

# Estimation, Fisher Information, and the Cramer-Rao Bound

Hossein Falaki  
Computer Science Department,  
University of California, Los Angeles  
falaki@cs.ucla.edu

## 1 Introduction

In this note we will present a geometric interpretation of the asymptotic variance of the estimator solved from an estimating equation. We will also introduce the Fisher information of a set of i.i.d samples. We will prove the Cramer-Rao bound on the variance of any unbiased estimator.

## 2 Geometric Interpretation

For a set of i.i.d samples:

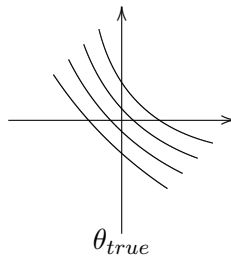
$$X_1, X_2, \dots, X_n \sim p(x, \theta_{true})$$

consider an unbiased estimator of  $\theta$

$$\hat{\theta} = h(X)$$

Because  $h()$  is an unbiased estimator of  $\theta$ :

$$\forall \theta \quad \frac{1}{n} \sum_{i=1}^n h(x_i, \theta) = 0 \tag{1}$$



Writing the Taylor expansion of this equation  $\theta_{true}$  we get:

$$\frac{1}{n} \sum_{i=1}^n h(x_i, \theta) = \underbrace{\frac{1}{n} \sum_{i=1}^n h(x_i, \theta_{true})}_{\text{intercept}} + \underbrace{\frac{1}{n} \sum_{i=1}^n h'(x_i, \theta_{true})(\theta - \theta_{true})}_{\text{slope}}$$

Therefore:

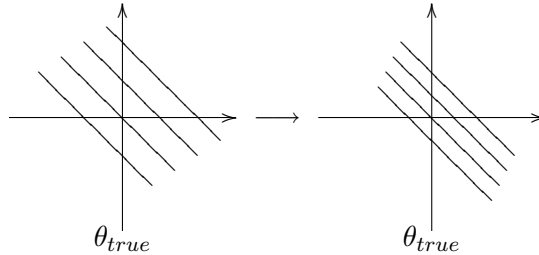
$$\hat{\theta} - \theta_{true} = -\frac{\frac{1}{n} \sum_{i=1}^n h(X_i, \Theta_{true})}{\frac{1}{n} \sum_{i=1}^n h'(X_i, \Theta_{true})}$$

In the previous note we proved that:

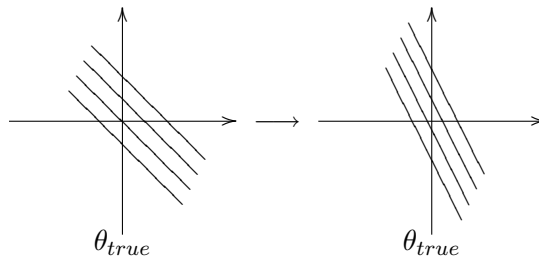
$$\sqrt{n}(\hat{\Theta} - \Theta_{true}) = -\frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n h(X_i, \Theta_{true})}{\frac{1}{n} \sum_{i=1}^n h'(X_i, \Theta_{true})} \sim N(0, \frac{Var(h(x, \theta_{true}))}{E_{\theta_{true}}(h'(x, \theta_{true}))^2})$$

This equation indicates that to minimize the variance of  $\hat{\theta}$  under hypothetical repeated sampling, we should minimize the intercept and at the same time maximize the slope. This situation can geometrically be interpreted in the following way:

- With fixed slope: the lower the intercept is, the lower is the variance of  $\hat{\theta}$ :



- With fixed intercept: the higher the slope is, the lower is the variance of  $\hat{\theta}$



### 3 Fisher Information

In this section we define the concept of *Fisher Information*. First consider the following important identity: Consider

$$m(\theta) = E_{\theta}(h(X, \theta)) = \int h(x, \theta)p(x, \theta)dx$$

By taking the derivative of we get:

$$\begin{aligned}
\frac{\partial}{\partial\theta}E_{\theta}(h(x, \theta)) &= \int \frac{\partial}{\partial\theta}[h(x, \theta)p(x, \theta)]dx \\
&= \int [h'(x, \theta)p(x, \theta) + h(x, \theta)p'(x, \theta)]dx \\
&= E_{\theta}[h'(x, \theta)] + \int h(x, \theta) \frac{\partial}{\partial\theta}(\log p(x, \theta))p(x, \theta)dx \\
&= E_{\theta}[h'(x, \theta)] + E_{\theta}(h(x, \theta) \frac{\partial}{\partial\theta} \log p(x, \theta)) \\
&= E_{\theta}[h'(x, \theta)] + Cov_{\theta}(h(x, \theta), \frac{\partial}{\partial\theta} \log p(x, \theta))
\end{aligned}$$

In other words:

$$\frac{\partial}{\partial\theta}E_{\theta}(h(x, \theta)) = E_{\theta}(h'(x, \theta)) + Cov_{\theta}(h(x, \theta), \frac{\partial}{\partial\theta} \log p(x, \theta)) \quad (2)$$

Recall that

$$\sqrt{n}(\hat{\Theta} - \Theta_{true}) \sim N(0, \frac{Var(h(x, \theta_{true}))}{E_{\theta_{true}}(h'(x, \theta_{true}))^2})$$

