STAT0008 Lecture 4
The Cramér-Rao Lower Bound and the Exponential Family of Distributions

Dr. Aidan O'Keeffe

Department of Statistical Science
University College London

22nd October 2018

Outline

- Score function – linear form and its relationship to the Cramér-Rao lower bound
- Minimum variance bound estimators (MVBEs)
- Minimum variance bound unbiased estimators (MVBUEs)
- Fisher information and the Cramér-Rao lower bound for k-dimensional parameters
- The exponential family of distributions

Cramér-Rao Lower Bound

Recall, from Lecture 3, the Cramér-Rao lower bound...

If T is an unbiased estimator for $m(\theta)$, then the variance of T is such that

$$\mathsf{Var}(T) \geq \frac{[m'(\theta)]^2}{\mathcal{I}(\theta)}.$$

Also, from Lecture 3, the inequality above was derived from the assumption that the absolute value of the Pearson correlation between T and the score function $U(\theta; \mathbf{X})$ is ≤ 1.

$$\text{i.e.} \quad |\mathsf{Corr}(T, U(\theta; \mathbf{X}))| \leq 1.$$

Cramér-Rao Lower Bound

We obtain equality where

$$\text{Corr}(T, U(\theta; \mathbf{X})) = 1.$$

In other words, the Cramér-Rao lower bound will be attained where the Pearson correlation between T and $U(\theta; \mathbf{X})$ is equal to 1.

A Pearson correlation of 1 means that there is a linear, deterministic relationship between T and $U(\theta; \mathbf{X})$.

We can show that the variance of T attains the Cramér-Rao lower bound if and only if T and $U(\theta; \mathbf{X})$ are linearly related such that

$$U(\theta; \mathbf{X}) = A(\theta)\,(T(\mathbf{X}) - m(\theta))$$

where $A(\theta)$ is a non-random function of θ.

Score Function: Linear Form

We see that, where the Cramér-Rao lower bound is attained by Var$(T(\mathbf{X}))$, the score function $U(\theta; \mathbf{X})$ can be written

$$U(\theta; \mathbf{X}) = A(\theta)\,(T(\mathbf{X}) - m(\theta))$$

$U(\theta; \mathbf{X})$, the score function, contains $\mathbf{X}$ and θ.

$A(\theta)$ is a non-random function of θ. i.e. it has θ as an argument but **not** $\mathbf{X}$.

$T(\mathbf{X})$ is a statistic that is a function of $\mathbf{X}$ only and does not contain θ.

$m(\theta)$ is a function of θ only and does not contain $\mathbf{X}$.

Note that the linear form must be as above so that $\mathbb{E}(U(\theta; \mathbf{X})) = 0$ holds.

Score Function: Linear Form

Why is this functional form useful...?

Well...if we can write the score function in this way we can:

1. Obtain an estimator of $m(\theta)$ that attains the Cramér-Rao lower bound (i.e. has the smallest possible variance).
2. Identify the Cramér-Rao lower bound (and Fisher information).

Point 1. is obvious. To verify Point 2., let us go back to the linear form of $U(\theta; \mathbf{X})$.

Score Function: Linear Form

Set

$$U(\theta; \mathbf{X}) = A(\theta)\,(T(\mathbf{X}) - m(\theta))$$

Score Function: Linear Form

Score Function: Linear Form

Hence, we see that

$$A(\theta) = \frac{\mathcal{I}(\theta)}{m'(\theta)}.$$

So if we write the score function in the form

$$\begin{aligned} U(\theta; \mathbf{X}) &= A(\theta)\left(T(\mathbf{X}) - m(\theta)\right) \\ &= \frac{\mathcal{I}(\theta)}{m'(\theta)}\left(T(\mathbf{X}) - m(\theta)\right). \end{aligned}$$

then we can easily recover the Fisher information and Cramér-Rao lower bound.

MVBE and MVBUE

Suppose that T is an estimator of $m(\theta)$.

If Var(T) attains the Cramér-Rao lower bound, then T is known as a **minimum variance bound estimator**, with acronym **MVBE**.

If T is **unbiased** for $m(\theta)$ **and** Var(T) attains the Cramér-Rao lower bound, then T is known as a **minimum variance bound unbiased estimator**, with acronym **MVBUE**.

Given a likelihood function, $\mathcal{L}(\theta \mid \mathbf{X})$, we can use the following steps to attempt to identify MVBUEs.

