Fisher information and Cramer-Rao bound

Contents

1 Score and Fisher information

Let X be a random variable (or a random vector) with pdf $f(x;\theta)$ indexed by an unknown parameter $\theta \in \Theta \subset \mathbb{R}$. The question of interest here is: how much information about θ can be obtained from observing X?

To explore this, assume that f is differentiable with respect to θ, and define the **score function** associated with f as

$$S(\theta, x) := \partial_\theta \log f(x;\theta) = \frac{\partial_\theta f(x;\theta)}{f(x;\theta)}.$$

Observe that, by definition, for any fixed $\theta_0 \in \Theta$,

$$S(\theta_0, x) = \lim_{\epsilon \to 0} \frac{\frac{1}{\epsilon}\left[f(x;\theta_0+\epsilon) - f(x;\theta_0)\right]}{f(x;\theta_0)}.$$

Thus, the score $S(\theta_0, x)$ can be interpreted as the *relative instantaneous rate of change* of the function $\theta \mapsto f(x;\theta)$ at the point θ_0. In particular, if f changes rapidly in a neighborhood of θ_0, the score will have a large absolute value; conversely, if f is relatively flat, the score will be small. In other words, a large value of $|S(\theta_0, x)|$ indicates that the observation x is highly informative for distinguishing θ_0 from nearby parameter values.

A remarkable property of the score function is given in the following proposition.

Proposition 1.1. *Suppose that*

(I) the support of f, i.e. the set $\{x : \ f(x;\theta) > 0\}$, does not depend on θ, and

(II) the operations of integration (or summation) and differentiation by θ can be interchanged in $\int f(x;\theta)dx$. Thus, $\partial_\theta \int f(x;\theta)dx = \int \partial_\theta f(x;\theta)dx$[1].

Then

$$E_\theta S(\theta, X) = 0, \ \forall \theta \in \Theta.$$

The expected score is zero because, when the parameter is set to its correct value, a tiny change in that parameter makes the density increase for some values of x and decreases for others. These opposing changes must perfectly balance out, because the total probability — the area under the density curve — must always sum to 1. Consequently, the average effect of that small parameter shift, weighted by the data probabilities, is zero, which is precisely what the expected score expresses.

Attention. *From now on, unless explicitly stated otherwise or unless it is evident that they fail to hold, we will assume that conditions (I) and (II) are satisfied. Notably, these conditions are satisfied by the (regular) exponential family.*

[1]For more details on this condition, see the discussion of the Leibniz integral rule here

Taking the square (of S) and averaging we obtain $I_X(\theta)$:

$$I_X(\theta) := E_\theta \left[S^2(\theta, X)\right] = Var_\theta \left[S(\theta, X)\right]$$

which is known as the (expected) ***Fisher information*** (**FI**) that X contains about θ, or the FI for θ based on X.

FI attempts to *quantify the average sensitivity of the random variable X to the value of the parameter θ*. If small changes in θ result in large changes in the values of X, then observing the latter can tell us a lot about θ. In this case the FI would be quite large. In other words, FI attempts to quantify how easy one can guess the θ that produced the observed X.

Attention. Note that the subscript X in $I_X(\theta)$ serves solely as a label indicating that the FI is computed with respect to the random variable X; it does **not** imply that the Fisher information itself is random. In contrast, the score $S(\theta, X)$ *is* a random variable, since it is defined as a function of X. *In what follows, whenever no ambiguity about the underlying random variable arises, we will simply write $I(\theta)$ instead of $I_X(\theta)$.*

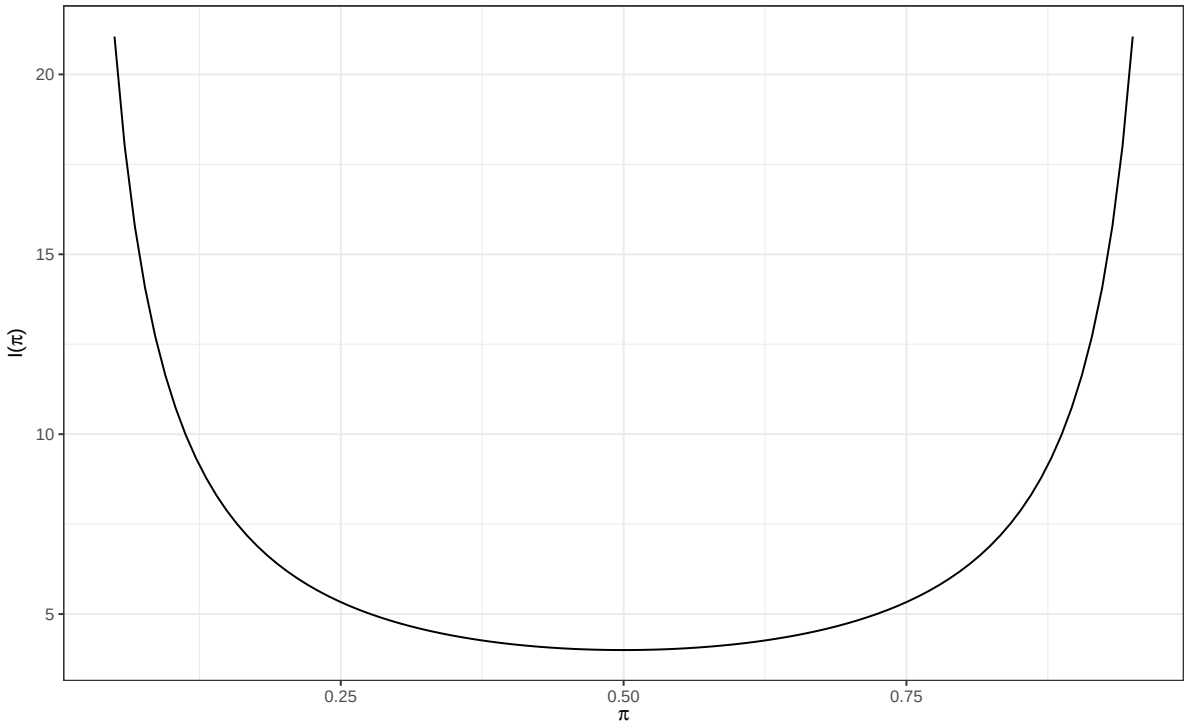
Example 1.1 (Calculating Fisher Information 1)**.**

- Bernoulli distribution: $X \sim Be(\pi)$ with $\pi \in (0,1)$. $f(x;\pi) = \pi^x(1-\pi)^{1-x}$, $x = 0,1$.

$$S(\pi, X) = \partial_\pi \{X \log(\pi) + (1-X)\log(1-\pi)\} = \frac{X-\pi}{\pi(1-\pi)}.$$

$$\implies I(\pi) = E_\pi \left[S^2(\pi, X)\right] = \frac{1}{\pi(1-\pi)}.$$

Here, the Fisher information turns out to be the reciprocal of the variance of a Bernoulli. This is not unusual. In fact, as we will see later, the Fisher information is typically inversely proportional to the variance. Intuitively, when the data exhibit greater variability, it becomes harder to infer the true value of the parameter, and the Fisher information decreases. Conversely, when the variance is small, the data are more concentrated and provide more precise information about the parameter.



- Normal distribution: $X \sim N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. $f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.

– For μ :

$$S(\mu, X) = \partial_\mu \left\{ -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(X-\mu)^2 \right\} = \frac{X-\mu}{\sigma^2}.$$
$$\implies I(\mu) = E\left[S^2(\mu, X)\right] = \frac{1}{\sigma^2}.$$

Note that, in this case, the FI about μ does not depend on μ but only on σ^2. It decreases with σ^2.

– For σ^2 :

$$S(\sigma^2, X) = \partial_{\sigma^2} \left\{ -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(X-\mu)^2 \right\} = -\frac{1}{2\sigma^2} + \frac{1}{2}\frac{(X-\mu)^2}{\sigma^4}.$$
$$\implies I(\sigma^2) = E\left[S^2(\sigma^2, X)\right] = \frac{1}{4}E\left(\frac{(X-\mu)^2}{\sigma^4} - \frac{1}{\sigma^2}\right)^2$$
$$= \frac{1}{4}\left(\frac{1}{\sigma^8}E(X-\mu)^4 + \frac{1}{\sigma^4} - \frac{2}{\sigma^6}E(X-\mu)^2\right) = \frac{1}{2\sigma^4},$$

where we have used the fact that, for $X \sim N(\mu, \sigma^2)$, $E(X-\mu)^4 = 3\sigma^4$. □

The following proposition offers a useful alternative for calculating the FI. Rather than requiring the second moment of the score, it expresses $I(\theta)$ as the negative expected derivative of the score function itself.

Proposition 1.2. *Let $X \sim f(x;\theta)$. Assume that $f(x;\theta)$ is twice differentiable with respect to θ, and that double differentiation and integration (or summation) can be interchanged, so that $\partial_\theta^2 \int f(x;\theta)dx = \int \partial_\theta^2 f(x;\theta)dx$. Then,*

$$I(\theta) = -E_\theta[\partial_\theta S(\theta, X)].$$

Note that the equality above is equivalent to $I(\theta) = -E_\theta\left[\partial_\theta^2 \log f(X;\theta)\right]$.

