A Few Notes on Fisher Information (WIP)

David Meyer

dmm@{1-4-5.net,uoregon.edu}

Last update: November 22, 2020

1 Definitions

There are so many interesting things about Fisher Information [1] and its theoretical properties (and for that matter, its applications), but I'll start here with definitions. Let $\{P_\theta\}_{\theta \in \Theta}$ denote a parametric family of distributions on a space $\mathcal{X}$, where each $\theta \in \Theta \subset \mathbb{R}^d$ indexes $\{P_\theta\}$. Assume (with no real loss of generality, AFAICT) that each P_θ has a density p_θ. Then the Fisher information associated with the *model* is the matrix given by

$$\mathbb{I}_\theta := \mathbb{E}_\theta\Big[\nabla_\theta \log p_\theta(X) \nabla_\theta \log p_\theta(X)^{\mathrm{T}}\Big] = \mathbb{E}_\theta[l_\theta l_\theta^T] \tag{1}$$

where l_θ is the *score function* $\nabla_\theta \log p_\theta(X)$. The score function turns up in perhaps unexpected places (e.g., score function gradient in policy gradient reinforcement learning) and has various applications.

Essentially the Fisher Information tells us how much information a random variable $\mathcal{X}$ contains about the parameter vector θ, where $\mathcal{X}$ is distributed according to a probability distribution parameterized by θ. Intuitively, the Fisher information captures the variability of the gradient of the score function, $\nabla_\theta \log p_\theta(x)$; in a family of distributions for which the score function has high variability, we expect estimation of the parameter θ to be easier; essentially (and perhaps counter-Intuitively) events with lower probability contain more information.

Fisher Information can be described in various ways. One way to think about the Fisher Information is that it is the variance of the score function l_θ. Recall that the mean $\mu = \mathbb{E}[X]$ and the variance $\sigma^2 = \mathbb{E}[(X - \mu)^2]$. First, consider $\mathbb{E}\big[\nabla_\theta \log p_\theta(X)\big] = \mathbb{E}\big[l_\theta\big]$.

$$\mathbb{E}_\theta\big[l_\theta\big] = \int p_\theta(x)\nabla_\theta \log p_\theta(x)dx \qquad \text{\# Definition of Expectation} \tag{2}$$

$$= \int \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)} p_\theta(x)dx \qquad \text{\#}\ \nabla_\theta \log p_\theta(x) = \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)} \tag{3}$$

$$= \int \nabla_\theta p_\theta(x)dx \qquad \text{\#}\ \frac{p_\theta(x)}{p_\theta(x)} = 1 \tag{4}$$

$$= \nabla_\theta \int p_\theta(x)dx \qquad \text{\# Assume } \int \text{ and } \nabla \text{ can be exchanged} \tag{5}$$

$$= \nabla_\theta 1 \qquad \text{\#}\ p_\theta(x) \text{ is a density so } \int_x p_\theta(x)dx = 1 \tag{6}$$

$$= 0 \qquad \text{\# for c any constant } \nabla c = 0 \tag{7}$$

So the mean μ of the score function, $\mathbb{E}_\theta\big[\nabla_\theta \log p_\theta(X)\big] = 0$. This condition (expected value is zero) is the *regularity condition* required by the Cramér Rao Bound. See Section 3 for more on the Cramér Rao Bound (sometimes Cramér Rao Lower Bound or CRLB).

Now, recall that the variance $\sigma^2 = \mathbb{E}[(X-\mu)^2]$. However, we know that $\mathbb{E}_\theta\big[l_\theta\big] = 0$, so $\mathbb{I}_\theta = \mathrm{Var}(l_\theta) = \mathbb{E}\big[\big(\nabla_\theta \log p_\theta(x)\big)^2\big]$. Under relatively mild conditions ($\frac{\partial^2}{\partial\theta}p_\theta(x)$ exists) and using the chain rule we see that

$$\nabla^2_\theta \log p_\theta(x) = \frac{\nabla^2_\theta p_\theta(x)}{p_\theta(x)} - \frac{\nabla_\theta p_\theta(x)\nabla_\theta p_\theta(x)^T}{p_\theta(x)^2} \qquad \text{\# finite differences} \tag{8}$$

$$= \frac{\nabla^2_\theta p_\theta(x)}{p_\theta(x)} - \nabla_\theta \log p_\theta(x)\nabla_\theta \log p_\theta(x)^T \qquad \text{\#}\ \nabla_\theta \log p_\theta(x) = \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)} \tag{9}$$

$$= \frac{\nabla^2_\theta p_\theta(x)}{p_\theta(x)} - l_\theta l_\theta^T \qquad \text{\# definition of } l_\theta \tag{10}$$

$$l_\theta l_\theta^T = -\nabla^2_\theta \log p_\theta(x) + \frac{\nabla^2_\theta p_\theta(x)}{p_\theta(x)} \qquad \text{\# algebra} \tag{11}$$

