

Sufficiency

Aleksandar Dogandžić

2017-05-16

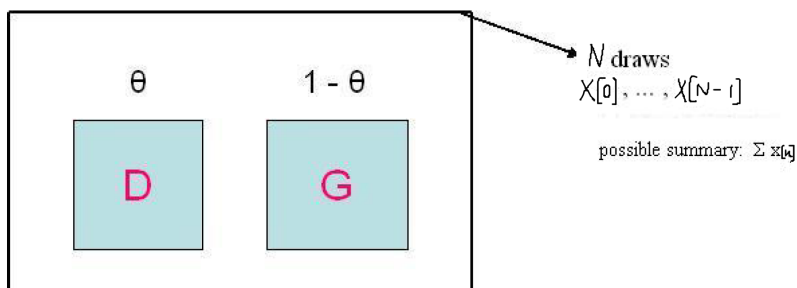
Contents

Binary source	1
Digital communications	6
Packet arrivals	8
Estimating a parameter of uniform distribution	9
Detection	9
Classification: Discrete likelihood ratios	10

READING: §5.3–5.4 in the textbook and (Hero 2015, §3.5).

A function $T(X)$ of the observations X only is called a statistic.

Binary source



A MACHINE produces N items in succession, with probability θ of producing a defective product. Suppose that there is no dependence in quality of the produced items.

$X[n]$ is a 0/1-valued variable with $\Pr\{X[n] = 1\} = \theta$ and $\Pr\{X[n] = 0\} = 1 - \theta$. Then, our statistical model is

$$\begin{aligned} p_{\mathbf{X}|\Theta}(\mathbf{x} | \theta) &= \prod_{n=0}^{N-1} \theta^{x[n]} (1 - \theta)^{1-x[n]} \\ &= \theta^{\sum_{n=0}^{N-1} x[n]} (1 - \theta)^{N - \sum_{n=0}^{N-1} x[n]} \end{aligned} \quad (1)$$

where $\mathbf{x} = (x[n])_{n=0}^{N-1}$.

* NOTE: $K = K(X) \triangleq \sum_{n=0}^{N-1} X[n]$ is a binomial random variable taking values in $\{0, 1, \dots, N\}$:

$$p_{K|\Theta}(k|\theta) = \text{Bin}(k|N, \theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}. \quad (2)$$

See the table of distributions.

The joint probability mass function (pmf) of X and K is

$$\begin{aligned} p_{X,K|\Theta}(\mathbf{x}, k|\theta) &= \Pr\{(X[n])_{n=0}^{N-1} = (x[n])_{n=0}^{N-1}, K = k\} \\ &= \begin{cases} p_{X|\Theta}(\mathbf{x}|\theta), & \text{if } k = \sum_{n=0}^{N-1} x[n] \\ 0, & \text{otherwise} \end{cases} \\ &= p_{X|\Theta}(\mathbf{x}|\theta) \mathbb{1}\left(k = \sum_{n=0}^{N-1} x[n]\right) \\ &= \theta^k (1-\theta)^{N-k} \mathbb{1}\left(k = \sum_{n=0}^{N-1} x[n]\right). \end{aligned} \quad \text{see (1)}$$

Therefore,

$$\begin{aligned} p_{X|K,\Theta}(\mathbf{x}|k, \theta) &= \frac{p_{X,K|\Theta}(\mathbf{x}, k|\theta)}{p_{K|\Theta}(k|\theta)} \\ &= \frac{\theta^k (1-\theta)^{N-k} \mathbb{1}\left(k = \sum_{n=0}^{N-1} x[n]\right)}{\binom{N}{k} \theta^k (1-\theta)^{N-k}} \\ &= \frac{1}{\binom{N}{k}} \mathbb{1}\left(k = \sum_{n=0}^{N-1} x[n]\right) \end{aligned} \quad \text{see (2)}$$


i.e., X given $\sum_{n=0}^{N-1} X[n] = k$ is uniformly distributed over the $\binom{N}{k}$ sequences that have exactly k ones. Consequently,

- the conditional distribution of X given $\sum_{n=0}^{N-1} X[n] = k$ is *not* a function of θ and
- $\sum_{n=0}^{N-1} X[n]$ carries all relevant information about θ .

