Note on Cramér–Rao Lower Bound *

Kenji I. Sato

December 23, 2009

1 Introduction

Let $Y = \{y_1, \ldots, y_m\}$ is a collection of given data which are realizations of i.i.d. random trials. We assume that the probabilistic events occur with probability density $p(y|\theta)$, depending on an unknown parameter $\theta \in \mathbb{R}^n$. Our concern is to estimate θ from Y. We might be interested particularly in the best performance we can expect for an estimation. One of lower bounds is obtained from analyzing the Fisher information; it is the Cramér–Rao bound.

Next section provides basic definitions and the problem formulation. The third section gives the proof for the Cramér–Rao bound.

2 Problem

We assume that every function we will encounter has sufficient smoothness. The expectation operator with respect to the density $p(y|\theta)$ is denoted by $\mathbb{E}_\theta[\cdot]$. Thus, $\mathbb{E}_\theta[X] = \int X(y)p(y|\theta)dy$.

Let $\hat{\theta}(Y)$ be an unbiased estimator of θ; i.e., $\mathbb{E}_\theta[\hat{\theta}(Y)] = \theta$ holds. To begin with, let us define several concepts concerning estimation performance of $\hat{\theta}$. Define the score of $p(y|\theta)$ by

$$V_\theta(Y) := \nabla_\theta \log p(Y|\theta) := \sum_{i=1}^{m} \nabla_\theta \log p(y_i|\theta).$$

Note that $V_\theta(Y)$ has zero mean because

$$\begin{aligned}
\mathbb{E}_\theta[V_\theta(Y)] &= \int p(Y|\theta)\nabla_\theta \log p(Y|\theta)dY \\
&= \int p(Y|\theta)\sum_{i=1}^{m} \frac{\nabla_\theta p(y_i|\theta)}{p(y_i|\theta)}dY \\
&= \int \nabla_\theta \left(\prod_{i=1}^{m} p(y_i|\theta)\right) dy_1 \cdots dy_m \\
&= \nabla_\theta \left[\prod_{i=1}^{m} \left(\int p(y_i|\theta)dy_i\right)\right] \\
&= 0.
\end{aligned}$$

Definition 2.1 (Fisher Information). The Fisher information I is defined to be the covariance matrix of $V_\theta(Y)$:

$$I := \mathbb{E}_\theta[V_\theta(Y)V_\theta(Y)^T].$$

The covariance matrix of $\hat{\theta}(Y)$ will be denoted by

$$B := \mathbb{E}_\theta\left[(\hat{\theta}(Y) - \mathbb{E}_\theta[\hat{\theta}(Y)])(\hat{\theta}(Y) - \mathbb{E}_\theta[\hat{\theta}(Y)])^T\right]$$

Definition 2.2 (Efficiency). Efficiency of estimate $\hat{\theta}(Y)$ is defined by

$$e := 1/\det IB.$$

Since both I and B are positive definite, $e > 0$.

Theorem 2.3 (Cramér–Rao). *The efficiency e of any unbiased estimator $\hat{\theta}(Y)$ is bounded above by 1; thus,*

$$e \leq 1.$$

This upper bound is considered to be the lower bound of the variance B of $\hat{\theta}$. For the scalar case, the variance B is bounded from below by $1/I$. In this sense, the Cramér–Rao bound tells us up to how good the estimation can be.

3 Proof

Lemma 3.1. *Let $A, B \in \mathbb{C}^{n\times n}$ be Hermitian and positive semi-difinite. It holds that*

$$\det(A + B) \geq \det A + \det B.$$

*This note is based on Schölkopf and Smola [2].

Proof. This follows from the Minkowski's determinantal inequality

$$[\det(A+B)]^{1/n} \geq (\det A)^{1/n} + (\det B)^{1/n}.$$

///

Lemma 3.2 (Everitt [1]). *Let $A_1, A_2, B \in \mathbb{C}^{n\times n}$. A_1, A_2 are assumed to be Hermitian. If*

$$A = \begin{bmatrix} A_1 & B \\ B^* & A_2 \end{bmatrix}$$

is positive semi-definite,

$$|\det B|^2 \leq \det A_1 \cdot \det A_2.$$

Proof. Firstly, we suppose that A_1 and A_2 are non-singular. Consider the Shur complement of A. It is well-known that A is similar to the matrix

$$\tilde{A} = \begin{bmatrix} A_1 - BA_2^{-1}B^* & 0 \\ 0 & A_2 \end{bmatrix}.$$

By Lemma 3.1,

$$\begin{aligned} \det A_1 &= \det(A_1 - BA_2^{-1}B^* + BA_2^{-1}B^*) \\ &\geq \det(A_1 - BA_2^{-1}B^*) + \det BA_2^{-1}B^* \\ &\geq \det BA_2^{-1}B^*. \end{aligned}$$