So far we have the following forms of the Fisher Information $\mathbb{I}_\theta$:

$$\mathbb{I}_\theta = \mathrm{Var}(l_\theta) = \mathbb{E}_\theta\big[l_\theta l_\theta^{\mathrm{T}}\big] = \mathbb{E}\big[\big(\nabla_\theta \log p_\theta(x)\big)^2\big] \tag{12}$$

It turns out that using Equation 11 we can see that $\mathbb{I}_\theta$ also turns out to equal $-\mathbb{E}_\theta\big[\nabla^2 \log p_\theta(x)\big]$, which we can see as follows:

$$
\begin{align}
\mathbb{I}_\theta &= \mathbb{E}_{x\sim p_\theta(x)}\big[l_\theta l_\theta^T\big] && \#\ \text{ definition of } \mathbb{I}_\theta \tag{13}\\
&= \mathbb{E}_\theta\left[-\nabla_\theta^2 \log p_\theta(x) + \frac{\nabla_\theta^2 p_\theta(x)}{p_\theta(x)}\right] && \#\ \text{ Equation 11} \tag{14}\\
&= -\int p_\theta(x)\nabla^2 \log p_\theta(x)dx + \int \frac{p_\theta(x)\nabla^2 p_\theta(x)dx}{p_\theta(x)} && \#\ \text{ Defn Expectation} \tag{15}\\
&= -\int p_\theta(x)\nabla^2 \log p_\theta(x)dx + \int \nabla^2 p_\theta(x)dx && \#\ \frac{p_\theta(x)}{p_\theta(x)} = 1 \tag{16}\\
&= -\mathbb{E}_\theta\big[\nabla^2 \log p_\theta(x)\big] + \nabla^2 \underbrace{\int p_\theta(x)dx}_{1} && \#\, p_\theta(x) \text{ is a density} \tag{17}\\
&= -\mathbb{E}_\theta\big[\nabla^2 \log p_\theta(x)\big] + \nabla^2 1 && \#\ \text{ Equation 17} \tag{18}\\
&= -\mathbb{E}_\theta\big[\nabla^2 \log p_\theta(x)\big] + 0 && \#\, \nabla c = 0 \tag{19}\\
&= -\mathbb{E}_\theta\big[\nabla^2 \log p_\theta(x)\big] \tag{20}
\end{align}
$$

where the c in Equation 19 is any constant.

So in summary, we know that that the Fisher Information $\mathbb{I}_\theta$ equals

$$
\mathbb{I}_\theta = \mathbb{E}_\theta\big[l_\theta l_\theta^T\big] = -\mathbb{E}_\theta\big[\nabla^2 \log p_\theta(x)\big] \tag{21}
$$

This representation also makes clear the additional fact that, if we have n i.i.d. observations from the model P_θ then the information content grows linearly, that is, $\log p_\theta(X_1^n) = \sum_{i=1}^{n} \log p_\theta(X_i)$.

Perhaps surprisingly, $\mathbb{I}_\theta$ also equals $\mathrm{Cov}_x\big[\nabla_\theta \log p_\theta(x)\big]$, which we'll see in Section 3. But there's more. The Fisher Information is related to the KL-divergence, as follows:

$$
\mathbb{I}_\theta = \nabla_{\theta'}^2 D_{\mathrm{KL}}(p_{\theta'}(x) \mid\mid p_\theta(x)) = \nabla_{\theta'}^2 D_{\mathrm{KL}}(p_\theta(x) \mid\mid p_{\theta'}(x)) \tag{22}
$$

It is unusual for the KL to be symmetrical in this way. We will see more about this in Section 2.