Is there a loss of information by keeping and recording only $\sum_{n=0}^{N-1} x[n]$ in the above example?

Yes we are dropping a lot of information. But

No in terms of inference about θ .

 NOTE: $\sum_{n=0}^{N-1} x[n]$ compresses $\{0, 1\}^N$ (N bits) to $\{0, 1, \dots, N\}$ ($\log N$ bits).

* WE wish to separate out any aspects of the data that are irrelevant in the context of our model, i.e., reduce the data and deal only with the statistics whose use involves no loss of information. For example, we could save memory and store only the reduced data. What we mean by “no loss of information” is quantified in the following definition.

Definition 1. $T = T(X)$ is a sufficient statistic for θ if the conditional distribution of X given $T(X)$ does not involve θ :

$$p_{X|T(X),\Theta}(x | T(x) = t, \theta).$$

not a function of θ

Think of sufficient statistics as not throwing away information about θ .

☞ TRIVIAL sufficient statistic. Keeping all measurements is always sufficient: $T(X) = X$.

In general, checking sufficiency directly is difficult because we need to compute conditional distributions. Fortunately, the following theorem has conditions that are easy to verify.

Theorem 1 (Factorization Theorem). A statistic $T(X)$ is sufficient for θ if and only if there exists functions $g(t, \theta)$ and $h(x)$ such that

$$f_{X|\Theta}(x | \theta) = \underbrace{g(T(x), \theta)}_{\substack{\text{parameters} \\ \text{coupled with} \\ \text{sufficient} \\ \text{statistics}}} \underbrace{h(x)}_{\substack{\text{does not} \\ \text{contain} \\ \text{parameters}}}.$$

☞ NOTE: $T(x)$ must be a *statistic*, a function of data x only.

Proof: To illustrate the idea of the proof and for simplicity, we concentrate on the discrete case. Suppose that $T(X)$ is a sufficient statistic. Then, we have

$$\begin{aligned} p_{X|\Theta}(x | \theta) &= \Pr_X\{X = x\} \\ &= \Pr_X\{X = x, T(X) = T(x)\} && \{X = x\} \subseteq \{T(X) = T(x)\}. \\ &= \Pr_X\{T(X) = T(x)\} \underbrace{\Pr_X\{X = x | T(X) = T(x)\}}_{h(x), \text{ by sufficiency}} \\ &= g(T(x), \theta)h(x). \end{aligned}$$

Conversely,

$$\begin{aligned} \Pr_X\{X = x | T(X) = T(x)\} &= \frac{\Pr_X\{X = x, T(X) = T(x)\}}{\Pr_X\{T(X) = T(x)\}} && \{X = x\} \subseteq \{T(X) = T(x)\}. \\ &= \frac{\underbrace{\Pr_X\{X = x\}}_{p_{X|\Theta}(x|\theta)}}{\Pr_X\{T(X) = T(x)\}} && \\ &= \frac{p_{X|\Theta}(x|\theta)}{\sum_{y: T(y)=T(x)} p_{X|\Theta}(y|\theta)} && \Pr_X\{T(X) = T(x)\} = \sum_{y: T(y)=T(x)} p_{X|\Theta}(y|\theta) \\ &= \frac{g(T(x), \theta)h(x)}{\sum_{y: T(y)=T(x)} \underbrace{g(T(y), \theta)h(y)}_{\text{by the assumption}}} && \end{aligned}$$

$$\begin{aligned}
&= \frac{g(T(x), \theta) h(x)}{g(T(x), \theta) \sum_{y: T(y)=T(x)} h(y)} \\
&= \frac{h(x)}{\sum_{y: T(y)=T(x)} h(y)}
\end{aligned}$$

which is *not* a function of θ . □

* EXAMPLE (binary source).

$$p_{X|\Theta}(\mathbf{x} | \theta) = \theta^{\sum_{n=0}^{N-1} x[n]} (1 - \theta)^{N - \sum_{n=0}^{N-1} x[n]}.$$

Hence, $\sum_{n=0}^{N-1} x[n]$ is a sufficient statistic for θ .