To see why this proposition is true, consider the following calculation where we have simplified the notation by writing S and f instead of $S(\theta, X)$ and $f(X;\theta)$, respectively.

$$\partial_\theta S = \partial_\theta \frac{\partial_\theta f}{f} = \frac{\partial_\theta^2 f \times f - (\partial_\theta f)^2}{f^2} = \frac{\partial_\theta^2 f}{f} - S^2.$$

But $E_\theta\left(\frac{\partial_\theta^2 f(X;\theta)}{f(X;\theta)}\right) = \int \partial_\theta^2 f(x;\theta)dx = \partial_\theta^2 \int f(x;\theta)dx = 0$. Thus, $E_\theta(\partial_\theta S) = -E_\theta S^2$.

Example 1.2 (Calculating Fisher Information 2)**.**

- Bernoulli distribution: $X \sim Be(\pi)$ with $\pi \in (0,1)$.

$$\partial_\pi S(\pi, X) = \frac{1}{\pi^2(1-\pi)^2}\Big(-\pi(1-\pi) - (\pi(1-\pi))'(X-\pi)\Big).$$
$$\implies I(\pi) = -E[\partial_\pi S(\pi, X)] = \frac{1}{\pi(1-\pi)}.$$

- Normal distribution: $X \sim N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

$$\partial_\mu S(\mu, X) = -\frac{1}{\sigma^2}.$$
$$\implies I(\mu) = -E[\partial_\mu S(\mu, X)] = \frac{1}{\sigma^2}.$$

$$\partial_{\sigma^2} S(\sigma^2, X) = \frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}.$$
$$\implies I(\sigma^2) = -E[\partial_{\sigma^2} S(\sigma^2, X)] = \frac{1}{2\sigma^4}.$$

2 FI contained in a statistic

The above definitions of the score and FI can be directly applied to any statistics. In fact, let $T \equiv T(\boldsymbol{X}) = T(X_1, \ldots, X_n)$ be a statistic whose pd is given by $h_n(t;\theta)$. The score associated with h_n and its corresponding FI are given by

$$S(\theta, T) = \partial_\theta \log h_n(T;\theta)$$
$$I_T(\theta) = E_\theta \left[S^2(\theta, T)\right]$$

$I_T(\theta)$ is the (Fisher) information about θ that we can extract from T.
Assuming the interchangeability of integration and differentiation twice, this FI can also be expressed as

$$I_T(\theta) = -E_\theta \left[\partial_\theta S(\theta, T)\right].$$

Example 2.1.

- Let $X_i, i = 1, \ldots, n$, be an iid sample from $Be(\pi)$. Let us define the statistic $T = \sum_{i=1}^n X_i$. Since $T \sim Bin(n, \pi)$,

the pd of T is given by $h_n(t;\pi) = P_\pi(T = t) = C_n^t \pi^t (1-\pi)^{n-t}$. It follows that

$$I_T(\pi) = E\left[S^2(\pi, T)\right] = E\left[\frac{T - n\pi}{\pi(1-\pi)}\right]^2 = \frac{n}{\pi(1-\pi)}.$$

- Let $X_i, i = 1, \ldots, n$, be an iid sample from a Normal distribution $N(\mu, \sigma^2)$. Since $\overline{X}_n \sim N(\mu, \sigma^2/n)$, it follows that

$$I_{\overline{X}_n}(\mu) = \frac{n}{\sigma^2}. \square$$

An important fact about FI is its ***additivity***. Let T_1 and T_2 be two statistics with pd h_1 and h_2, and with FI $I_{T_1}(\theta)$ and $I_{T_2}(\theta)$, respectively. If T_1 and T_2 are *independent*, i.e. if $h(t_1, t_2; \theta) = h_1(t_1;\theta)h_2(t_2;\theta)$, $\forall t$, with h being the joint pd of (T_1, T_2), then

$$\begin{aligned} I_{(T_1,T_2)}(\theta) &= E(\partial_\theta \log h(T_1, T_2; \theta))^2 = E(\partial_\theta \log h_1(T_1;\theta) + \partial_\theta \log h_2(T_2;\theta))^2 \\ &= E(\partial_\theta \log h_1(T_1;\theta))^2 + E(\partial_\theta \log h_2(T_2;\theta))^2 + 2E(\partial_\theta \log h_1(T_1;\theta)\partial_\theta \log h_2(T_2;\theta)) \\ &= I_{T_1}(\theta) + I_{T_2}(\theta). \end{aligned}$$

As consequence we have the following result.

Proposition 2.1. *Let* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *be an iid sample from a distribution with density* $f(x;\theta)$*, and let* $f_n(\boldsymbol{x};\theta) = \prod_{i=1}^n f(x_i;\theta)$ *denote the joint density of* $\boldsymbol{X}$*. Define the individual and joint Score and FI functions by*

$$S(\theta, X_i) = \partial_\theta \log f(X_i;\theta) \qquad I_{X_i}(\theta) = E\big[S^2(\theta, X_i)\big]$$
$$S(\theta, \boldsymbol{X}) = \partial_\theta \log f_n(\boldsymbol{X};\theta) \qquad I_{\boldsymbol{X}}(\theta) = E\big[S^2(\theta, \boldsymbol{X})\big]$$

Then, $S(\theta, \boldsymbol{X}) = \sum_{i=1}^n S(\theta, X_i)$*, and* $I_{\boldsymbol{X}}(\theta) = \sum_{i=1}^n I_{X_i}(\theta) = n\, I_{X_1}(\theta)$.

Attention. From now on, if no confusion is possible, we use the notation I_n to denote the joint FI, i.e. $I_n = I_{\boldsymbol{X}}$, and write I for the individual FI, i.e. $I = I_{X_i} = I_{X_1}$. With this convention, the equality above becomes $I_n(\theta) = nI(\theta)$.

For any statistic $\boldsymbol{T} = (T_1(\boldsymbol{X}), T_2(\boldsymbol{X}), \ldots, T_d(\boldsymbol{X}))$, $d \geq 1$, it can be shown that

$$0 \leq I_{\boldsymbol{T}}(\theta) \leq I_n(\theta).$$

This inequality expresses a fundamental principle: no statistic computed from the sample can contain more information about θ than the sample itself.

3 Sufficient statistic

One may naturally ask under what circumstances a given statistic can attain the maximal possible Fisher information I_n. To answer this question, we must introduce the notion of sufficiency.

The definition below applies to both the single-parameter case and the multi-parameter case, where $\boldsymbol{\theta}$ may represent a vector of several parameters.

Definition 3.1 (Sufficiency)**.** Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample with joint pd $f_n(\boldsymbol{x}; \boldsymbol{\theta})$. A statistic $\boldsymbol{T}$ is sufficient for $\boldsymbol{\theta}$ if the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{T}$ does not depend on $\boldsymbol{\theta}$.

Put another way, given a sufficient statistic $\boldsymbol{T}$ for $\boldsymbol{\theta}$, the sample $\boldsymbol{X}$ provides no additional information about $\boldsymbol{\theta}$. Sufficient statistics are especially valuable when their dimension d is much smaller than the sample size n, because they reduce a large dataset to one or a few numbers while preserving all the information about the parameter.

Example 3.1. Let $X_i, i = 1, \ldots, n$, be an iid sample from $Be(\pi)$. Define the statistic $T = \sum_{i=1}^{n} X_i$. We have that

$$P(X_1 = x_1, \ldots, X_n = x_n | T = t) = \frac{P(X_1 = x_1, \ldots, X_n = x_n)}{P(T = t)} I(\sum_{i=1}^{n} x_i = t)$$
$$= \frac{\prod_i \pi^{x_i}(1-\pi)^{1-x_i}}{C_n^t \pi^t (1-\pi)^{n-t}} I(\sum_{i=1}^{n} x_i = t) = \frac{I(\sum_{i=1}^{n} x_i = t)}{C_n^t}.$$

This expression does **not** depend on π. Therefore, by the definition of sufficiency, $\sum_{i=1}^{n} X_i$ is sufficient for π. □

Sufficiency and transformation

- If T is sufficient for θ, then T is also sufficient for any transformation $k(\theta)$, regardless of the form of k.
- If T is sufficient for θ, then any statistic $U = k(T)$ remains sufficient for θ, provided that the function k is bijective (or at least one-to-one). In such cases, U contains exactly the same information as T, just expressed on a different scale. For example, in the $Ber(\pi)$ model, since $\sum_{i=1}^{n} X_i$ is sufficient for π, the sample mean $\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ is also sufficient for π. □

The following result, known as the ***Factorization Theorem***, makes it very easy to identify sufficient statistics (this theorem applies to both the single-parameter and multi-parameter cases).

Theorem 3.1 (Factorization Theorem). *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample with joint pd $f_n(\boldsymbol{x}; \boldsymbol{\theta})$. A statistic $\boldsymbol{T}(\boldsymbol{X})$ sufficient for $\boldsymbol{\theta}$ if and only if there exist nonnegative functions φ and h such that f_n can be factorized as*

$$f_n(\boldsymbol{x}; \boldsymbol{\theta}) = \varphi(\boldsymbol{T}(\boldsymbol{x}); \boldsymbol{\theta}) h(\boldsymbol{x}), \ \forall \boldsymbol{x}, \boldsymbol{\theta}.$$

where h does not depend on $\boldsymbol{\theta}$, and φ depends on $\boldsymbol{x}$ only through $\boldsymbol{T}(\boldsymbol{x})$.