2 Estimation and Fisher information

The Fisher information has intimate connections to estimation, both in terms of classical estimation and various "information games". The canonical motivating example appears to be estimation of the mean of a sample of n i.i.d. $X_1^n \sim \text{Bernoulli}(p), p \in [0,1]$. The standard formulation of the Bernoulli probability mass function (pmf) f for $x \in [0,1]$ looks like

$$f(x;p) = \begin{cases} p & \text{if } x = 1, \\ 1-p & \text{if } x = 0. \end{cases} \tag{23}$$

$$= p^x(1-p)^{1-x} \tag{24}$$

So if we take the pmf of the Bernoulli distribution to be $P(X = x) = p^x(1-p)^{1-x}$, then $\nabla P(X = x) = \frac{x}{p} - \frac{1-x}{1-p}$, so that

$$\mathbb{I}_p = \mathbb{E}_p\left[\left(\frac{X}{p} - \left(\frac{1-X}{1-p}\right)\right)^2\right] \tag{25}$$

$$= \frac{1}{p} + \frac{1}{1-p} \tag{26}$$

$$= \frac{1}{p(1-p)} \tag{27}$$

where $\mathbb{I}_p$ is the Fisher Information for a single sample.

The variance of a Bernoulli distribution with sample mean μ looks like this:

$$\mathbb{E}\big[(X-\mu)^2\big] = \frac{1}{n}\text{Var}(X) \qquad \#\ \mu = p \text{ (Bernoulli disribution)} \tag{28}$$

$$= \frac{p(1-p)}{n} \qquad \#\ \text{Variance of the Bernoulli distribution} \tag{29}$$

$$= \frac{1}{\mathbb{I}_p} \cdot \frac{1}{n} \qquad \#\ p(1-p) = \frac{1}{\mathbb{I}_p},\ \text{Equation 27} \tag{30}$$

$$\text{Var}(X) = \frac{1}{\mathbb{I}_\theta} \tag{31}$$

It turns out that this inverse dependence of the variance on Fisher Information (Equation 31) is unavoidable, and is quantified by the Cramér Rao Bound, which provides lower bounds on the mean squared error of all unbiased estimators.

3 The Cramér Rao Lower Bound

The Cramér Rao Lower Bound (CRLB) answers the question: Given an estimation problem, what is the variance of the best possible estimator?

3.1 Brief Review of Terms

By way of review, recall that a *statistic* is a single measure of some attribute of a sample (e.g., its arithmetic mean value). It is calculated by applying a function (statistical algorithm) to the values of the items of the sample, which are known together as a set of data. So suppose we have a sample $D = X_1^n$ then $f : D \to \mathbb{R}$ is a *statistic on* D. Statistics have all kinds of interesting properties that we won't review here, including observability[1] completeness, consistency, sufficiency, unbiasedness. Other examples of properties of a statistic include minimum mean square error, low variance, robustness, and computational convenience, and measures of information such as the Fisher Information.

An *estimator* is a rule for calculating an estimate of a statistic based on observed data. An estimator is usually described in terms of an (unknown) parameter θ. So an "estimator" or "point estimate" is a statistic (that is, a function of the data, e.g., f above) that is used to infer the value of an unknown parameter in a statistical model. The parameter being estimated is sometimes called the estimand. It can be either finite-dimensional (in parametric and semi-parametric models), or infinite-dimensional (semi-parametric and non-parametric models). In any cases the parameter is denoted θ and the estimator is traditionally written by adding a circumflex over the symbol: $\widehat{\theta}$. Being a function of the data, the estimator is itself a random variable; a particular realization of this random variable is called the "estimate", the words "estimator" and "estimate" are used interchangeably.

An example of an estimator is the *bias*: The bias of an estimator $\widehat{\theta}$ is defined as $\text{Bias}(\widehat{\theta}) = \mathbb{E}[\widehat{\theta}] - \theta$. It is the distance between the average of the collection of estimates and the true value of the parameter being estimated. Note that the bias of $\widehat{\theta}$ is a function of the true value of θ so saying that the bias of $\widehat{\theta}$ is b means that for every θ the bias of $\widehat{\theta}$ is b. An *unbiased* estimator has the property that $\mathbb{E}[\widehat{\theta}] = \theta$, that is, $\text{Bias}(\widehat{\theta}) = 0$. Note that interestingly, the *Mean Squared Error* (MSE) of a sample, its variance, and its bias, are related in particular, $\text{MSE}(\widehat{\theta}) = \text{Var}(\widehat{\theta}) + \text{Bias}(\widehat{\theta})^2$, That is, the MSE = Variance + square of the Bias. This suggests the important result that for an unbiased estimator $\widehat{\theta}$, $\text{Var}(\widehat{\theta}) = \text{MSE}(\widehat{\theta})$. You can show this pretty easily:

[1] A statistic is an observable random variable, which differentiates it both from a parameter that is a generally unobservable quantity describing a property of a statistical population, and from an unobservable random variable, such as the difference between an observed measurement and a population average, that is., $\frac{1}{n} \sum_{i=1}^{n} X_i - \mu$, where the X_i are the observations and μ is the true population mean.

$$
\begin{aligned}
\mathrm{MSE}(\widehat{\theta}) &= \mathbb{E}\big[(\theta - \widehat{\theta})^2\big] && (32)\\
&= \mathbb{E}\big[(\theta - \mathbb{E}[\widehat{\theta}] + \mathbb{E}[\widehat{\theta}] - \widehat{\theta})^2\big] && (33)\\
&= \mathbb{E}\big[(\theta - \mathbb{E}[\widehat{\theta}])^2 + \mathbb{E}\big[(\mathbb{E}[\widehat{\theta}] - \widehat{\theta})^2\big] + \mathbb{E}\big[(2(\theta - \mathbb{E}[\widehat{\theta}])(\mathbb{E}[(\widehat{\theta}] - \widehat{\theta})\big] && (34)\\
&= \mathbb{E}\big[\underbrace{(\theta - \mathbb{E}[\widehat{\theta}])^2}_{\mathrm{Bias}(\widehat{\theta})^2}\big] + \mathbb{E}\big[\underbrace{(\mathbb{E}[\widehat{\theta}] - \widehat{\theta})]^2}_{\mathrm{Var}(\widehat{\theta})}\big] + \underbrace{2(\theta - \mathbb{E}[\widehat{\theta}])(\mathbb{E}[\widehat{\theta}] - \mathbb{E}[\widehat{\theta}])}_{\mathbb{E}[\widehat{\theta}] - \mathbb{E}[\widehat{\theta}] = 0} && (35)\\
\mathrm{MSE}(\widehat{\theta}) &= \mathrm{Bias}(\widehat{\theta})^2 + \mathrm{Var}(\widehat{\theta}) && (36)
\end{aligned}
$$

Of course, if $\mathrm{Bias}(\widehat{\theta}) = 0$ ($\widehat{\theta}$ is unbiased), then $\mathrm{MSE}(\widehat{\theta}) = \mathrm{Var}(\widehat{\theta})$.

Back to the CRLB. Recall that the question at had was given an estimation problem, what is the variance of the best possible estimator? The CRLB gives and answer to this question, as well as a somewhat impractical method for finding the best estimator. To get some intuition about the CRLB, consider the the problem of estimating the DC level A in the model

$$x[n] = \mathrm{A} + w[n] \tag{37}$$

where the signal we are looking for is embedded in the White Gaussian Noise (WGN) $w[n] \sim \mathcal{N}(0, \sigma^2)$. For simplicity, suppose that we are using only a single observation to do this, and let this observation be $x[0]$. Then the pdf of $x[0]$ is Gaussian:

$$p(x[n]; A) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[0] - \mathrm{A})^2\right] \tag{38}$$

One we observe some value of $x[0]$, say $x[0] = 3$, some values of A become more likely. In fact, the pdf of A has the same form as the pdf of $x[0]$.

$$p(A) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[0] - 3)^2\right] \tag{39}$$

Figure 1 shows the pdf plotted for A = 3 and $\sigma^2 = \frac{1}{3}$ and $\sigma^2 = 1$. The main observation here is that that we will probably get more accurate results for the pdf in the upper case, but why? Before answering this question, here are a few observations