* EXAMPLE: Mean and variance of scalar independent, identically distributed (i.i.d.) Gaussian measurements. Conditional on a and σ^2 , $(X[n])_{n=0}^{N-1}$ are i.i.d. $\mathcal{N}(a, \sigma^2)$. Define $\boldsymbol{\theta} = (a, \sigma^2)$ and $\mathbf{X} = (X[n])_{n=0}^{N-1}$. Then, we have

$$\begin{aligned}
f_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x} | \boldsymbol{\theta}) &= (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - a)^2\right\} \\
&= (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{Na^2}{2\sigma^2}\right) \\
&\quad \cdot \exp\left\{-\frac{1}{2\sigma^2} \left(\sum_{n=0}^{N-1} x^2[n] - 2a \sum_{n=0}^{N-1} x[n]\right)\right\}. \quad (3)
\end{aligned}$$

Applying the factorization theorem leads to the sufficient statistics for $\boldsymbol{\theta}$:

$$\mathbf{T}_1(\mathbf{x}) = \begin{bmatrix} \sum_{n=0}^{N-1} x[n] \\ \sum_{n=0}^{N-1} x^2[n] \end{bmatrix}.$$

Here, $h(\mathbf{x})$ is trivial: $h(\mathbf{x}) = 1$.

Define the sample mean

$$\bar{x} = \frac{1}{N} \sum_{n=0}^{N-1} x[n].$$

A frequently used equivalent sufficient statistic

$$\mathbf{T}_2(\mathbf{x}) = \begin{bmatrix} \bar{x} \\ \frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \bar{x})^2 \end{bmatrix}$$

can be obtained by suitably arranging the terms in the expression for $f_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x} | \boldsymbol{\theta})$:

$$f_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x} | \boldsymbol{\theta}) = (2\pi\sigma^2)^{-N/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} [(x[n] - \bar{x}) + (\bar{x} - a)]^2\right\}$$

and expanding the squares in the exponent:

$$f_{\mathbf{X}|\boldsymbol{\Theta}}(\mathbf{x} | \boldsymbol{\theta}) = (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{N}{2\sigma^2} \left\{ \frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \bar{x})^2 \right\} - \frac{N}{2\sigma^2} (\bar{x} - a)^2 \right].$$

- * **EXAMPLE.** Mean of scalar i.i.d. Gaussian measurements. Suppose $(X[n])_{n=0}^{N-1}$ are i.i.d. $\mathcal{N}(a, 1)$ given a . Then, using (3) with $\sigma^2 = 1$, we obtain

$$f_{\mathbf{X}|A}(\mathbf{x} | a) = \overbrace{\exp[Na(\bar{x} - 0.5a)] (2\pi)^{-0.5N} \exp\left(-0.5 \sum_{n=0}^{N-1} x^2[n]\right)}^{h(\mathbf{x})}$$

and, by the factorization theorem, \bar{x} is a sufficient statistic for the parameter a .

- * **EXAMPLE:** Mean vector and covariance matrix of vector i.i.d. Gaussian measurements. Conditional on \mathbf{a} and Σ , $(\mathbf{X}[n])_{n=0}^{N-1}$ are i.i.d. $d \times 1$ vectors $\mathcal{N}(\mathbf{a}, \Sigma)$. Define $\boldsymbol{\theta} = (\mathbf{a}, \Sigma)$. Then, we have

$$\begin{aligned} f_{(\mathbf{X}[n])_{n=0}^{N-1} | \boldsymbol{\theta}}((\mathbf{x}[n])_{n=0}^{N-1} | \boldsymbol{\theta}) \\ &= \prod_{n=0}^{N-1} \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp[-0.5(\mathbf{x}[n] - \mathbf{a})^\top \Sigma^{-1}(\mathbf{x}[n] - \mathbf{a})] \\ &= [\det(2\pi\Sigma)]^{-N/2} \exp\left\{-0.5 \sum_{n=0}^{N-1} (\mathbf{x}[n] - \mathbf{a})^\top \Sigma^{-1}(\mathbf{x}[n] - \mathbf{a})\right\}. \end{aligned} \quad (4)$$