Example 3.2.

- Let $X_i, i = 1, \ldots, n$, be an iid sample from $Be(\pi)$. The joint pd of $(X_1, \ldots, X_n)$ is

$$\prod_{i=1}^{n} \pi^{x_i} (1-\pi)^{1-x_i} = \pi^{\sum_i x_i} (1-\pi)^{n - \sum_i x_i}.$$

  It follows that $\sum_{i=1}^{n} X_i$ is sufficient for π.

- Let $X_i, i = 1, \ldots, n$, be an iid sample from $N(\mu, \sigma^2)$. The joint pd of $(X_1, \ldots, X_n)$ is

$$f(x_1, \ldots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$
$$= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2\right)\right)$$

It follows that $\left(\sum_{i=1}^n X_i,\ \sum_{i=1}^n X_i^2\right)$ is sufficient for (μ, σ^2). This implies that $(\overline{X}, S^2)$ is sufficient for $\boldsymbol{\theta}$. □

It can be shown that, for both the single-parameter and multi-parameter cases,

$$\boldsymbol{T} \text{ is sufficient for } \boldsymbol{\theta} \implies \boldsymbol{I}_{\boldsymbol{T}}(\boldsymbol{\theta}) = \boldsymbol{I}_n(\boldsymbol{\theta}).$$

Under some regularity conditions, the reverse is also true[2]. In the multi-parameter case, $\boldsymbol{I}$ (in bold) denotes the Fisher information matrix; see Section 8 for more details.

Example 3.3. We have seen that for the Bernoulli model, the statistic $T = \sum_{i=1}^n X_i$ is sufficient for π. Consequently, the Fisher information contained in T is equal to the Fisher information in the full sample, that is, $I_{\sum_{i=1}^n X_i}(\pi) = I_n(\pi)$. This equality can be verified directly by the calculations carried out in Example 1.1 and Example 2.1. □

[2]See Pollard, D. (2013). A note on insufficiency and the preservation of Fisher information. DOI: 10.1214/12-IMSCOLL919

As a direct consequence of the Factorization theorem, we have the following result that allows one to obtain sufficient statistics when data come from an exponential family.

Proposition 3.1. *If $\boldsymbol{X} = (X_1, \ldots, X_n)$ is an iid sample from a J-parameter exponential family ($J \geq 1$) with pd*

$$h(x) \exp\big(\boldsymbol{\eta}^t(\boldsymbol{\theta})\boldsymbol{T}(x) - B(\boldsymbol{\theta})\big),$$

then the statistic $\sum_{i=1}^n \boldsymbol{T}(X_i)$ is sufficient for $\boldsymbol{\theta}$.

Example 3.4. Let $X_1, \ldots, X_n$ be an iid sample from $f(x;\theta) = \theta x^{\theta-1}$, where $x \in (0,1)$ and $\theta > 0$. Let's show that $\prod_{i=1}^n X_i$ is sufficient for θ. To see this, we can write $f(x;\theta)$ as

$$f(x;\theta) = I(0 < x < 1)x^{-1} \exp(\theta \log x + \log \theta),$$

which is an exponential family with natural statistic $T = \log X$. It follows that $L = \sum_i \log(X_i)$ is sufficient for θ. And since $\prod_{i=1}^n X_i = \exp(L)$ is a bijective function of L, it is also sufficient for θ. □

4 FI and re-parametrization

We have already seen that statistical models can be parameterized in different ways. It is important to realize that FI depends on the chosen parameterization.

Proposition 4.1 (FI re-parametrization)**.** *Let $\eta : \theta \mapsto \eta(\theta)$ be a differentiable, one-to-one transformation of θ. Define the reparameterized pd*

$$f^*(x;\eta) = f(x;\theta),\ \forall x,\ \text{where } \eta = \eta(\theta).$$

Let $I(\theta) = E[\partial_\theta \log f(X;\theta)]^2$ be the FI about θ (when the parameterization in θ is used) and $I(\eta) = E\left[\partial_\eta \log f^*(X;\eta)\right]^2$ be the FI about η (when the parameterization in η is used). Then, $I(\theta)$ and $I(\eta)$ are related by $I(\theta) = I(\eta)\,(\eta'(\theta))^2$.

The proof of this proposition is straightforward and is a direct consequence of the chain rule :

$$I(\theta) = E[\partial_\theta \log f(X;\theta)]^2 = E[\partial_\theta \log f^*(X;\eta)]^2 = E\left[\partial_\theta \eta(\theta) \partial_\eta \log f^*(X;\eta)\right]^2 = (\eta'(\theta))^2 I(\eta).$$

Attention. In the following, for a given model $\{f(x;\theta) : \theta \in \Theta\}$, the notation $I(\eta(\theta))$ will always refer to the FI computed under the reparameterized model $f^*(x;\eta)$ with $\eta = \eta(\theta)$. It does **not** denote the quantity obtained by simply substituting θ with $\eta(\theta)$ inside $I(\theta)$.

Example 4.1.

- Let $X \sim N(\mu, \sigma^2)$. $f(x;\mu,\sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$. We have seen (see Example 1.1 or 1.2) that $I(\sigma^2) = \frac{1}{2\sigma^4}$. Let's now compute $I(\sigma)$. Using $I(\sigma) = -E\left[\partial_\sigma^2 \log f\right]$, we obtain

$$I(\sigma) = -E\left[\partial_\sigma \left\{ -\frac{1}{\sigma} + \frac{(X-\mu)^2}{\sigma^3} \right\}\right] = -E\left[\frac{1}{\sigma^2} - 3\frac{(X-\mu)^2}{\sigma^4}\right] = \frac{2}{\sigma^2}.$$

  This matches the result one obtains by applying Proposition 4.1, in fact :

$$I(\sigma) = I(\sigma^2) \times \left(\partial_t \, t^2 \Big|_{t=\sigma} \right)^2 = I(\sigma^2)(2\sigma)^2 = \frac{2}{\sigma^2}.$$

  Or, equivalently,

$$I(\sigma^2) = I(\sigma) \times \left(\partial_t \, \sqrt{t} \Big|_{t=\sigma^2} \right)^2 = I(\sigma) \left(\frac{1}{2\sqrt{\sigma^2}} \right)^2 = \frac{1}{2\sigma^4}.$$

- Let $X \sim Pois(\theta)$. $f(x;\theta) = \dfrac{\theta^x}{x!} e^{-\theta}$, $x = 0, 1, \ldots$ and $\theta > 0$. Direct calculation leads to

$$I(\theta) = -E\left\{\partial_\theta^2 \log f\right\} = -E\left\{ -\frac{X}{\theta^2} \right\} = \frac{1}{\theta}$$

Now, let's consider the parametrization with $\eta = \log(\theta)$: $f^*(x;\eta) = \frac{e^{x\eta}}{x!}e^{-e^{\eta}}$, $\eta \in (-\infty, \infty)$.

$$I(\eta) = -E\left\{\partial_\eta^2 \log f^*\right\} = -E\{-e^\eta\} = e^\eta$$
$$\iff I(\log(\theta)) = \theta$$

This same result can be obtained by directly applying Proposition 4.1 as follows:

$$I(\theta) = I(\log(\theta)) \times \left(\partial_t\, log(t)\Big|_{t=\theta}\right)^2 = \frac{I(\log(\theta))}{\theta^2}.$$

Thus, $I(\log(\theta)) = \theta^2 I(\theta) = \theta$. □

This last example suggests that, in Poisson model, it is easier to estimate $\eta = \log(\theta)$ than θ when the latter is large. Let's check this out. A natural estimator of θ is $\hat{\theta} := \overline{X}$ and a natural estimator of η is $\hat{\eta} := \log(\hat{\theta}) = \log(\overline{X})$. We have that $MSE(\hat{\theta}) = \frac{\theta}{n}$ and, by first order Taylor polynomial approximation, i.e. $\log(\hat{\theta}) \approx \log(\theta) + (\hat{\theta} - \theta)\frac{1}{\theta}$, we can write $MSE(\hat{\eta}) \approx \frac{1}{n\theta}$. These mean-square errors cannot be compared, as they are of different scales/magnitudes. However, we can see that as θ increases, the MSE performance of $\hat{\theta}$ becomes worse and worse compared to that of $\hat{\eta}$. The following simulation confirms this fact.

```r
n <- 100; theta <- 10
hat.theta <- estSim(dataFun = \() rpois(n, theta), estFun = mean)
```

- Performances of $\hat{\theta}$:

```r
mse(hat.theta, theta)
```

```
    bias      var      mse     mare
-0.00218  0.10112  0.10113  0.02537
```

- Performances of $\hat{\eta}$:

```r
mse(log(hat.theta), log(theta))
```

```
     bias       var       mse      mare
-0.000724  0.001012  0.001013  0.011024
```

Remark. We have seen that, in general, $I(g(\theta)) \neq I(\theta)$. We may ask the question what happen with FI when the data itself, or a statistic from it, are transformed. The answer depends on the type of transformation used. For example in the case of strictly monotonic and differentiable transformation of the data, FI does change. More precisely, if

$U = k(T)$, where k is a differentiable and strictly monotonic function that does not depend on θ, then $I_T(\theta) = I_U(\theta)$. This is a direct consequence of the *Change of Variable(s) Formula*: $f_T(T;\theta) = f_U(U;\theta)\left|k'(U)\right|$. □

5 Information Inequality: The Cramer-Rao Lower bound (CRLB)

We will develop a lower bound for the variance of any given statistic, which can be mainly used (i) as a benchmark for comparing estimator performance, and (ii) to find the MVUE (Minimum Variance Unbiased Estimator). The bound we are interested in is called the Cramer-Rao Lower Bound (**CRLB**), and is given in the following theorem.