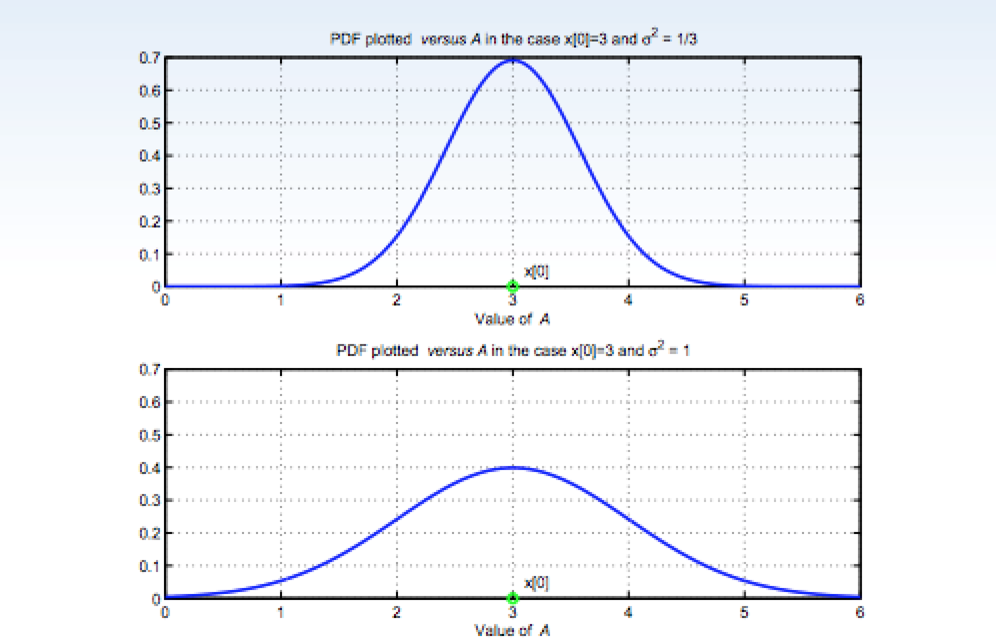

Figure 1: pdfs for A = 3, and $\sigma^2 = 1/3$ and $\sigma^2 = 1$

- When the pdf is viewed as a function of the unknown parameter θ and with $x[.]$ fixed. the pdf is called the *likelihood function*.
- The curvature of the likelihood function determines how accurately we can estimate the parameter. For example, for the pdf on top in Figure 1 it is easier to estimate θ. In fact, if a single sample is observed as $x[0] = \mathrm{A} + w[0]$, then we can expect a better estimate if σ^2 is small.

But still, *why* should we expect a better estimate where σ^2 is small, and can we quantify this effect? Further, is there any measure that would be common for all possible estimators for a specific estimation problem? Specifically, we?d like to find the smallest possible variance. We know that the second derivative of the likelihood function (or log-likelihood) is one alternative for measuring the curvature of a function, so this is a good candidate. Let's see what happens in this case. BTW, what this is saying in some sense is that the minimum variance $\sigma^2_{\min} = -\frac{1}{curvature}$.

First, notice that in this case the minimum variance of all estimators is σ^2 since we are using only one sample. We also know that the likelihood function; this is given in Equation

38. Given this likelihood, the log-likelihood is

$$\ln p(x[0]; \mathrm{A}) = -\ln \sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}(x[0] - A)^2 \tag{40}$$

The derivative with respect A to a is given by

$$\frac{\partial \ln p(x[0]; \mathrm{A})}{\partial \mathrm{A}} = \nabla_{\mathrm{A}} \ln p(x[0]; \mathrm{A}) = \frac{1}{\sigma^2}(x[0] - \mathrm{A}) \tag{41}$$

and the second derivative is

$$\frac{\partial^2 \ln p(x[0]; \mathrm{A})}{\partial^2 \mathrm{A}} = \nabla_{\mathrm{A}}^2 \ln p(x[0]; \mathrm{A}) = -\frac{1}{\sigma^2} \tag{42}$$

Since σ^2 is the smallest possible variance, in this case we have an alternative way of finding the minimum variance of all estimators, in paricular

$$\sigma_{\mathrm{min}}^2 = -\frac{1}{\frac{\partial^2 \ln p(x[0];\mathrm{A})}{\partial^2 \mathrm{A}}} = -\frac{1}{\text{curvature of the likelihood}} \tag{43}$$

In order to generalize the result, the following regularity conditions must hold:

- If the function depends on the data X, you must take the expectation with respect to X
- If the function depends on the parameter θ, then evaluate the derivative at the true value of θ

All of this put together gives us the Minimum Variance Unbiased Estimator (MVUE):

$$\mathrm{MVUE} = \frac{1}{-\mathbb{E}\left[\frac{\partial^2 \ln p(x[0];\theta)}{\partial^2 \theta}\right]}$$

3.2 CRLB Scalar Parameter

Theorem: CRLB Scalar Parameter Case

Assuming that the pdf $p(x;\theta)$ satisfies the regularity condition that $\forall\theta$, $\mathbb{E}_{x\sim p(x;\theta)}\left[\frac{\partial \ln p(x;\theta)}{\partial\theta}\right] = 0$, then the *variance* of any unbiased estimator $\widehat{\theta}$ must satisfy

$$\mathrm{Var}(\widehat{\theta}) \geqslant \frac{1}{-\mathbb{E}\left[\frac{\partial^2 \ln p(x;\theta)}{\partial^2\theta}\right]} \tag{44}$$