Define the sample mean vector and covariance matrix

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{x}[n], \quad S(\mathbf{x}) = \frac{1}{N} \sum_{n=0}^{N-1} (\mathbf{x}[n] - \bar{\mathbf{x}})(\mathbf{x}[n] - \bar{\mathbf{x}})^\top.$$

Focus on the sum in the exponent term in (4):

$$\begin{aligned} \sum_{n=0}^{N-1} (\mathbf{x}[n] - \mathbf{a})^\top \Sigma^{-1}(\mathbf{x}[n] - \mathbf{a}) &= \sum_{n=0}^{N-1} (\mathbf{x}[n] - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \mathbf{a})^\top \Sigma^{-1}(\mathbf{x}[n] - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \mathbf{a}) \\ &= \sum_{n=0}^{N-1} (\mathbf{x}[n] - \bar{\mathbf{x}})^\top \Sigma^{-1}(\mathbf{x}[n] - \bar{\mathbf{x}}) \\ &\quad + 2 \underbrace{\sum_{n=0}^{N-1} (\mathbf{x}[n] - \bar{\mathbf{x}})^\top \Sigma^{-1}(\bar{\mathbf{x}} - \mathbf{a})}_{0} \\ &\quad + \sum_{n=0}^{N-1} (\bar{\mathbf{x}} - \mathbf{a})^\top \Sigma^{-1}(\bar{\mathbf{x}} - \mathbf{a}) \\ &= N \operatorname{tr}[\Sigma^{-1} S(\mathbf{x})] + N(\bar{\mathbf{x}} - \mathbf{a})^\top \Sigma^{-1}(\bar{\mathbf{x}} - \mathbf{a}). \end{aligned}$$

Hence,

$$\begin{aligned} f_{(\mathbf{X}[n])_{n=0}^{N-1} | \boldsymbol{\theta}}((\mathbf{X}[n])_{n=0}^{N-1} | \boldsymbol{\theta}) &= (2\pi)^{-Nd/2} (\det \Sigma)^{-N/2} \exp\{-0.5N \operatorname{tr}[\Sigma^{-1} S(\mathbf{x})]\} \\ &\quad \cdot \exp[-0.5N(\bar{\mathbf{x}} - \mathbf{a})^\top \Sigma^{-1}(\bar{\mathbf{x}} - \mathbf{a})] \end{aligned}$$

and, consequently, $\bar{\mathbf{x}}$ and $S(\mathbf{x})$ are the sufficient statistics for $\boldsymbol{\theta}$.

* EXAMPLE. Rank-ordered sample. $X_{(1)}, \dots, X_{(N)}$ is sufficient when $(X_i)_{i=1}^N$ are i.i.d..

Proof:

$$\begin{aligned} f(x_1, \dots, x_N | \theta) &= \prod_{i=1}^N f(x_i | \theta) \\ &= \prod_{i=1}^N f(x_{(i)} | \theta). \end{aligned} \quad (X_i)_{i=1}^N \text{ are i.i.d.}$$

□

Digital communications

CONSIDER the following signal-plus-noise measurement model:

$$x(t) = \underbrace{s(t)}_{\text{signal}} + \underbrace{w(t)}_{\text{noise}}$$

where the signal $s(t)$ is usually represented using orthonormal basis functions $\varphi_k(t)$:

$$s(t) = \sum_{k=1}^K \alpha_k \varphi_k(t).$$

NOTE: The signal $s(t)$ is unknown, but it has known structure, incorporated in this basis-function expansion. *We wish to use this structure for data reduction.*

If $\varphi_k(t)$ are orthonormal, α_k can be computed as:

$$\alpha_k = \int s(t) \varphi_k(t) dt.$$

Here, our goal at the receiver is to decide which $s(t)$ (α_k 's) has been transmitted.