Theorem 5.1 (Information Inequality)**.** *Let $\boldsymbol{X} \equiv \boldsymbol{X}_n = (X_1, \ldots, X_n)$ be a random sample with joint pd $f_n(\boldsymbol{x};\theta)$, $\theta \in \Theta$. Assume that assumptions (I) and (II) as given above (see Proposition 1.1) hold for $f_n(\boldsymbol{x};\theta)$, the joint dp of $\boldsymbol{X}$. Let $T \equiv T(\boldsymbol{X})$ be a statistic. Assume that (III) $\partial_\theta E_\theta(T)$ exists and can be obtained by differentiating under the integral (or sum) sign. i.e., $\partial_\theta \int T(\boldsymbol{x}) f_n(\boldsymbol{x};\theta) d\boldsymbol{x} = \int T(\boldsymbol{x}) \partial_\theta f_n(\boldsymbol{x};\theta) d\boldsymbol{x}$. Then*

$$Var_\theta(T) \geq \frac{\left(\partial_\theta E_\theta(T)\right)^2}{I_n(\theta)}, \ \forall \theta \in \Theta.$$

The inequality above is a direct consequence of the Cauchy-Schwarz inequality. The proof goes as follows. Let $S_n \equiv S(\theta, \boldsymbol{X}) = \partial_\theta \log f_n(\boldsymbol{X}; \theta)$ be the Score associated with f_n. By the Cauchy-Schwarz inequality,

$$[Cov(T, S_n)]^2 = [E(TS_n) - E(T)E(S_n)]^2 = [E(TS_n)]^2 \leq Var(T)Var(S_n) = Var(T)I_n(\theta).$$

The final result is the consequence of the fact that $E(TS_n) = \int T(\boldsymbol{x})\partial_\theta \log f_n(\boldsymbol{X}; \theta) f_n(\boldsymbol{x}; \theta) d\boldsymbol{x} = \int T(\boldsymbol{x})\partial_\theta f_n(\boldsymbol{x}; \theta) d\boldsymbol{x} = \partial_\theta E(T)$.

As an immediate consequence, if $\hat{\theta} \equiv T$ is an *unbiased* estimator of θ, then $Var_\theta(\hat{\theta}) \geq \dfrac{1}{I_n(\theta)}$.

More generally, if $\widehat{g(\theta)} \equiv T$ is an unbiased estimator of $g(\theta)$, then, $\forall \theta \in \Theta$,

$$Var_\theta\left(\widehat{g(\theta)}\right) \geq I_n^{-1}(g(\theta)),$$

where, $I_n^{-1}(g(\theta)) = \dfrac{1}{I_n(g(\theta))}$, and $I_n(g(\theta)) := \dfrac{I_n(\theta)}{(g'(\theta))^2}$.[3]

The right-hand side of the inequality above is known as the **Cramér–Rao Lower Bound (CRLB)** for estimating $g(\theta)$,

[3] g does not need to be one-to-one. If it is one-to-one, then $I_n(g(\theta))$ as defined above coincides with the FI of the reparameterization $\theta \mapsto g(\theta)$.

i.e. $CRLB(g(\theta)) = I_n^{-1}(g(\theta))$. *It represents the smallest possible variance that any* ***unbiased estimator*** *of* $g(\theta)$ *can achieve under the stated regularity conditions.*

Note. All (regular) exponential families meet the theorem's assumptions; in particular, Assumption (III) holds for any statistic T, eliminating the need for case-by-case verification. □

An *unbiased* estimator $\widehat{g(\theta)}$ of $g(\theta)$ is called ***efficient*** if its variance equals the CRLB for $g(\theta)$, i.e. if $Var\left(\widehat{g(\theta)}\right) = CRLB(g(\theta)) := I_n^{-1}(g(\theta))$; otherwise its (absolute) efficiency is defined to be

$$Eff\left(\widehat{g(\theta)}\right) := \frac{CRLB(g(\theta))}{Var\left(\widehat{g(\theta)}\right)} = \frac{\left(g^{'}(\theta)\right)^2}{I_n(\theta)Var\left(\widehat{g(\theta)}\right)}.$$

$Eff\left(\widehat{g(\theta)}\right) \in (0, 1]$, and equals 1 if and only if $\widehat{g(\theta)}$ is efficient.

By definition, an efficient estimator is (1) unbiased and (2) its variance is uniformly lower than (or equal to) the variance of any other unbiased estimator. Thus, *an efficient estimator, when it exists, is the uniformly minimum variance unbiased estimator (MVUE).*

Efficiency is a stronger requirement than being the MVUE. Indeed, while every efficient estimator is necessarily an

MVUE, the converse is not true. An estimator can be MVUE without attaining the CRLB, either because the bound is not attainable or because regularity conditions fail and the CRLB does not apply. In many models, no unbiased estimator reaches the CRLB, yet a unique MVUE still exists with variance strictly larger than $1/I_n(\theta)$ for some θ. Thus, efficiency implies MVUE, but MVUE does not imply efficiency.

$$\text{Efficient} \implies \text{MVUE}$$

Example 5.1. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from $N(\mu, \sigma^2)$ (exponential family).

- Suppose that μ is our parameter of interest. We have seen that $I_n(\mu) = n/\sigma^2$. So for any unbiased estimator $\hat{\mu}_n$ of μ,

  $$Var(\hat{\mu}_n) \geq \frac{\sigma^2}{n}.$$

  Now, since $\overline{X}_n$ is an unbiased estimator of μ and $Var(\overline{X}_n) = \frac{\sigma^2}{n}$, we conclude that $\overline{X}_n$ is efficient for μ. And so, $\overline{X}_n$ is the MVUE of μ.

- Suppose that σ^2 is our parameter of interest and μ is known. We known that $I_n(\sigma^2) = \frac{n}{2\sigma^4}$. So for any unbiased estimator $\hat{\sigma}_n^2$ of σ^2,

  $$Var(\hat{\sigma}_n^2) \geq 2\sigma^4/n.$$

On the other hand, we known that $\tilde{\sigma}_n^2 = n^{-1}\sum_i (X_i - \mu)^2$ is an unbiased estimator of σ^2. And, using the fact that $E(X-\mu)^4 = 3\sigma^4$, we have that $Var\left(\tilde{\sigma}_n^2\right) = n^{-1}Var(X-\mu)^2 = n^{-1}\left(E(X-\mu)^4 - \sigma^4\right) = 2\sigma^4/n$. So, we conclude that $\tilde{\sigma}_n^2$ is efficient for σ^2. And so it is the MVUE of σ^2.

- Suppose that μ^2 is our parameter of interest. By the information inequality, with $g : \mu \mapsto \mu^2$, we have that

$$Var(\hat{\delta}) \geq \frac{1}{I_n(\mu^2)} = \frac{(2\mu)^2}{I_n(\mu)} = \frac{(2\mu)^2}{n/\sigma^2} = \frac{4\mu^2\sigma^2}{n}.$$

for any unbiased estimator $\hat{\delta}$ of μ^2. Thus, $CRLB(\mu^2) = 4n^{-1}\mu^2\sigma^2$. But, for now, we cannot say if this limit is attainable or not and thus if there is an efficient estimator for μ^2 or not.

Example 5.2 (Importance of assumptions)**.** Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from $Unif(0,\theta)$, $\theta > 0$. Thus $f(x;\theta) = \frac{1}{\theta}$, $0 < x < \theta$. Since $\partial_\theta \log f(x;\theta) = -1/\theta$, if we apply Theorem 5.1, we could conclude that, for any unbiased estimator $\hat{\theta}$ of θ,

$$Var(\hat{\theta}) \geq \frac{\theta^2}{n}.$$

But, we learned earlier that the estimator $\hat{\theta}_2 = \frac{n+1}{n}X_{(n)}$ is unbiased and its variance is $\frac{\theta^2}{n(n+2)}$, which is uniformly smaller than θ^2/n !

