A GEOMETER'S VIEW OF THE CRAMÉR-RAO BOUND ON ESTIMATOR VARIANCE

ANTHONY D. BLAOM

Abstract. The classical Cramér-Rao inequality gives a lower bound for the variance of an unbiased estimator of an unknown parameter, in some statistical model of a random process. In this note we rewrite the statement and proof of the bound using contemporary geometric language.

The Cramér-Rao inequality gives a lower bound for the variance of a unbiased estimator of a parameter in some statistical model of a random process. Below is a restatement and proof in sympathy with the underlying geometry the problem. While our presentation is mildly novel, its mathematical content is very well-known.

Assuming some very basic familiarity with Riemannian geometry, and that one has reformulated the bound appropriately, the essential parts of the proof boil down to half a dozen lines. For completeness we explain the connection with log-likelihoods, and show how to recover the more usual statement in terms of the Fisher information matrix. We thank Jakob Stroh for helpful feedback.

1. The Cramér-Rao inequality

The mathematical setting of statistical inference consists of: (i) a smooth manifold \( X \), the sample space, which we will suppose is finite-dimensional; and (ii) a set \( \mathcal{P} \) of probability measures on \( X \), called the space of models or parameters. The objective is to make inferences about an unknown model \( p \in \mathcal{P} \), given one or more observations \( x \in X \), drawn at random from \( X \) according to \( p \).

Under certain regularity assumptions detailed below, this data suffices to make \( X \) into a Riemannian manifold, whose geometric properties are related to problems of statistical inference. It seems that Calyampudi Radhakrishna Rao was the first to articulate this connection between geometry and statistics [2].

In formulating the Cramér-Rao inequality, we suppose that \( \mathcal{P} \) is a smooth finite-dimensional manifold (i.e., we are doing so-called parametric inference). We say that \( \mathcal{P} \) is regular if the probability measures \( p \in \mathcal{P} \) are all Borel measures on \( X \), and if there exists some positive Borel measure \( \mu \) on \( X \), hereafter called a reference measure, such that

\[
p = f_p \mu,
\]

for some collection of smooth functions \( f_p, p \in \mathcal{P} \), on \( X \). The definition of regularity furthermore requires that we may arrange \( (x, p) \mapsto f_p(x) \) to be jointly smooth.

\[\text{In this note 'smooth' means } C^2.\]
An unbiased estimator of some smooth function $\theta: \mathcal{P} \to \mathbb{R}$ (the “parameter”) is a smooth function $\hat{\theta}: \mathcal{X} \to \mathbb{R}$ whose expectation under each $p \in \mathcal{P}$ is precisely $\theta(p)$:

$$\theta(p) = \mathbb{E}(\hat{\theta} \mid p) := \int_{x \in \mathcal{X}} \hat{\theta}(x) \, dp; \quad p \in \mathcal{P}. \quad (2)$$

**Theorem** (Rao-Cramér 2 [1]). The space of models $\mathcal{P}$ determines a natural Riemannian metric on $\mathcal{X}$, known as the Fisher-Rao metric, with respect to which there is the following lower bound on the variance of an unbiased estimator $\hat{\theta}$ of $\theta$:

$$\nabla(\hat{\theta} \mid p) \geq |\nabla \theta(p)|^2; \quad p \in \mathcal{P}. \quad (3)$$

More informally: The parameter space $\mathcal{P}$ comes equipped with a natural way of measuring distances, leading to a well-defined notion of steepest rate of ascent, for any function $\theta$ on $\mathcal{P}$. The square of this rate is precisely the lower bound for the variance of an unbiased estimator $\hat{\theta}$.

2. Observation-dependent one-forms on the space of models

It is fundamental to the present geometric point of view that each observation $x \in \mathcal{X}$ determines a one-form $\lambda_x$ on the space $\mathcal{P}$ of models in the following way: Let $v \in T_{p_0}\mathcal{P}$ be a tangent vector, understood as the derivative of some path $t \mapsto p_t \in \mathcal{P}$ through $p_0$:

$$v = \frac{d}{dt} p_t \bigg|_{t=0}. \quad (4)$$

Then, recalling that each $p_t$ is a probability measure on $\mathcal{X}$ (and $\mathcal{P}$ is regular) we may write $p_t = g_t p_0$, for some smooth function $g_t: \mathcal{X} \to \mathbb{R}$, and define

$$\lambda_x(v) = \frac{d}{dt} g_t(x) \bigg|_{t=0}. \quad \text{Lemma. } \mathbb{E}(\lambda_x(v) \mid p) = 0 \text{ for all } p \in \mathcal{P} \text{ and } v \in T_p\mathcal{P}.$$