In communication receivers, the received data $x(t)$ are *matched* to the basis functions, i.e.,

$$\hat{\alpha}_k = \int x(t) \varphi_k(t) dt \quad (5) \quad k = 1, 2, \dots, K$$

are computed and utilized for demodulation.

QUESTION: Are the $\hat{\alpha}_k$ s sufficient statistics for deciding about $s(t)$ or, more precisely, for inference about the α_k s?

📖 NOTE: In some applications, sampled data $(x[n])_{n=0}^{N-1}$ are available and

$$(\hat{\alpha}_k)_{k=1}^K = \hat{\alpha}_k(x) = \sum_{n=0}^{N-1} x[n] \varphi_k[n]$$

are used to approximate (up to a scaling factor) the integrals in (5). We focus on this scenario, having in mind that we can easily switch from sums to integrals by letting the sampling interval go to zero and N to infinity. Clearly, N is much larger than the number of basis functions K , i.e.,

$$K \ll N.$$

In the sampled-data case, our model is

$$X[n] = \underbrace{s[n]}_{\text{signal}} + \underbrace{W[n]}_{\text{noise}}$$

where

$$s[n] = \sum_{k=1}^K \alpha_k \varphi_k[n].$$

Define

$$\mathbf{X} = \begin{bmatrix} X[0] \\ X[1] \\ \vdots \\ X[N-1] \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} W[0] \\ W[1] \\ \vdots \\ W[N-1] \end{bmatrix}$$

and

$$\boldsymbol{\mu}(\boldsymbol{\alpha}) = \begin{bmatrix} s[0] \\ s[1] \\ \vdots \\ s[N-1] \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_K \end{bmatrix}$$

implying

$$\mathbf{X} = \boldsymbol{\mu}(\boldsymbol{\alpha}) + \mathbf{W}.$$

If the noise is additive zero-mean Gaussian with known covariance matrix

$$\mathbf{C} = \mathbb{E}_{\mathbf{W}}(\mathbf{W}\mathbf{W}^\top)$$

then

$$\{\mathbf{X} \mid \boldsymbol{\alpha}\} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\alpha}), \mathbf{C})$$

and

$$\boldsymbol{\mu}(\boldsymbol{\alpha}) = \begin{bmatrix} s[0] \\ s[1] \\ \vdots \\ s[N-1] \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^K \alpha_k \varphi_k[0] \\ \sum_{k=1}^K \alpha_k \varphi_k[1] \\ \vdots \\ \sum_{k=1}^K \alpha_k \varphi_k[N-1] \end{bmatrix} = \mathbf{F}\boldsymbol{\alpha}$$

where

$$\mathbf{F} = \begin{bmatrix} \varphi_1[0] & \varphi_2[0] & \cdots & \varphi_K[0] \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1[N-1] & \varphi_2[N-1] & \cdots & \varphi_K[N-1] \end{bmatrix}$$

is an $N \times K$ matrix of basis functions. Hence,

$$f_{\mathbf{X}|\boldsymbol{\alpha}}(\mathbf{x}|\boldsymbol{\alpha}) = \frac{1}{\sqrt{\det(2\pi\mathbf{C})}} \exp[-0.5(\mathbf{x} - \mathbf{F}\boldsymbol{\alpha})^\top \mathbf{C}^{-1}(\mathbf{x} - \mathbf{F}\boldsymbol{\alpha})]. \quad (6)$$

What are the sufficient statistics for inference on $\boldsymbol{\alpha}$? Since \mathbf{C} is *known*, then the vector of sufficient statistics for $\boldsymbol{\alpha}$ is¹

$$\mathbf{F}^\top \mathbf{C}^{-1} \mathbf{x} \quad (7)$$

which is a $K \times 1$ vector. Since $K \ll N$, (7) achieves dimensionality reduction compared with the raw data \mathbf{x} .