An important consequence of the Theorem is that any unbiased estimator can attain this lower bound $\forall\theta$ if and only if

$$\frac{\partial \ln p(x;\theta)}{\partial\theta} = \mathbb{I}_\theta(g(x) - \theta) \tag{45}$$

for some functions g and $\mathbb{I}$. That estimator, which is the MVUE has $\widehat{\theta} = g(x)$ with minimum variance is $\frac{1}{\mathbb{I}_\theta}$. The proof of Equation 45 can be found in [2] and elsewhere around the web.

Notes on the CRLB Scalar Parameter Case

- Note that the Fisher Information $\mathbb{I}_\theta = \mathbb{E}\Big[\frac{\partial^2 \ln p(x;\theta)}{\partial^2\theta}\Big]$ so that by the Theorem: $\mathrm{Var}(\widehat{\theta}) \geqslant \mathbb{I}_\theta^{-1}$.
- The CRLB provides a lower bound for an estimator's error variance
- We need to know the pdf to evaluate CRLB. Often we don?t know this information and cannot evaluate this bound. OTOH, if the data is multivariate Gaussian or i.i.d. with known distribution then we can find the CRLB.
- If the estimator reaches the CRLB, it is called *efficient*.
- The MVUE may or may not be efficient, and it is not guaranteed that MVUE exists or even is realizable.

3.3 Example: Estimation of the DC Level in WGN

Consider the DC level in White Gaussian Noise with N data points: $x[n] = A + w[n], n = 0, 1, \dots, N-1$. The question we want to answer is: What is the minimum variance of any unbiased estimator using N samples? Recall that the pdf and the likelihood function are both the product of the individual densities so that

$$p(x;A) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big[-\frac{1}{2\sigma^2}(x[n]-A)^2\Big] \tag{46}$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\Big[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\Big] \tag{47}$$

The log-likelihood is now

$$\ln p(x;A) = -\ln\big[(2\pi\sigma^2)^{\frac{N}{2}}\big] - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2 \tag{48}$$

and the first derivative is

$$\frac{\partial \ln p(x;A)}{\partial A} = \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A) \tag{49}$$

$$= \frac{N}{\sigma^2}(\tilde{\mathbf{x}}-A) \tag{50}$$

where $\tilde{\mathbf{x}}$ is the sample mean. The second derivative has the simple form

$$\frac{\partial^2 \ln p(x;A)}{\partial A^2} = -\frac{N}{\sigma^2} \tag{51}$$

hence the minimum variance of any *unbiased* estimator is

$$\mathrm{Var}(\hat{A}) \geqslant \frac{N}{\sigma^2} \tag{52}$$

Theorem: CRLB Vector Parameter Case

It is again assumed that the pdf $p(x;\theta)$ satisfies the *regularity* conditions as above. Then the covariance matrix of any unbiased estimator $\widehat{\theta}$ not surprisingly satisfies

$$\mathbf{C}_{\widehat{\theta}} - \mathbb{I}_{\theta}^{-1} \geqslant 0 \tag{53}$$

where $\geqslant$ is interpreted as *positive semidefinite*. Here the Fisher Information matrix is given by

$$\left[\mathbb{I}_{\theta}\right]_{ij} = -\mathbb{E}_{x\sim p(x;\theta)}\left[\frac{\partial \ln p(x;\theta)}{\partial \theta_i \partial \theta_j}\right] \tag{54}$$

Furthermore, an unbiased estimator may be found that attains the bound in that $\mathbf{C}_{\widehat{\theta}} = \mathbb{I}_{\theta}^{-1}$ *iff*

$$\frac{\partial \ln p(x;\theta)}{\partial \theta} = \mathbb{I}_{\theta}(g(x)-\theta) \tag{55}$$

for some function g and some $m \times m$ matrix $\mathbb{I}_{\theta}$. That estimator, which s the MVUE, is given by $\widehat{\theta} = g(x)$ and the covariance matrix is $\mathbb{I}_{\theta}^{-1}$.

References

[1] A. Ly, M. Marsman, J. Verhagen, R. Grasman, and E.-J. Wagenmakers, "A Tutorial on Fisher Information," *ArXiv e-prints*, May 2017.

[2] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory.* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.