Now if $v \in T_{p_0}\mathcal{P}$ is a tangent vector as in (4), and if (2) holds, then

$$d\theta(v) = \frac{d}{dt} \int_{x \in \mathcal{X}} \hat{\theta}(x) \, dp_t \bigg|_{t=0} = \frac{d}{dt} \int_{x \in \mathcal{X}} \hat{\theta}(x) g_t(x) \, dp_0 \bigg|_{t=0} = \int_{x \in \mathcal{X}} \hat{\theta}(x) \lambda_x(v) \, dp_0,$$

giving us:

**Proposition.** For any unbiased estimator $\hat{\theta}: \mathcal{X} \to \mathbb{R}$ of $\theta: \mathcal{P} \to \mathbb{R}$, one has

$$d\theta(v) = \int_{x \in \mathcal{X}} \hat{\theta}(x) \lambda_x(v) \, dp; \quad v \in T_{p_0}\mathcal{P}.$$
3. Log-likelihoods

As an aside, we shall now see that the observation-dependent one-forms $\lambda_x$ are exact, and at the same time give their more usual interpretation in terms of log-likelihoods.

Choosing a reference measure $\mu$, and defining $f_p$ as in (1), one defines the log-likelihood function $(x, p) \mapsto L_x(p): \mathcal{X} \times \mathcal{P} \to \mathbb{R}$ by

$$L_x(p) = \log f_p(x).$$

While the log-likelihood depends on the reference measure $\mu$, its derivative $dL_x$ (a one-form on $\mathcal{P}$) does not, for in fact:

**Lemma.** $dL_x = \lambda_x$.

**Proof.** With a reference measure fixed as in (1), we have, along a path $t \mapsto p_t$, $p_t = g_t p_0$, where $g_t = f_{p_t}/f_{p_0}$. Applying the definition of $\lambda_x$, we compute

$$\lambda_x \left( \frac{d}{dt} p_t \bigg|_{t=0} \right) = \frac{d}{dt} f_{p_t}(x) \bigg|_{t=0} = \frac{d}{dt} \left( \frac{e^{L_x(p_t)}}{e^{L_x(p_0)}} \right) \bigg|_{t=0} = dL_x \left( \frac{d}{dt} p_t \bigg|_{t=0} \right).$$

$\square$

In particular, local maxima of $L_x$ (points of so-called maximum likelihood) do not depend on the reference measure.

4. The metric and derivation of the bound

With the observation-dependent one-forms in hand, we may now define the Fisher-Rao Riemannian metric on $\mathcal{P}$. It is given by

$$\mathbb{I}(u, v) = \int_{x \in \mathcal{X}} \lambda_x(u) \lambda_x(v) \, dp, \text{ for } u, v \in T_p \mathcal{P}.$$

Now that we have a metric, it is natural to consider $\nabla \theta$ instead of $d\theta$ in Proposition 2. By the definition of gradient, we have

$$|\nabla \theta(p)|^2 = d\theta (\nabla \theta(p)).$$

This equation and Proposition 2 now gives, for any $v \in T_p \mathcal{P},$

$$|\nabla \theta(p)|^2 = \int_{x \in \mathcal{X}} \hat{\theta}(x) \lambda_x(\nabla \theta(p)) \, dp = \int (\hat{\theta}(x) - \theta(p)) \lambda_x(\nabla \theta(p)) \, dp.$$

The second equality holds because $\int_{x \in \mathcal{X}} \lambda_x(\nabla \theta(p)) \, dp = 0$, by Lemma 2. Applying the Cauchy-Schwartz inequality to the right-hand side gives

$$|\nabla \theta(p)|^2 \leq \left( \int_{x \in \mathcal{X}} (\hat{\theta}(x) - \theta(p))^2 \, dp \right)^{1/2} \left( \int_{x \in \mathcal{X}} \lambda_x(\nabla \theta(p)) \lambda_x(\nabla \theta(p)) \, dp \right)^{1/2}$$

$$= \sqrt{\mathbb{V}(\hat{\theta} | p)} \sqrt{\mathbb{I}(\nabla \theta(p), \nabla \theta(p))} = \sqrt{\mathbb{V}(\hat{\theta} | p)} |\nabla \theta(p)|.$$

