) = \begin{cases} \frac{x_n}{\sigma^2} \exp(-x_n^2/2\sigma^2) & x_n > 0 \\ 0 & \text{otherwise} \end{cases}$$

Find a sufficient statistic for σ^2.

Solution Let $u(t)$ be the unit step function. The joint PDF of $x_1, x_2, \ldots, x_n$ is given by (because they are independent),

$$\begin{aligned} p(\mathbf{x}; \sigma^2) &= \prod_{n=1}^{N} p(x_n; \sigma^2) \\ &= \prod_{n=1}^{N} \frac{x_n}{\sigma^2} \exp(-x_n^2/2\sigma^2) u(x_n) \\ &= \left(\prod_{n=1}^{N} x_n u(x_n) \right) \left(\frac{1}{\sigma^2} \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^{N} x_n^2 \right) \right) \\ &= h(\mathbf{x}) g(T(\mathbf{x}), \sigma^2) \end{aligned}$$

whence, the sufficient statistic for σ^2 is $T(\mathbf{x})$

$$T(\mathbf{x}) = \sum_{n=1}^{N} x_n^2$$

4.2 Problem 5.5

The IID observations x_n for $n = 1, 2, \ldots, N$ are distributed according to $\mathcal{U}[-\theta, \theta]$, where $\theta > 0$. Find a sufficient statistic for θ.

Solution The individual sample p.d.f. is given by

$$p(x_n; \theta) = \begin{cases} 1/2\theta & -\theta < x_n < \theta \\ 0 & \text{otherwise} \end{cases}$$

The joint p.d.f is given by

$$\begin{aligned} p(\mathbf{x}; \theta) &= \prod_{n=1}^{N} p(x_n; \theta) \\ &= \begin{cases} 1/(2\theta)^N & -\theta < x_n < \theta,\, n \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Define a function bool(S) for any mathematical statement S such that

$$\text{bool}(S) = \begin{cases} 1 & S \text{ is true} \\ 0 & S \text{ is false} \end{cases}$$

(This is also called as Indicator function, see Wikipedia). Then

$$p(\mathbf{x}; \theta) = \frac{1}{(2\theta)^N} \text{bool}(-\theta < x_n < \theta,\, \forall \in \mathbb{N})$$

But,

$$\begin{aligned} x_n < \theta &\implies \theta > x_1 \text{ and } \theta > x_2 \cdots \text{ and } \theta > x_N \\ &\implies (\theta > x_1) \cap (\theta > x_2) \cap \cdots \cap (\theta > x_n) \\ &\implies \theta > \max\{x_1, x_2, \ldots, x_N\} \end{aligned}$$

Similarly,

$$\begin{aligned} -\theta < x_n &\implies \theta > -x_n \\ &\implies \theta > \max\{-x_1, -x_2, \ldots, -x_N\} \end{aligned}$$

Combining both of the above,

$$\begin{aligned} -\theta < x_n < \theta &\implies (-\theta < x_n) \cap (\theta > x_n) \\ &\implies (\theta > \max(-\mathbf{x})) \cap (\theta > \max(\mathbf{x})) \\ &\implies \theta > \max\{|x_1|, |x_2|, \dots, |x_N|\} \end{aligned}$$

So, the joind p.d.f. becomes

$$\begin{aligned} p(\mathbf{x}; \theta) &= \frac{1}{(2\theta)^N} \operatorname{bool}(\max\{|x_1|, |x_2|, \dots, |x_N|\} < \theta) \\ &= g(T(\mathbf{x}), \theta) h(\mathbf{x}) \end{aligned}$$

where $h(\mathbf{x}) = 1$ and

$$\begin{aligned} T(\mathbf{x}) &= \max\{|x_1|, |x_2|, \dots, |x_N|\} \\ g(T(\mathbf{x}), \theta) &= \frac{1}{(2\theta)^N} \operatorname{bool}(T(\mathbf{x}) < \theta) \end{aligned}$$

Hence, by Neyman-Fisher factorization theorem, $T(\mathbf{x})$, as given above, is the sufficient statistic. **Note:** The sample mean is *not* a sufficient statistic for uniform distribution!

5 Chapter 7: MLE

The MLE for a scalar parameter is defined as the value of parameter t that maximizes $p(\mathbf{x}; t)$ for a *given, fixed* $\mathbf{x}$, i.e., the value that maximizes the likelihood function. The maximization is performed over the allowable range of t.

To find the MLE, solve the equation

$$\frac{\partial}{\partial t} \ln p(\mathbf{x}; t) = 0$$

for t. This equation may have multiple solutions and you should choose the one appropriately.

Theorem. *If an efficient estimator (the estimator which attains CRLB) exists, then MLE procedure will find it.*

The MLE is

- asymptotically unbiased i.e., $E(\hat{t}\,) \to t$ as $N \to \infty$.

- asymptotically efficient i.e., $\mathrm{var}(\hat{t}\,) \to CRLB$ as $N \to \infty$.
- asymptotically optimal i.e., both of the above are true

Theorem. *If the pdf $p(\mathbf{x}; t)$ is twice differentiable and the Fisher information $I(t)$ is nonzero, then the MLE of the unknown parameter t is asymptotically distributed (for large N) according to*

$$\hat{t} \sim\sim \mathcal{N}\left(t, I^{-1}(t)\right)$$

i.e., Gaussian distributed with mean equal to true value t and variance equal to CRLB (= inverse of Fisher information).

Theorem. *Assume that the MLE $\hat{t}$ of unknown parameter t is known. Consider a transformation function of t,*

$$\tau = f(t)$$

for any function $f(\cdot)$. Then the MLE $\hat{\tau}$ of τ is nothing but

$$\hat{\tau} = f(\hat{t}\,)$$