The apparent contradiction occurs because Assumption (I) (support independent of θ) is violated. Consequently, the

FI computed above is incorrect and the CRLB does not apply to this model, so no conflict exists !
Note that it can be shown that $\hat{\theta}_2$ is actually the MVUE of θ. □

Remark. Efficiency is not generally preserved under transformations: $\hat{\theta}$ is efficient for $\theta \not\Longrightarrow g(\hat{\theta})$ is efficient for $g(\theta)$. The only transformations that preserve efficiency are affine maps $g(\theta) = a\theta + b; (a \neq 0)$. Specifically,

$$\hat{\theta} \text{ is efficient for } \theta \implies a\hat{\theta} + b \text{ is efficient for } a\theta + b. \square$$

6 Efficiency in exponential families

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from a one-parameter exponential family, written under two parametrizations:[4]

$$f(x;\theta) = h(x) \exp\left(\eta(\theta)T(x) - B(\theta)\right) \qquad \text{(original parameterization)}$$
$$f^*(x;\eta) = h(x) \exp\left(\eta T(x) - A(\eta)\right) \qquad \text{(canonical parameterization)}$$

[4]For convenience, the symbol η denotes hereafter both the mapping $g : \theta \mapsto \eta(\theta)$ and the natural parameter in the canonical form. The intended meaning will always be clear from the context: when η stands alone it refers to the natural (acnonical) parameter, whereas $\eta(\theta)$ explicitly indicates the original parametrisation.

For the statistic $T_i = T(X_i)$, $i = 1, \ldots, n$, we have seen that (for the canonical representation)

$$E(T_i) = A^{'}(\eta), \text{ and } Var(T_i) = A^{''}(\eta).$$

The FI about η is

$$I(\eta) = E\big(\partial_\eta \log f^*\big)^2 = E\Big(T_1 - A^{'}(\eta)\Big)^2 = A^{''}(\eta).$$

Define the sample mean of the sufficient statistic : $\overline{T} = n^- \sum_i T_i$. Then,

$$E\big(\overline{T}\big) = A^{'}(\eta), \text{ and } Var\big(\overline{T}\big) = \frac{A^{''}(\eta)}{n}.$$

When we regard $A'(\eta)$ as the parameter of interest, the CRLB for estimating this quantity is

$$I^{-1}(A^{'}(\eta)) = \frac{\Big(A^{''}(\eta)\Big)^2}{I_n(\eta)} = \frac{A^{''}(\eta)}{n}.$$

$\Longrightarrow \overline{T}$ is efficient for $A^{'}(\eta)$.

Under the original parametrization, this result is equivalent to say that $\overline{T}$ is efficient for $\dfrac{B'(\theta)}{\eta'(\theta)}$.

Example 6.1. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from the exponential distribution. Thus, for some $\theta > 0$, the pd of X_i is given by

$$\begin{aligned} f(x;\theta) &= \frac{1}{\theta} e^{-\frac{x}{\theta}} I(x > 0) \\ &= I(x > 0) \exp\left(-\frac{1}{\theta} x - \log(\theta) \right) \qquad \text{(exponential family)} \end{aligned}$$

from the result above, we can directly conclude that $\overline{T} := n^{-1} \sum_{i=1}^{n} X_i$ is is efficient for

$$\frac{(log(\theta))'}{(-1/\theta)'} = \frac{1/\theta}{1/\theta^2} = \theta = E(X_1). \square$$

7 CRLB Attainment

A natural question to ask is under what conditions a given unbiased estimator, say $T(\boldsymbol{X})$, of $g(\theta)$ can attain the CRLB? It turns out that the CRLB is achieved only when the definition of the estimator $T(\boldsymbol{X})$ has the special form given in the following theorem.

Theorem 7.1 (CRLB Attainment). *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample with a joint pd $f_n(\boldsymbol{x}; \theta), \theta \in \Theta$. Suppose the regularity conditions of Theorem 5.1 are satisfied. Then a statistic $T(\boldsymbol{X})$ is efficient for $g(\theta)$ if and only if, $\forall \theta \in \Theta$, there exists a function $a_n(\theta) \neq 0$, such that*

$$\partial_\theta \log f_n(\boldsymbol{x}; \theta) = a_n(\theta)[T(\boldsymbol{x}) - g(\theta)]. \tag{1}$$

Moreover, when (1) holds: (i) $a_n(\theta) = I_n(\theta)/g^{'}(\theta)$, and (ii) $f_n(\boldsymbol{x}; \theta)$ belongs to a one-parameter exponential family

Let's proof this result. Let $S_n = \partial_\theta \log f_n(\boldsymbol{X}; \theta)$ be the Score associated with f_n. Remember that, under the stated assumptions, $E(S_n) = 0$, $Var(S_n) = I_n(\theta)$ and $Cov(S_n, T) = \partial_\theta E(T)$.
Suppose that $T(\boldsymbol{X})$ is an efficient estimator of $g(\theta)$. Then, $\forall \theta \in \Theta$, $E(T) = g(\theta)$, and $Var(T) = (g'(\theta))^2/I_n(\theta)$. So, $Cov^2(S_n, T) = (g'(\theta))^2 = Var(T)Var(S_n)$. Since Cauchy-Schwarz inequality become an equality only in the case of linear dependence, we conclude that $\exists a \equiv a_n(\theta) \neq 0$ and $b \equiv b_n(\theta)$, such that $S_n = aT + b$. But since $E(S_n) = 0$, we have that $b = -ag(\theta)$. Thus, $S_n = a(T - g(\theta))$.
Conversely, suppose that, $\forall \theta \in \Theta$, $\exists a \equiv a_n(\theta) \neq 0$ such that $S_n = a(T - g(\theta))$. Observe that $E(S_n) = a(E(T) - g(\theta)) \Rightarrow E(T) = g(\theta)$, $Var(S_n) = a^2 Var(T) \Rightarrow Var(T) = I_n(\theta)/a^2$, and $Cov(S_n, T) = Cov(a(T - g(\theta)), T) = aVar(T) \Rightarrow g'(\theta) = aVar(T)$. This implies that $a = I_n(\theta)/g^{'}(\theta)$, which in turn implies that $Var(T) = (g'(\theta))^2/I_n(\theta)$. This concludes the proof.

The CRLB attainment theorem above leads to an explicit constructive procedure for deriving the (efficient) MVUE of

$g(\theta)$ *when it exists*. Namely, put

$$T = g(\theta) + \frac{g'(\theta)}{I_n(\theta)} \partial_\theta \log f_n(\boldsymbol{X}; \theta).$$

If the expression on the right hand side of the equality above *does not depend on* θ, i.e. T as defined above is a statistic, then T is efficient for $g(\theta)$.

Example 7.1. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from the exponential distribution. Thus, for some $\theta > 0$, the pd of X_i is given by $f(x; \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$, for $x > 0$. It is easy to check that

$$\partial_\theta \log f_n(\boldsymbol{X}; \theta) = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n X_i$$

$$\partial_\theta^2 \log f_n(\boldsymbol{X}; \theta) = \frac{n}{\theta^2} - \frac{2}{\theta^3} \sum_{i=1}^n X_i \Rightarrow I_n(\theta) = -E\left[\partial_\theta^2 \log f_n(\boldsymbol{X}; \theta)\right] = \frac{n}{\theta^2}$$

- Let's try to find the efficient estimator for θ. To do so, put

$$\begin{aligned} T &:= \theta + \partial_\theta \log f_n(\boldsymbol{X}; \theta) / I_n(\theta) \\ &= \theta + \frac{\theta^2}{n} \left(-\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n X_i \right) = n^{-1} \sum_{i=1}^n X_i \end{aligned}$$

$\to n^{-1}\sum_i X_i$ is the desired estimator.

- Let's now try to find an efficient estimator for $\delta = \dfrac{1}{\theta}$. Following the same procedure, let

$$\begin{aligned} T &:= \frac{1}{\theta} + \frac{\left(\frac{1}{\theta}\right)'}{\frac{n}{\theta^2}}\left(-\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} X_i\right) \\ &= \frac{2}{\theta} - \frac{1}{\theta^2} n^{-1} \sum_{i=1}^{n} X_i. \end{aligned}$$

  The latter is not a statistic, so there is no efficient estimator for $1/\theta$. Note, however, that this does not mean that there is no MVUE. □

8 Multiparameter case

The above theory extends naturally to parametric models whose pd $f(x;\boldsymbol{\theta})$ depend on several parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \boldsymbol{\Theta} \subset \mathbb{R}^d$. In what follows, we present the main results without working through the full details of the calculations.

In the sequel, we assume that the following two regularity conditions are satisfied:

(I) *The set* $\{x : \ f(x;\boldsymbol{\theta}) > 0\}$ *does not depend on* $\boldsymbol{\theta}$.

(II) *The operations of integration (or summation) and differentiation by* θ_j *can be interchanged in* $\int f(x;\boldsymbol{\theta})dx$. i.e., $\partial_{\theta_j} \int f(x;\boldsymbol{\theta})dx = \int \partial_{\theta_j} f(x;\boldsymbol{\theta})dx$, $\forall j = 1, \ldots, d$.

The Score vector

The Score vector of f is defined as the gradient of $\boldsymbol{\theta} \mapsto \log f(X;\boldsymbol{\theta})$, i.e.

$$\boldsymbol{S} := \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log f(X;\boldsymbol{\theta}) = (S_1, \ldots, S_d)^t,$$

where $S_j = \partial_{\theta_j} \log f(X;\boldsymbol{\theta})$ is the score for θ_j. It is easy to see that $E(S_j) = 0$, $\forall j$. Thus, $E(\boldsymbol{S}) = \boldsymbol{0}$.