The Cramér-Rao bound now follows.
5. The bound in terms of Fisher information

Theorem 1 is coordinate-free formulation. To recover the more usual statement of the Cramér-Rao bound, let \( \phi_1, \ldots, \phi_k \) be local coordinates on \( \mathcal{P} \), the space of models on \( \mathcal{X} \), and \( \frac{\partial}{\partial \phi_1}, \ldots, \frac{\partial}{\partial \phi_k} \) the corresponding vector fields on \( \mathcal{P} \), characterised by

\[
d\phi_i \left( \frac{\partial}{\partial \phi_j} \right) = \delta_i^j.
\]

Here \( \delta_i^j = 1 \) if \( i = j \) and is zero otherwise. Applying Lemma 3, the coordinate representation \( I_{ij} \) of the Fisher-Rao metric \( \mathbb{I} \) is given by

\[
I_{ij}(p) = \mathbb{I} \left( \frac{\partial}{\partial \phi_i}(p), \frac{\partial}{\partial \phi_j}(p) \right) = \int_{x \in \mathcal{X}} d\lambda_x \left( \frac{\partial}{\partial \phi_i}(p) \right) d\lambda_x \left( \frac{\partial}{\partial \phi_j}(p) \right) dp
\]

\[
= \int_{x \in \mathcal{X}} \left( \frac{\partial L_x}{\partial \phi_i}(p) \right) \left( \frac{\partial L_x}{\partial \phi_j}(p) \right) dp,
\]

where \( L_x(p) = \log f_p(x) \) is the log-likelihood. In statistics \( I_{ij} \) is known as the Fisher information matrix.

For the moment we continue to let \( \theta \) denote an arbitrary function on \( \mathcal{P} \), and \( \hat{\theta} \) an unbiased estimate. Now \( \nabla \theta \) is the gradient of \( \theta \), with respect to the metric \( \mathbb{I} \). Since the coordinate representation of the metric is \( I_{ij} \), a standard computation gives the local coordinate formula

\[
\nabla \theta = \sum_{i,j} I^{ij} \frac{\partial \theta}{\partial \phi_i} \frac{\partial}{\partial \phi_j},
\]

where \( \{I^{ij}\} \) is the inverse of \( \{I_{ij}\} \). Regarding the lower bound in Theorem 1 we compute

\[
|\nabla \theta(p)|^2 = \mathbb{I}(\nabla \theta(p), \nabla \theta(p)) = \sum_{i,j,m,n} \mathbb{I} \left( I^{ij}(p) \frac{\partial \theta}{\partial \phi_i}(p), I^{mn}(p) \frac{\partial \theta}{\partial \phi_m}(p) \right)
\]

\[
= \sum_{i,j,m,n} I^{ij}(p) I^{mn}(p) \frac{\partial \theta}{\partial \phi_i}(p) \frac{\partial \theta}{\partial \phi_m}(p) \mathbb{I} \left( \frac{\partial}{\partial \phi_j}(p), \frac{\partial}{\partial \phi_n}(p) \right)
\]

\[
= \sum_{i,j,m,n} I^{ij}(p) I^{jn}(p) I^{mn}(p) \frac{\partial \theta}{\partial \phi_i}(p) \frac{\partial \theta}{\partial \phi_m}(p)
\]

\[
= \sum_{i,j,m,n} \delta_i^j \delta^m_n I^{mn}(p) \frac{\partial \theta}{\partial \phi_i}(p) \frac{\partial \theta}{\partial \phi_m}(p) = \sum_{i,m} I^{mi}(p) \frac{\partial \theta}{\partial \phi_i}(p) \frac{\partial \theta}{\partial \phi_m}(p).
\]

Theorem 1 now reads

\[
\nabla(\hat{\theta} \mid p) \geq \sum_{i,m} I^{mi}(p) \frac{\partial \theta}{\partial \phi_i}(p) \frac{\partial \theta}{\partial \phi_m}(p).
\]
In particular, if we suppose $\theta$ is one of the coordinate functions, say $\theta = \phi_j$, then we obtain
\[
\mathbb{V}(\hat{\phi}_j | p) \geq \sum_{i,m} I^{mi}(p) \frac{\partial \phi_j}{\partial \phi_i}(p) \frac{\partial \phi_j}{\partial \phi_m}(p) = \sum_{i,m} I^{mi} \delta_j^i \delta_j^m = I^{jj}(p),
\]
the version of the Cramér-Rao bound to be found in statistics textbooks.

References

[1] Harald Cramér. Mathematical Methods of Statistics. Princeton Mathematical Series, vol. 9. Princeton University Press, Princeton, N. J., 1946.

[2] C. Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc., 37:81–91, 1945.

E-MAIL: anthony.blaom@gmail.com