For white noise ($\mathbf{C} = \sigma^2 \mathbf{I}$) with known variance σ^2 , (7) simplifies to (up to a known proportionality factor):

$$\mathbf{F}^\top \mathbf{x} = \begin{bmatrix} \sum_{n=0}^{N-1} \varphi_1[n]x[n] \\ \sum_{n=0}^{N-1} \varphi_2[n]x[n] \\ \vdots \\ \sum_{n=0}^{N-1} \varphi_K[n]x[n] \end{bmatrix} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_K \end{bmatrix}.$$

If σ^2 is *unknown*, then

$$f_{\mathbf{X}|\boldsymbol{\alpha}, \sigma^2}(\mathbf{x}|\boldsymbol{\alpha}, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^N}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{x} - \mathbf{F}\boldsymbol{\alpha})^\top (\mathbf{x} - \mathbf{F}\boldsymbol{\alpha})\right]$$

where $\boldsymbol{\alpha}$ and σ^2 are parameters and \mathbf{x} is the measurement vector.

Now

$$\mathbf{x}^\top \mathbf{x} = \sum_{n=0}^{N-1} x^2[n] \quad \text{and} \quad \mathbf{F}^\top \mathbf{x} = [\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_K]^\top$$

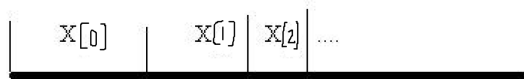
are jointly sufficient for $\boldsymbol{\alpha}$ and σ^2 .

Packet arrivals

SUPPOSE that elements of

$$\mathbf{X} = [X[0], X[1], \dots, X[N-1]]^\top$$

are i.i.d. inter-arrival times of packets arriving at a node in a communication network.



If \mathbf{C} is *unknown*, we cannot separate out any non-trivial sufficient statistics for both $\boldsymbol{\alpha}$ and \mathbf{C} .

¹ Expand (6) and use identities $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$ and $(\mathbf{C}^{-1})^\top = (\mathbf{C}^\top)^{-1}$.

Figure 1: Packet arrivals.

Conditional on θ , $(X[n])_{n=0}^{N-1}$ come from $\text{Expon}(\theta)$ distribution:

$$\begin{aligned} f_{\mathbf{X}|\Theta}(\mathbf{x} | \theta) &= \prod_{n=0}^{N-1} \theta e^{-\theta x[n]} \mathbb{1}_{[0,\infty)}(x[n]) \\ &= \theta^N \exp\left(-\theta \underbrace{\sum_{n=0}^{N-1} x[n]}_{T(\mathbf{x})}\right) \mathbb{1}_{[0,\infty)}(\min_n x[n]) \end{aligned}$$

where $\mathbb{1}_A(x) = \begin{cases} 1, & x \in A, \\ 0, & \text{otherwise} \end{cases}$ denotes an indicator function. Note

that

$$\prod_{n=0}^{N-1} \mathbb{1}_{[0,+\infty)}(x[n]) = \mathbb{1}_{[0,+\infty)}(\min_n x[n])$$

because

$$x[n] \geq 0 \quad \forall n \quad \Longleftrightarrow \quad \min_n x[n] \geq 0.$$

Estimating a parameter of uniform distribution

CONDITIONAL on θ , $X = (X[n])_{n=0}^{N-1}$ are i.i.d. $U(0, \theta)$:

$$\begin{aligned} f_{\mathbf{X}|\Theta}(\mathbf{x} | \theta) &= \prod_{n=0}^{N-1} \left[\frac{1}{\theta} \mathbb{1}_{[0,\theta]}(x[n]) \right] \\ &= \frac{1}{\theta^N} \prod_{n=0}^{N-1} \mathbb{1}_{[0,\theta]}(x[n]) \\ &= \frac{1}{\theta^N} \underbrace{\mathbb{1}_{(-\infty, \theta]}(\max_n x[n])}_{g(T(\mathbf{x}), \theta)} \underbrace{\mathbb{1}_{[0,+\infty)}(\min_n x[n])}_{h(\mathbf{x})} \end{aligned}$$

$U(a, b)$ stands for uniform probability density function (pdf) between a and b , see the table of distributions.