These inequalities are true because A is positive semi-definite and so is $\tilde{A}$. Therefore,

$$\begin{aligned} \det A_1 \cdot \det A_2 &\geq \det BA_2^{-1}B^* \cdot \det A_2 \\ &= \det BB^* = |\det B|^2. \end{aligned}$$

Next we suppose $\det A_1 \cdot \det A_2 = 0$. Let us define, for $\delta > 0$,

$$\hat{A}(\delta) = A + \delta E_{2n},$$

where E_m is $m \times m$ unit matrix for $m \in \mathbb{N}$. By the fact that we have proved in the above,

$$\begin{aligned} |\det B|^2 &\leq \det(A_1 + \delta E_n) \cdot \det(A_2 + \delta E_n) \\ &\to 0, \quad (\text{as } \delta \to 0). \end{aligned}$$

The target inequality trivially holds. ///

The next lemma, which plays a key role to prove the theorem, is a special case of Lemma 3.2.

Lemma 3.3. *Let x, y be $\mathbb{R}^n$-valued random variables with mean zero. Then,*

$$(\det \mathbb{E}[xy^T])^2 \leq \det \mathbb{E}[xx^T] \cdot \det \mathbb{E}[yy^T].$$

Proof. Notice that covariance matrix

$$\mathbb{E}\left[\begin{bmatrix} x \\ y \end{bmatrix}\begin{bmatrix} x^T & y^T \end{bmatrix}\right] = \begin{bmatrix} \mathbb{E}[xx^T] & \mathbb{E}[xy^T] \\ \mathbb{E}[yx^T] & \mathbb{E}[yy^T] \end{bmatrix}$$

is positive semi-definite and symmetric. Applying Lemma 3.2, we have

$$(\det \mathbb{E}[xy^T])^2 \leq \det \mathbb{E}[xx^T] \cdot \det \mathbb{E}[yy^T].$$

///

Proof of the Theorem. Note that

$$\begin{aligned} &\det \mathbb{E}_\theta\left[(V_\theta(Y) - \mathbb{E}_\theta[V_\theta(Y)])(\hat{\theta}(Y) - \mathbb{E}_\theta[\hat{\theta}(Y)])^T\right] \\ &= \det \mathbb{E}_\theta[V_\theta(Y)\hat{\theta}(Y)^T] \leq \det I \cdot \det B = \det IB. \end{aligned}$$

The first equality is by $\mathbb{E}_\theta[V_\theta(Y)] = 0$; the inequality holds by Lemma 3.3. The rest of the proof is devoted to showing that $\mathbb{E}_\theta[V_\theta(Y)\hat{\theta}(Y)^T]$ is the identity matrix, that immediately gives its determinant is unity. To see this, let us consider the ij-th component. Let us denote the derivative of V and $p(y|\theta)$ in terms of θ_i by V_{θ_i} and $p_{\theta_i}(y|\theta)$, respectively. Straightforward calculation gives us

$$\begin{aligned} \mathbb{E}_\theta[V_{\theta_i}(Y)\hat{\theta}_j(Y)] &= \int p(Y|\theta)V_{\theta_i}(Y)\hat{\theta}_j(Y)dY \\ &= \frac{\partial}{\partial\theta_i}\int p(Y|\theta)\hat{\theta}_j(Y)dY \\ &= \frac{\partial}{\partial\theta_i}\theta_j(Y) \\ &= \delta_{i,j}, \end{aligned}$$

where $\delta_{i,j}$ is the Kronecker's delta. ///

In this proof we have used a short-hand notation for simultaneous distribution. It is not hard to construct a rigorous proof in the same manner as the proof for $\mathbb{E}_\theta[V_\theta(Y)] = 0$ in the previous page.

References

[1] Everitt, W. N. (1958) "A Note on Positive Definite Matrices," *Glasgow Mathematical Journal*, Vol. 3, pp. 173–175.

[2] Schölkopf, B. and A. Smola (2002) *Lerning with Kernels*: MIT Press.