The FI matrix

The ***FI matrix*** contained in X about $\boldsymbol{\theta}$ is defined as

$$\boldsymbol{I}(\boldsymbol{\theta}) := E(\boldsymbol{S}\boldsymbol{S}^t) = Var\,(\boldsymbol{S})\,.$$

Thus, $\boldsymbol{I}(\boldsymbol{\theta}) = \left[I_{jk}(\boldsymbol{\theta})\right]_{1\leq j,k\leq d}$, where $I_{jk}(\boldsymbol{\theta}) := E\left(S_j \times S_k\right) = Cov\left(S_j, S_k\right)$.

Any FI matrix is symmetric and positive semidefinite by construction (it is the covariance matrix of the score vector). It becomes positive definite when the statistical model is regular and identifiable.

Attention. Here, we restrict our attention to regular models, for which the FI matrix is positive definite—and therefore invertible.

FI matrix–Hessian form

If $f(x;\boldsymbol{\theta})$ is twice differentiable and double integration and differentiation under the integral sign can be interchanged, i.e., $\partial_{\theta_k\theta_j}\int f(x;\boldsymbol{\theta})dx = \int \partial_{\theta_k\theta_j} f(x;\boldsymbol{\theta})dx$, $\forall j,k = 1,\dots,d$, then $I_{jk}(\boldsymbol{\theta}) = -E\left(\partial_{\theta_k} S_j\right)$. Thus,

$$\boldsymbol{I}(\boldsymbol{\theta}) = -E\left(\boldsymbol{\nabla}^2_{\boldsymbol{\theta}} \log f(X;\boldsymbol{\theta})\right),$$

where $\boldsymbol{\nabla}^2_{\boldsymbol{\theta}} \log f(X;\boldsymbol{\theta})$ denotes the Hessian of $\boldsymbol{\theta} \mapsto \log f(X;\boldsymbol{\theta})$.

To be more explicit about the formulas given above. Let's consider the special case of a two-parameter model with $\boldsymbol{\theta} = (\theta_1, \theta_2)$, $S_1 = \partial_{\theta_1} \log f(X;\boldsymbol{\theta})$, and $S_2 = \partial_{\theta_2} \log f(X;\boldsymbol{\theta})$. Under the regularity assumptions stated above, we can write the FI matrix in any of the following equivalent expressions:

$$\boldsymbol{I}(\theta_1,\theta_2) = E\begin{pmatrix} S_1^2 & S_1S_2 \\ S_1S_2 & S_2^2 \end{pmatrix} = \begin{pmatrix} Var(S_1) & Cov(S_1,S_2) \\ Cov(S_1,S_2) & Var(S_2) \end{pmatrix} = -E\begin{pmatrix} \partial_{\theta_1}S_1 & \partial_{\theta_2}S_1 \\ \partial_{\theta_1}S_2 & \partial_{\theta_2}S_2 \end{pmatrix}.$$

FI matrix in an iid Sample

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample with joint pd $f_n(\boldsymbol{x}; \boldsymbol{\theta})$. The joint Score and associated FI are defined as

$$\boldsymbol{S}_n = \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log f_n(\boldsymbol{X}; \boldsymbol{\theta}), \;\; \boldsymbol{I}_n(\boldsymbol{\theta}) = E\left(\boldsymbol{S}_n \boldsymbol{S}_n^t\right).$$

Because of the iid assumption, the FI is additive, yielding

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = n\boldsymbol{I}(\boldsymbol{\theta}).$$

Hereafter, we'll use use $\boldsymbol{I}_n^{-1}(\boldsymbol{\theta})$ to denote $[\boldsymbol{I}_n(\boldsymbol{\theta})]^{-1} = n^{-1}[\boldsymbol{I}(\boldsymbol{\theta})]^{-1}$, the inverse matrix of $\boldsymbol{I}_n(\boldsymbol{\theta})$.

Example 8.1. Normal distribution $N(\mu, \sigma^2)$ with $\log f(x; \mu, \sigma^2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$.

$$S_1 = \partial_\mu \log f = \frac{X-\mu}{\sigma^2} \qquad S_2 = \partial_{\sigma^2} \log f = -\frac{1}{2\sigma^2} + \frac{(X-\mu)^2}{2\sigma^4}$$

$$\partial_\mu S_1 = -\frac{1}{\sigma^2} \qquad \partial_\mu S_2 = -\frac{X-\mu}{\sigma^4}$$

$$\partial_{\sigma^2} S_1 = -\frac{X-\mu}{\sigma^4} \qquad \partial_{\sigma^2} S_2 = \frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}$$

$$I_{11} = -E(\partial_\mu S_1) = \frac{1}{\sigma^2} \qquad I_{22} = -E(\partial_{\sigma^2} S_2) = \frac{1}{2\sigma^4}$$

$$I_{21} = -E(\partial_\mu S_2) = 0 \qquad I_{12} = -E(\partial_{\sigma^2} S_1) = 0$$

$$\Rightarrow \quad \boldsymbol{I}(\mu, \sigma^2) = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \text{ and } \boldsymbol{I}_n(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}. \square$$

FI under reparametrization – Multiparameter Case

Assume the model is reparametrized by a one-to-one differentiable mapping $\boldsymbol{\eta} : \boldsymbol{\theta} \mapsto \boldsymbol{\eta}(\boldsymbol{\theta}) = (\eta_1(\boldsymbol{\theta}), \ldots, \eta_d(\boldsymbol{\theta}))$.
Let $\boldsymbol{I}(\boldsymbol{\theta})$ and $\boldsymbol{I}(\boldsymbol{\eta}) \equiv \boldsymbol{I}(\boldsymbol{\eta}(\boldsymbol{\theta}))$ denote the FI matrices under the $\boldsymbol{\theta}$- and $\boldsymbol{\eta}$-parametrizations, respectively.

Then the two matrices are related by

$$\boldsymbol{I}(\boldsymbol{\theta}) = \dot{\boldsymbol{\eta}}^t(\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\eta})\,\dot{\boldsymbol{\eta}}(\boldsymbol{\theta}),$$

where $\dot{\boldsymbol{\eta}}(\boldsymbol{\theta}) = \partial_{\boldsymbol{\theta}}\boldsymbol{\eta}(\boldsymbol{\theta})$ is the $d \times d$ Jacobian matrix of $\boldsymbol{\theta} \mapsto \boldsymbol{\eta}(\boldsymbol{\theta})$, whose (j,k)-entry is $\partial_{\theta_k}\eta_j(\boldsymbol{\theta})$, and $\dot{\boldsymbol{\eta}}^t(\boldsymbol{\theta}) = \left[\dot{\boldsymbol{\eta}}(\boldsymbol{\theta})\right]^t$ denotes its transpose.

Example 8.2. Normal distribution $N(\mu, \sigma^2)$.
Let $\boldsymbol{\eta} : (\mu, \sigma) \mapsto (\eta_1, \eta_2)$, with $\eta_1(\mu, \sigma) = \mu$, and $\eta_2(\mu, \sigma) = \sigma^2$. We have that

$$\dot{\boldsymbol{\eta}} \equiv \dot{\boldsymbol{\eta}}(\mu, \sigma) := \begin{pmatrix} \partial_\mu \eta_1 & \partial_\sigma \eta_1 \\ \partial_\mu \eta_2 & \partial_\sigma \eta_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 2\sigma \end{pmatrix}.$$

$$\Rightarrow \boldsymbol{I}(\mu, \sigma) = \dot{\boldsymbol{\eta}}^t \boldsymbol{I}(\mu, \sigma^2)\dot{\boldsymbol{\eta}} = \begin{pmatrix} 1 & 0 \\ 0 & 2\sigma \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2\sigma \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}. \square$$

Information Inequality Theorem–Multiparameter CRLB

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample with joint pd $f_n(\boldsymbol{x}; \boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d$. Let $\boldsymbol{T} \equiv (T_1(\boldsymbol{X}), \ldots, T_p(\boldsymbol{X}))$ be an estimator of $\boldsymbol{g}(\boldsymbol{\theta})$, where $\boldsymbol{g}(\cdot) = (g_1(\cdot), \ldots, g_p(\cdot))$ is a differentiable mapping

from $\mathbb{R}^d$ to $\mathbb{R}^p$. Assume that
(I) and (II), as given above (see the beginning of the current section), hold for $f_n(\boldsymbol{x};\boldsymbol{\theta})$.
(III) $\partial_{\theta_k} E_{\boldsymbol{\theta}}(T_j)$ exists and can be obtained by differentiating under the integral sign, $\forall j,k,\boldsymbol{\theta}$.
If $\boldsymbol{T}$ is an unbiased estimator of $\boldsymbol{g}(\boldsymbol{\theta})$, i.e. if $E_{\boldsymbol{\theta}}(T_j) = g_j(\boldsymbol{\theta})$, $j = 1,\ldots,p$, then, $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$Var_{\boldsymbol{\theta}}(\boldsymbol{T}) \succeq \boldsymbol{I}_n^{-1}(\boldsymbol{g}(\boldsymbol{\theta})),$$

where $Var_{\boldsymbol{\theta}}(\boldsymbol{T})$ is the variance-covariance matrix of $\boldsymbol{T}$, $\boldsymbol{I}_n^{-1}(\boldsymbol{g}(\boldsymbol{\theta})) := \dot{\boldsymbol{g}}(\boldsymbol{\theta})\boldsymbol{I}_n^{-1}(\boldsymbol{\theta})\dot{\boldsymbol{g}}^t(\boldsymbol{\theta})$[5], $\dot{\boldsymbol{g}}$ is the $p \times d$ Jacobian matrix of $\boldsymbol{g}$, whose $(j,k)-$th element is $\partial_{\theta_k} g_j(\boldsymbol{\theta})$, and $\dot{\boldsymbol{g}}^t(\boldsymbol{\theta}) = [\dot{\boldsymbol{g}}(\boldsymbol{\theta})]^t$.
In particular, if $d = p$ and $\boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, i.e. if $\boldsymbol{T}$ is unbiased for $\boldsymbol{\theta}$, then $Var_{\boldsymbol{\theta}}(\boldsymbol{T}) \succeq \boldsymbol{I}_n^{-1}(\boldsymbol{\theta})$.