Here, we have used the facts that

$$\begin{aligned} (x[n])_{n=0}^{N-1} \leq \theta &\Longleftrightarrow \max_n x[n] \leq \theta \\ (x[n])_{n=0}^{N-1} \geq 0 &\Longleftrightarrow \min_n x[n] \geq 0. \end{aligned}$$

Detection

DETECTION problem: $\theta \in \{0, 1\}$ and

$$\begin{aligned} f_{\mathbf{X}|\Theta}(\mathbf{x} | \theta) &= \theta f_{\mathbf{X}|\Theta}(\mathbf{x} | 1) + (1 - \theta) f_{\mathbf{X}|\Theta}(\mathbf{x} | 0) \\ &= \left[\theta \underbrace{\frac{f_{\mathbf{X}|\Theta}(\mathbf{x} | 1)}{f_{\mathbf{X}|\Theta}(\mathbf{x} | 0)}}_{g(T(\mathbf{x}), \theta)} + (1 - \theta) \right] \underbrace{f_{\mathbf{X}|\Theta}(\mathbf{x} | 0)}_{h(\mathbf{x})} \end{aligned}$$

The likelihood ratio $T(\mathbf{x})$ is sufficient for θ : it is one-dimensional regardless of the nature of $f_{\mathbf{X}|\Theta}(\mathbf{x} | \theta)$. See also (Hero 2015, Ex. 3 in §3.5.2).

Classification: Discrete likelihood ratios

SUPPOSE the parameter θ belongs to the discrete parameter space

$$\text{sp}_{\Theta} = \{0, 1, \dots, M-1\}.$$

Then,

$$\begin{aligned} \mathbf{T}(\mathbf{X}) &= \left[\frac{f_{\mathbf{X}|1}(\mathbf{X} | 1)}{f_{\mathbf{X}|0}(\mathbf{X} | 0)}, \frac{f_{\mathbf{X}|2}(\mathbf{X} | 2)}{f_{\mathbf{X}|0}(\mathbf{X} | 0)}, \dots, \frac{f_{\mathbf{X}|M-1}(\mathbf{X} | M-1)}{f_{\mathbf{X}|0}(\mathbf{X} | 0)} \right]^{\top} \\ &= [\Lambda_1(\mathbf{X}), \dots, \Lambda_{M-1}(\mathbf{X})]^{\top} \end{aligned}$$

is sufficient for θ provided that $\mathbf{T}(\mathbf{X})$ is finite for all θ . An equivalent way to express this vector is as the sequence

$$(\Lambda_{\theta}(\mathbf{X}))_{\theta \in \text{sp}_{\Theta}} = (\Lambda_1(\mathbf{X}), \dots, \Lambda_{M-1}(\mathbf{X}))$$

called the *likelihood trajectory* over θ .

Indeed, define

$$\mathbf{u}_{\theta} = \mathbf{e}_k \quad \text{for } \theta = k$$

where $\mathbf{e}_k = [0, \dots, 0, 1, 0, \dots, 0]^{\top}$ is the k -th column of the $(p-1) \times (p-1)$ identity matrix. Now,

$$f_{\mathbf{X}|\Theta}(\mathbf{x} | \theta) = \underbrace{\mathbf{u}_{\theta}^{\top} \mathbf{T}(\mathbf{x})}_{g(\mathbf{T}(\mathbf{x}), \theta)} \underbrace{f_{\mathbf{X}|0}(\mathbf{X} | 0)}_{h(\mathbf{x})}.$$

Definition 2. *The statistic $T(\mathbf{x})$ is minimally sufficient if it is sufficient and provides a reduction of data greater than or equal to the data reduction achieved by any other sufficient statistic $S(\mathbf{x})$.*

References

Hero, Alfred O. (2015). *Statistical Methods for Signal Processing*. Lecture notes. Univ. Michigan, Ann Arbor, MI.