Substituting the Covariance with it equivalent we get:

$$\begin{aligned}
\frac{Var(h(x, \theta_{true}))}{E_{\theta_{true}}(h'(x, \theta_{true}))^2} &= \frac{Var(h(x, \theta_{true}))}{Var_{\theta}(h(x, \theta))Var(\frac{\partial}{\partial\theta} \log p(x, \theta))Corr(h(x, \theta), \frac{\partial}{\partial\theta} \log p(x, \theta))} \\
&= \frac{1}{Var(\frac{\partial}{\partial\theta} \log p(x, \theta))Corr(h(x, \theta), \frac{\partial}{\partial\theta} \log p(x, \theta))}
\end{aligned}$$

In the case of MLE:

$$h(x, \theta) = \frac{\partial}{\partial\theta} \log p(x, \theta)$$

Therefore:

$$\frac{Var(h(x, \theta_{true}))}{E_{\theta_{true}}(h'(x, \theta_{true}))^2} = \frac{1}{Var(\frac{\partial}{\partial\theta} \log p(x, \theta))}$$

In other words:

$$\sqrt{n}(\hat{\Theta} - \Theta_{true}) \sim N(0, \frac{1}{Var(\frac{\partial}{\partial\theta} \log p(x, \theta))}) \quad (3)$$

$Var(\frac{\partial}{\partial\theta} \log p(x, \theta)) = Info(\theta)$  is referred to as the Fisher information. The larger it is, the smaller is the variance of  $\hat{\theta}$ .

$Info(\theta)$  can be considered the information in a single observation. Here, we will give a geometric interpretation of the Fisher information based on the curvature of the log-likelihood function.

Using identity 2, taking the derivative of the MLE function we get:

$$\begin{aligned}\frac{\partial}{\partial\theta}E_{\theta}\left(\frac{\partial}{\partial\theta}\log p(x,\theta)\right) &= E_{\theta}\left(\frac{\partial^2}{\partial\theta^2}\log p(x,\theta)\right) \\ &= E_{\theta}(h'(x,\theta)) + \text{Info}(\theta)\end{aligned}$$

Note that  $E_{\theta}(h'(x,\theta)) = 0$ .

*Proof.*

$$\int p(x,\theta)dx = 1 \implies \int \frac{\partial}{\partial\theta}p(x,\theta)dx = 0$$

$$\begin{aligned}E_{\theta}\left(\frac{\partial}{\partial\theta}\log p(x,\theta)\right) &= \int \frac{\partial}{\partial\theta}\log p(x,\theta)p(x,\theta)dx \\ &= \int \frac{\partial}{\partial\theta}p(x,\theta)dx = 0\end{aligned}$$

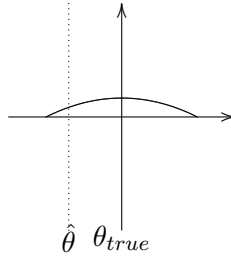
□

Therefore:

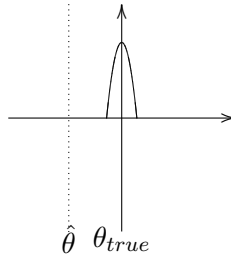
$$E_{\theta}\left(\frac{\partial^2}{\partial\theta^2}\log p(x,\theta)\right) = \text{Info}(\theta) \quad (4)$$

Equation 4 gives the geometric interpretation that we need.  $\text{Info}(\theta)$  is equivalent with the curvature of the log-likelihood function.

When the curvature is low ( $\text{Info}(\theta)$  is low) deviation from  $\theta_{true}$  will be plausible in the MLE method.



But when the curvature ( $\text{Info}(\theta)$ ) is high, small deviation from  $\theta_{true}$  is considered unacceptable.



## 4 The Cramer-Rao Bound

In this section we prove that MLE extracts the maximum information content of i.i.d samples. In other words the variance of any other estimator for  $\theta$  is higher than MLE. This result is known as the Cramer-Rao bound and is presented as follows:

Assume that  $X_1, X_2, \dots, X_n$  are samples (non necessarily i.i.d) from  $p(x, \theta_{true})$ , and consider any unbiased estimator of  $\theta$ :

$$\hat{\theta} = h(X)$$

$$E_{\theta}(h(x)) = \theta_{true}$$

The Cramer-Rao bound states that:

$$Var(\hat{\theta}) \geq \frac{1}{Info(\theta_{true})} \quad (5)$$

Before presenting the proof note that:

1. This is not an asymptotic bound
2. The bound is achieved by MLE

*Proof.*

$$E_{\theta}(h(X)) = \theta$$

$$\begin{aligned} \frac{\partial}{\partial \theta} E_{\theta}(h(X)) &= 1 \\ &= E_{\theta}(h'(X)) + cov(h(X), \frac{\partial}{\partial \theta} \log p(x, \theta))^2 \\ &= 0 + Cov(h(X), \frac{\partial}{\partial \theta} \log p(x, \theta))^2 \\ &\leq Var(h(X)) Var(\frac{\partial}{\partial \theta} \log p(x, \theta)) \end{aligned}$$

Considering that  $h(X) = \hat{\theta}$  we get:

$$Var(\hat{\theta}) \geq \frac{1}{Info(\theta_{true})}$$

□

## 5 Acknowledgment

This note is based on Prof. Ying Nian Wu's lectures on Theoretical Statistics at UCLA.