Above, the notation $\boldsymbol{A} \succeq \boldsymbol{B}$ means that $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite matrix. Consequently, writing $Var(\boldsymbol{T}) \succeq \boldsymbol{I}_n^{-1}$ is equivalent to say that $Var(\boldsymbol{a}^t\boldsymbol{T}) \geq \boldsymbol{a}^t\boldsymbol{I}_n^{-1}\boldsymbol{a}$, $\forall \boldsymbol{a}$.

This equivalence reveals a key interpretation: each choice of $\boldsymbol{a}$ selects a direction in the parameter space, corresponding to a particular linear combination of parameters. The CRLB states that, in every such direction, the variance of the corresponding estimator (i.e. $Var(\boldsymbol{a}^t\boldsymbol{T})$) cannot fall below the threshold $\boldsymbol{a}^t\boldsymbol{I}_n^{-1}\boldsymbol{a}$ determined by the FI matrix. In short, the CRLB provides a universal lower bound on how precisely we can estimate all parameters (or several functions thereof) simultaneously.

[5] $\boldsymbol{g}$ does not need to be one-to-one, but if it is, then $\boldsymbol{I}_n^{-1}(\boldsymbol{g}(\boldsymbol{\theta}))$, as defined above, coincides with $[\boldsymbol{I}_n(\boldsymbol{g}(\boldsymbol{\theta}))]^{-1}$, the inverse of the FI about $\boldsymbol{g}(\boldsymbol{\theta})$.

In the case of $p = 1$, i.e. when $g : \mathbb{R}^d \mapsto \mathbb{R}$ is a scalar-valued function, the CRLB simplifies to the following form: for any unbiased estimator T of $g(\boldsymbol{\theta})$,

$$Var(T) \geq \boldsymbol{\nabla}^t g(\boldsymbol{\theta}) I_n^{-1}(\boldsymbol{\theta}) \boldsymbol{\nabla} g(\boldsymbol{\theta}),$$

where where $\boldsymbol{\nabla} g(\boldsymbol{\theta}) = (\partial_1 g(\boldsymbol{\theta}), \ldots, \partial_d g(\boldsymbol{\theta}))^t$ is the gradient of g.

Example 8.3. Normal distribution $N(\mu, \sigma^2) \equiv N(\theta_1, \theta_2)$. For any unbiased estimator $\hat{\eta}$ of $\eta = g(\mu, \sigma^2) \in \mathbb{R}$, we have

$$\begin{aligned} Var(\hat{\eta}) &\geq \begin{pmatrix} \partial_\mu g(\mu, \sigma^2) & \partial_{\sigma^2} g(\mu, \sigma^2) \end{pmatrix} \boldsymbol{I}_n^{-1}(\mu, \sigma^2) \begin{pmatrix} \partial_\mu g(\mu, \sigma^2) \\ \partial_{\sigma^2} g(\mu, \sigma^2) \end{pmatrix} \\ &= \frac{\sigma^2}{n} \big(\partial_\mu g(\mu, \sigma)\big)^2 + \frac{2\sigma^4}{n} \big(\partial_{\sigma^2} g(\mu, \sigma^2)\big)^2. \end{aligned}$$

For example, if we are interested in estimating the coefficient of variation (CV), taking $g(\mu, \sigma^2) = \sigma/\mu$ and applying the CRLB gives

$$Var(\hat{\eta}) \geq \frac{\sigma^2}{n} \left(-\frac{\sigma}{\mu^2}\right)^2 + \frac{2\sigma^4}{n} \left(\frac{1}{2\mu\sigma}\right)^2 = \frac{\sigma^2}{n\mu^2} \left(\frac{\sigma^2}{\mu^2} + \frac{1}{2}\right).$$

$\longrightarrow$ no unbiased estimator (if any) of the CV can have variance smaller than $\dfrac{\sigma^2}{n\mu^2} \left(\dfrac{\sigma^2}{\mu^2} + \dfrac{1}{2}\right)$ — this is the best possible precision for any unbiased CV estimator under normality. □

Effect of nuisance parameters

The multiparameter CRLB provides important insight into *the effect of nuisance parameters on estimation precision*. To see this, consider again the two-parameter case ($d = 2$), with $\boldsymbol{\theta} = (\theta_1, \theta_2)$. Suppose θ_1 is our *primary parameter of interest* and θ_2 is a *nuisance parameter*. In the normal model, for example, we might be interested only in the mean μ, while σ^2 is a a nuisance parameter that is required for model correctness but is not of direct interest.
In such a context, the natural question is: can we estimate θ_1 with the same precision we would achieve if θ_2 were known ?
To investigate this, we start by writing the FI matrix for the two-parameter model, along with its inverse. Using the same notations for the score components as above (S_1 for θ_1 and S_2 for θ_2), we write

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = n \begin{pmatrix} Var(S_1) & Cov(S_1, S_2) \\ Cov(S_2, S_1) & Var(S_2) \end{pmatrix} \equiv n \begin{pmatrix} I_{11} & I_{12} \\ I_{12} & I_{22} \end{pmatrix}.$$

$$\boldsymbol{I}_n^{-1}(\boldsymbol{\theta}) = \frac{1}{n(I_{11}I_{22} - I_{12}^2)} \begin{pmatrix} I_{22} & -I_{12} \\ -I_{12} & I_{11} \end{pmatrix}.$$

We can think of two situations :

- θ_2 *is known*: In this case, the model reduces to a single parameter to $\theta = \theta_1$, and we return to the univariate setting. Applying the univariate CRLB theorem (Theorem 5.1), we have, for any unbiased estimator T_1 of θ_1,

$$Var(T_1) \geq \frac{1}{nI_{11}}.$$

- θ_2 *is unknown*: In this case, applying the multiparameter version of the CRLB theorem with $g : (\theta_1, \theta_2) \mapsto \theta_1$, we have, for any unbiased estimator T_1 of θ_1,

$$Var(T_1) \geq (1,0)\boldsymbol{I}_n^{-1}(\boldsymbol{\theta})(1,0)^t = \frac{I_{22}}{n(I_{11}I_{22} - I_{12}^2)} = \frac{1}{nI_{11}^*},$$

where $I_{11}^* := \dfrac{I_{11}I_{22} - I_{12}^2}{I_{22}} = I_{11}\left(1 - \dfrac{I_{12}^2}{I_{11}I_{22}}\right) = I_{11}\left(1 - Cor^2(S_1, S_2)\right)$.

Observe that $I_{11}^* \leq I_{11}$, with equality if and only if S_1 and S_2 are uncorrelated—that is, when the Fisher Information matrix is diagonal.

This shows that the presence of nuisance parameters can reduce the effective information available for estimating the parameters of interest. In other words, nuisance parameters not only complicate the estimation process but also diminish the attainable precision by lowering the corresponding FI.

Example 8.4. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from $N(\mu, \sigma^2)$, where $\boldsymbol{\theta} = (\mu, \sigma^2)$ is unknown. We have seen that

the FI matrix for this model is

$$\boldsymbol{I}_n(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix},$$

which is ***diagonal*** $\longrightarrow$ no loss from nuisance parameters.

- Suppose that μ is our parameter of interest. The information inequality states that $Var(\hat{\mu}) \geq \frac{\sigma^2}{n}$, for any unbiased estimator $\hat{\mu}$ of μ, *whether σ^2 is known or not*.
  On the other hand, since $E(\overline{X}_n) = \mu$ and $Var(\overline{X}_n) = \sigma^2/n$, $\overline{X}_n$ is efficient for μ, regardless of whether σ^2 is known.

- Suppose now that σ^2 is our parameter of interest. The information inequality tells us that $Var(\hat{\sigma}^2) \geq \frac{2\sigma^4}{n}$, for any unbiased estimator $\hat{\sigma}^2$ of σ^2, *whether μ is known or not*.

  – If μ is known, a natural estimator of σ^2 is $\tilde{\sigma}^2_n = n^{-1}\sum_i (X_i - \mu)^2$. We have seen that $E(\tilde{\sigma}^2_n) = \sigma^2$ and $Var(\tilde{\sigma}^2_n) = 2\sigma^4/n$. Hence, in this setting, $\tilde{\sigma}^2_n$ attains the CRLB and is therefore an **efficient estimator** of σ^2.

