STAT 700 November 19, 2022

Handout on Multidimensional Cramér-Rao Inequality

This handout contains two main topics by way of elaboration of class notes and the resuilts on necessary conditions for exactly achieving the lower variance bound of the Cramér-Rao Information Inequality. The first topic is the extension to multivariate parameter θ and statistic $S(X)$, and the second topic is the necessary and sufficient condition for exactly attaining the Cramér-Rao lower bound.

First consider the Cauchy-Schwarz Inequality for multivariate functions. Suppose $g(x)$ and $\underline{h}(x)$ are respectively scalar- and d-vector-valued square-integrable functions of x (which may also be multidimensional) with respect to a probability density $f(x)$, and assume that

$$M \;\equiv\; \int \underline{h}(x)^{\otimes 2}(x)\, f(x)\, dx \qquad \text{is positive-definite}$$

where throughout these pages, for any vector $\underline{v}$, we denote $\underline{v}^{\otimes 2} \;=\; \underline{v}\,\underline{v}^{tr}$. Then the ordinary Cauchy-Schwarz Inequality implies

$$\forall \quad \underline{v} \in \mathbb{R}^d, \qquad \Big(\int g(x)\, \{\underline{h}(x)^{tr}\, \underline{v}\}\, f(x)\, dx \Big)^2 \;\le\; \int g^2(x)\, f(x)\, dx \int \{\underline{v}^{tr}\, \underline{h}(x)^{\otimes 2}\, \underline{v}\}\, f(x)\, dx$$

By the positive definiteness assumed for M, it follows that for all nonzero vectors $\underline{v} \in \mathbb{R}^d$,

$$\int g^2(x)\, f(x)\, dx \;\ge\; \Big(\int g(x)\, \{\underline{h}(x)^{tr}\, \underline{v}\}\, f(x)\, dx \Big)^2 \Big/ \int \{\underline{v}^{tr}\, \underline{h}(x)^{\otimes 2}\, \underline{v}\}\, f(x)\, dx \tag{1}$$

Maximizing the right-hand side over vectors $\underline{v}$ such that $\underline{v}^{tr}\, M\, \underline{v} \;=\; 1$ using scalar Lagrange multiplier $\;-\lambda\;$ gives the equation

$$2 \int \{\underline{v}^{tr}\, \underline{h}(x)\}\, g(x)\, f(x)\, dx \int \underline{h}(x)\, g(x)\, f(x)\, dx \;=\; 2\lambda\, M\, \underline{v}$$

which implies that

$$\underline{v} \;=\; \text{const} \cdot M^{-1} \int \underline{h}(x)\, g(x)\, f(x)\, dx$$

Substituting this choice of $\underline{v}$ into (1) immediately implies

$$\int g^2(x)\, f(x)\, dx \;\ge\; (\int \underline{h}(x)\, g(x)\, f(x)\, dx)^{tr}\, M^{-1}\, (\int \underline{h}(x)\, g(x)\, f(x)\, dx)$$

In the context of the Cramér-Rao Information Inequality, as covered in class and in Section 3.4 of the Bickel-Doksum book, $T(X)$ is a scalar-valued statistic with $\psi(\theta) \equiv E_\theta(T(X))$ and $g(x) =$

$T(x) - \psi(\theta)$ and $\underline{h}(x) = \nabla_\theta \log f(x,\theta)$ for the parametric density-family under consideration, with $\theta \in \text{int}(\Theta) \subset \mathbb{R}^d$. The result is:

$$\text{Var}_\theta(T(X)) \geq \nabla_\theta^{tr} E_\theta(T(X)) \{I(\theta)\}^{-1} \nabla_\theta E_\theta(T(X)) \tag{2}$$

In case we want to consider a p-dimensional **vector**-valued statistic $S(X)$, let $\underline{c} \in \mathbb{R}^p \backslash \{\underline{0}\}$ and apply the previous result to $T(x) \equiv \underline{c}' S(x)$ to conclude

$$\underline{c}^{tr} \text{Var}_\theta(S(X)) \underline{c} \geq \nabla_\theta^{tr} E_\theta(\underline{c}^{tr} S(X)) \{I(\theta)\}^{-1} \nabla_\theta E_\theta(\underline{c}^{tr} S(X))$$

Now define $\psi(\theta) = E_\theta(S(X))$ and $J_\theta \psi(\theta) \equiv (\partial \psi_i(\theta)/\partial \theta_j)_{1 \leq i \leq p,\, 1 \leq j \leq d}$. In terms of these notations, the last displayed equation holding for all non-zero $\underline{c} \in \mathbb{R}^p$ says exactly the same as

$$\underline{c}^{tr} \text{Var}_\theta(S(X)) \underline{c} \geq \underline{c}^{tr} J_\theta \psi(\theta) \{I(\theta)\}^{-1} (J_\theta \psi(\theta))^{tr} \underline{c}, \qquad \forall \quad \underline{c} \in \mathbb{R}^p$$

or in the sense of nonnegative-definite ordering of $p \times p$ matrices,

$$\text{Var}_\theta(S(X)) \overset{pd}{\geq} J_\theta \psi(\theta) \{I(\theta)\}^{-1} (J_\theta \psi(\theta))^{tr} \tag{3}$$