  – If μ is unknown, $\tilde{\sigma}^2_n$ is no longer usable because it depends on the (unknown) true value of μ. In this case, an unbiased estimate of σ^2 is $S^2_n = \frac{1}{n-1}\sum_i (X_i - \overline{X}_n)^2$. However, $Var(S^2_n) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}$. So, S^2_n is unbiased but **slightly inefficient**. Nevertheless, it can be shown that S^2_n is the MVUE of σ^2. □

Example 8.5. Let $\boldsymbol{X} = (X_1, X_2, X_3)$ be a random vector drawn from the trinomial distribution $Mult(n, (\pi_1, \pi_2, \pi_3))$. Put $\boldsymbol{x} = (x_1, x_2, x_3)$, and $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$. The joint pd of $\boldsymbol{X}$ is

$$f_n(\boldsymbol{x}; \boldsymbol{\pi}) := P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{n!}{x_1! x_2! x_3!} \pi_1^{x_1} \pi_2^{x_2} \pi_3^{x_3},$$

with $x_1, x_2, x_3 \in \{0, \ldots, n\}$, $x_3 = n - x_1 - x_2$, and $\pi_3 = 1 - \pi_1 - \pi_2$. Because of these constraints, $f(\boldsymbol{x}; \boldsymbol{\pi})$ depends effectively only on (x_1, x_2) and (π_1, π_2). The parameter space is $\boldsymbol{\Theta} = \{(\pi_1, \pi_2) \in [0,1]^2 : \pi_1 \geq 0,\ \pi_2 \geq 0,\ \pi_1 + \pi_2 \leq 1\}$. It is known that $E(\boldsymbol{X}) = n\boldsymbol{\pi}$ and $Var(\boldsymbol{X}) = n(diag(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^t)$.
We have that

$$\begin{aligned}
S_1 &= \partial_{\pi_1} \log f_n(\boldsymbol{X}; \boldsymbol{\pi}) = X_1 \partial_{\pi_1} log(\pi_1) + X_2 \partial_{\pi_1} log(\pi_2) + X_3 \partial_{\pi_1} log(\pi_3) \\
&= \frac{X_1}{\pi_1} + X_3 \partial_{\pi_1}(\pi_3) \partial_{\pi_3}(log(\pi_3)) = \frac{X_1}{\pi_1} - \frac{X_3}{\pi_3}. \\
\partial_{\pi_1} S_1 &= X_1 \partial_{\pi_1} \pi_1^{-1} - X_3 \partial_{\pi_1} \pi_3^{-1} = -\frac{X_1}{\pi_1^2} - X_3 \partial_{\pi_1}(\pi_3) \partial_{\pi_3}(\pi_3^{-1}) = -\frac{X_1}{\pi_1^2} - \frac{X_3}{\pi_3^2}. \\
\partial_{\pi_2} S_1 &= -X_3 \partial_{\pi_2} \pi_3^{-1} = -\frac{X_3}{\pi_3^2}.
\end{aligned}$$

Similarly, by simetry, $S_2 := \partial_{\pi_2} \log f_n(\boldsymbol{X}; \boldsymbol{\pi}) = \frac{X_2}{\pi_2} - \frac{X_3}{\pi_3}$, $\partial_{\pi_1} S_2 = -\frac{X_3}{\pi_3^2}$, and $\partial_{\pi_2} S_2 = -\frac{X_2}{\pi_2^2} - \frac{X_3}{\pi_3^2}$.

Using the fact that $E(X_j) = n\pi_j$, $j = 1, 2, 3$, we get the FI matrix

$$\boldsymbol{I}_n(\pi_1, \pi_2) = -E \begin{pmatrix} \partial_{\pi_1} S_1 & \partial_{\pi_2} S_1 \\ \partial_{\pi_1} S_2 & \partial_{\pi_2} S_2 \end{pmatrix} = n \begin{pmatrix} \frac{1}{\pi_1} + \frac{1}{\pi_3} & \frac{1}{\pi_3} \\ \frac{1}{\pi_3} & \frac{1}{\pi_2} + \frac{1}{\pi_3} \end{pmatrix}.$$

The inverse of thi matrix is

$$\boldsymbol{I}_n^{-1}(\pi_1, \pi_2) = n^{-1}(\pi_1\pi_2\pi_3) \begin{pmatrix} \frac{1}{\pi_2} + \frac{1}{\pi_3} & -\frac{1}{\pi_3} \\ -\frac{1}{\pi_3} & \frac{1}{\pi_1} + \frac{1}{\pi_3} \end{pmatrix} = n^{-1} \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) \end{pmatrix}.$$

Suppose that π_1 is our parameter of interest and π_2 is a nuisance parameter. Given the FI matrix above and its inverse, according and the information inequality, we can say that, for any unbiased estimator $\hat{\pi}_1$ of π_1,

$$Var(\hat{\pi}_1) \geq \frac{1}{n\left(\frac{1}{\pi_1} + \frac{1}{\pi_3}\right)} = \frac{\pi_1(1-\pi_1-\pi_2)}{n(1-\pi_2)}, \text{ if } \pi_2 \text{ is known.}$$

$$Var(\hat{\pi}_1) \geq \frac{\pi_1(1-\pi_1)}{n}, \text{ if } \pi_2 \text{ is unknown.}$$

And because $\frac{\pi_1(1-\pi_1-\pi_2)}{n(1-\pi_2)} < \frac{\pi_1(1-\pi_1)}{n}$, knowing π_2 gives a smaller (tighter) lower bound, meaning that knowing π_2 allows potentially more precise estimation of π_1. □

An unbiased estimator $\boldsymbol{T} = (T_1, \ldots, T_k)^\top$ of a vector-valued parameter $\boldsymbol{g}(\boldsymbol{\theta})$ is efficient if its covariance matrix attains the CRLB for all $\boldsymbol{\theta}$ in the parameter space, i.e.

$$Var_{\boldsymbol{\theta}}(\boldsymbol{T}) = \dot{\boldsymbol{g}}(\boldsymbol{\theta}) \boldsymbol{I}_n^{-1}(\boldsymbol{\theta}) \dot{\boldsymbol{g}}^t(\boldsymbol{\theta})$$

The multiparameter version of the **CRLB attainment theorem** provides a practical criterion for establishing the existence of an efficient estimator and, when one exists, determining its explicit form. Under regularity conditions (I)–(III), the theorem states that if the quantity

$$\boldsymbol{T} := \boldsymbol{g}(\boldsymbol{\theta}) + \dot{\boldsymbol{g}}(\boldsymbol{\theta}) \boldsymbol{I}_n^{-1}(\boldsymbol{\theta}) \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log f_n(\boldsymbol{X}; \boldsymbol{\theta}),$$

is free of $\boldsymbol{\theta}$, then this $\boldsymbol{T}$ constitutes an efficient estimator of $\boldsymbol{g}(\boldsymbol{\theta})$.

Example 8.6. Consider again the trinomial model as in our previous example. We seek to determine whether an efficient estimator exists for $\boldsymbol{\pi} = (\pi_1, \pi_2)$. For that, we need to compute

$$\boldsymbol{\pi} + \boldsymbol{I}_n^{-1}(\boldsymbol{\pi}) \boldsymbol{\nabla}_{\boldsymbol{\pi}} \log f_n(\boldsymbol{X}; \boldsymbol{\pi}) = \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix} + n^{-1} \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) \end{pmatrix} \begin{pmatrix} \frac{X_1}{\pi_1} - \frac{X_3}{\pi_3} \\ \frac{X_2}{\pi_2} - \frac{X_3}{\pi_3} \end{pmatrix}$$

The first component of the matrix by the vector product above is

$$\begin{aligned}\pi_1(1-\pi_1)\left(\frac{X_1}{\pi_1}-\frac{X_3}{\pi_3}\right)-\pi_1\pi_2\left(\frac{X_2}{\pi_2}-\frac{X_3}{\pi_3}\right) &= (1-\pi_1)X_1-\pi_1X_2+(\pi_1(\pi_1+\pi_2)-\pi_1)\frac{X_3}{\pi_3}\\ &= (1-\pi_1)X_1-\pi_1X_2+(\pi_1(1-\pi_3)-\pi_1)\frac{X_3}{\pi_3}\\ &= X_1-\pi_1(X_1+X_2+X_3)=X_1-n\pi_1.\end{aligned}$$

As for the second competent, exactly the same algebra yields to

$$-\pi_1\pi_2\left(\frac{X_1}{\pi_1}-\frac{X_3}{\pi_3}\right)+\pi_2(1-\pi_2)\left(\frac{X_2}{\pi_2}-\frac{X_3}{\pi_3}\right)=X_2-n\pi_2$$

Thus,

$$\boldsymbol{\pi}+\boldsymbol{I}_n^{-1}(\boldsymbol{\pi})\boldsymbol{\nabla}_{\boldsymbol{\pi}}\log f_n(\boldsymbol{X};\boldsymbol{\pi})=\begin{pmatrix}\pi_1\\ \pi_2\end{pmatrix}+n^{-1}\begin{pmatrix}X_1-n\pi_1\\ X_2-n\pi_2\end{pmatrix}=\begin{pmatrix}\frac{X_1}{n}\\ \frac{X_2}{n}\end{pmatrix}.$$

$\Longrightarrow \begin{pmatrix}\frac{X_1}{n}\\ \frac{X_2}{n}\end{pmatrix}$ is efficient for (π_1,π_2). $\square$