Finally, we return to the case of scalar $T(X)$ and consider in detail the necessary and sufficient condition for equality in (2). We begin by stating and proving the condition when $p = d = 1$. In that case, assume continuous differentiability of $f(x,\theta)$ with respect to θ and square integrability of $\nabla_\theta \log f(x,\theta)$ with respect to $f(x,\theta)dx$ for every θ along with the conditions:

$$\textbf{(I)} \qquad C \equiv \{x : f(x,\theta) > 0\} \qquad \text{does not depend on} \quad \theta$$

$$\textbf{(II)} \qquad \nabla_\theta \int_S T(x) f(x,\theta)) \, dx = \int_S T(x) \nabla_\theta f(x,\theta) \, dx \quad \forall \ \theta$$

Then equality holds in the Cauchy-Schwarz inequality (1) for fixed θ if and only if $g(x) = T(x) - \psi(\theta)$ is (almost everywhere) proportional to $h(x) = \nabla_\theta \log f(x,\theta)$, and that proportionality constant can depend on θ, so that for some $k(\theta) > 0$, $\quad k(\theta)(T(x) - \psi(\theta)) = \nabla_\theta \log f(x,\theta)$. It is easy to argue that $k(\theta)$ must also be continuous in θ (since the other terms in the equality are). Therefore, varying t between arbitrary θ_0, θ, we have

$$\log f(x,\theta) - \log f(x,\theta_0) = \int_{\theta_0}^{\theta} k(t) (T(x) - \psi(t)) \, dt = \eta(\theta) T(x) - B(\theta) + \eta(\theta_0) T(x) \tag{4}$$

where $\eta(\theta) = \int_{\theta_0}^{\theta} k(t) dt$, and this form for $\log f(x,\theta)$ is obviously an exponential-family form, with $B(\theta)$ determined by the requirement that the density integrates to 1. *There is a somewhat technical (measure-theoretic) argument needed to show that the exceptional set C_θ of x's (of $f(x,\theta)dx$ probability 0) for which (4) fails to hold can in fact be taken the same for all θ. We omit that argument, which can be found on p. 183 of Bickel & Doksum.*

It remains only to see how the argument generalizes to the case where $T(x)$ is scalar and $\theta \in \Theta \subset \mathbb{R}^d$ is a vector parameter. This is a setting not handled explicitly in Bickel and Doksum. In this case, equality in (1) is equivalent to the statement that for all differentiable one-parameter curves $\theta = \theta(t)$ in $\text{int}(\Theta)$, with $\underline{v} = \theta'(t)$,

$$T(X) \, - \, \psi(\theta(t)) \, \propto (\theta'(t))^{tr} \, \nabla_\theta \log f(X, \theta(t)) \, = \, \frac{d}{dt} \log f(X, \theta(t)) \tag{5}$$

Exactly as in the reasoning leading up to (4), this is equivalent to $\log f(X, \theta(t))$ being of the form $\gamma(t)\, T(X) \, - \, \beta(t) \, + \, q(X)$. And this must be true at $t = 0$ for all different one-parameter curves such that $\theta(0) \, = \, \theta_0$, a particular point in $\text{int}(\Theta)$. From this, we can see that $\gamma(t)$ and therefore $\beta(t)$ must depend only on θ_0. How can the gradients $\theta'(0)$ be related, for all the different smooth curves $\theta(t) \in \Theta \subset \mathbb{R}^d$ when $d > 1$? Consider the special case where $f(x, \theta) = \exp\left(\eta(\theta)^{tr} \, S(x) \, - \, B(\theta)\right) h(x)$, and note that the proportionality (5) leads for each smooth curve through $\theta_0 = \theta(0)$ to

$$T(X) \, - \, \psi(\theta_0) \, = \, k(\theta_0)\, \theta'(0)^{tr} \, S(X)$$

and this can happen only if $k(\theta_0)\, \theta'(0) = \underline{v}$ for a vector $\underline{v}$ not depending on θ_0, with $T(X) = \underline{v}^{tr} \, S(X)$. I have not seen a complete proof of this, but I believe it to be correct: *under suitable smoothness and regularity conditions (involving (I)-(II) and finiteness and nonsingularity of $I(\theta)$), the only way that a nontrivial scalar statistic $T(X)$ can achieve the Cramér-Rao lower variance bound exactly, in finite samples, is for the parametric family $f(x, \theta)$ to be an exponential family (not necessarily canonical), so that $p = \dim(S)$ might be greater than $d = \dim(\Theta)$.*