MVBE and MVBUE

1. Write the score function $U(\theta; \mathbf{X}) = \frac{\partial}{\partial\theta} \log \mathcal{L}(\theta \mid \mathbf{X})$ in the form
$$U(\theta; \mathbf{X}) = A(\theta)\left(T(\mathbf{X}) - m(\theta)\right).$$
2. Then $T(\mathbf{X})$ is unbiased for $m(\theta)$ and $\mathrm{Var}(T(\mathbf{X}))$ attains the Cramér-Rao lower bound.
3. The Cramér-Rao lower bound $\left(\frac{[m'(\theta)]^2}{\mathcal{I}(\theta)}\right)$ can be easily identified using the equation
$$A(\theta) = \frac{\mathcal{I}(\theta)}{m'(\theta)} \implies \mathsf{CRLB} = \frac{m'(\theta)}{A(\theta)}.$$

T unbiased for θ

Note that when $T(\mathbf{X})$ is unbiased for θ, we have $m(\theta) = \theta$ and $m'(\theta) = 1$ so

$$U(\theta; \mathbf{X}) = \mathcal{I}(\theta)\left(T(\mathbf{X}) - \theta\right).$$

In this case, the Cramér-Rao lower bound is

$$\text{CRLB} = \frac{1}{\mathcal{I}(\theta)}.$$

MVBUE: Example 1

Suppose that $X_1, \ldots, X_n$ are a set of iid random variables where each $X_i \sim \text{Bernoulli}(\theta)$. Find a MVBUE for θ and determine the variance of this estimator.

MVBUE: Example 1

MVBUE: Example 2

Suppose that $X_1, \ldots, X_n$ are a set of iid random variables where each $X_i \sim \text{Poi}(\lambda)$. Find a MVBUE for λ and determine the variance of this estimator.

MVBUE: Example 2

$k-$dimensional Parameter

Suppose now that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\intercal$ is a parameter of dimension k (i.e. a $k \times 1$ vector).

We define a sample $X_1, \ldots, X_n$ such that each $X_i \sim \mathcal{D}(\boldsymbol{\theta})$, with corresponding likelihood function $\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{X})$. The log-likelihood function is

$$\ell(\boldsymbol{\theta} \mid \mathbf{X}) = \log[\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{X})].$$

The score function takes the form of a $k \times 1$ vector, $U(\boldsymbol{\theta}; \mathbf{X})$, where

$$U(\boldsymbol{\theta}; \mathbf{X}) = \begin{pmatrix} \frac{\partial \ell(\boldsymbol{\theta}|\mathbf{X})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\theta}|\mathbf{X})}{\partial \theta_k} \end{pmatrix}$$

$k-$dimensional Parameter

The score function is such that

$$\mathbb{E}\left[U(\boldsymbol{\theta};\mathbf{X})\right] = \begin{pmatrix} \mathbb{E}\left[\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{X})}{\partial \theta_1}\right] \\ \vdots \\ \mathbb{E}\left[\frac{\partial \ell(\boldsymbol{\theta}|\mathbf{X})}{\partial \theta_k}\right] \end{pmatrix} = \mathbf{0}.$$

As in the one-dimensional parameter case, the Fisher information is computed as the variance of the score function.

However, since $\boldsymbol{\theta}$ is a k-dimensional vector, the variance of the score function is a $k \times k$ variance-covariance matrix, known as the **Fisher information matrix**.

$k-$dimensional Parameter

The Fisher information matrix is

$$
\begin{aligned}
\mathcal{I}(\boldsymbol{\theta}) &= \mathsf{Var}(U(\boldsymbol{\theta};\mathbf{X})) \\
&= \mathbb{E}\left[U(\boldsymbol{\theta};\mathbf{X})U(\boldsymbol{\theta};\mathbf{X})^{\mathsf{T}}\right] \\
&= \begin{pmatrix} \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta_1}\right)^2\right] & \dots & \mathbb{E}\left[\frac{\partial \ell}{\partial \theta_1}\frac{\partial \ell}{\partial \theta_k}\right] \\ \vdots & \ddots & \vdots \\ \mathbb{E}\left[\frac{\partial \ell}{\partial \theta_k}\frac{\partial \ell}{\partial \theta_1}\right] & \dots & \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta_k}\right)^2\right] \end{pmatrix}
\end{aligned}
$$

where, for brevity, $\ell = \ell(\boldsymbol{\theta} \mid \mathbf{X})$. We note that the $(r,s)^{\text{th}}$ element of $\mathcal{I}(\boldsymbol{\theta})$ is

$$
[\mathcal{I}(\boldsymbol{\theta})]_{rs} = \mathbb{E}\left[\frac{\partial \ell(\boldsymbol{\theta} \mid \mathbf{X})}{\partial \theta_r}\frac{\partial \ell(\boldsymbol{\theta} \mid \mathbf{X})}{\partial \theta_s}\right].
$$

$k-$dimensional Parameter

Under certain regularity conditions, the following result holds

$$[\mathcal{I}(\boldsymbol{\theta})]_{rs} = \mathbb{E}\left[-\frac{\partial^2 \ell(\boldsymbol{\theta} \mid \mathbf{X})}{\partial\theta_r \partial\theta_s}\right].$$

If $T(\mathbf{X})$ is an estimator for the $k-$dimensional vector, $\boldsymbol{\theta}$, we can establish the Cramér-Rao lower bound for the variance-covariance matrix of Var$(T(\mathbf{X}))$.

Unlike in the one-dimensional case, we shall concentrate only on unbiased estimators of $\boldsymbol{\theta}$.

Suppose that $T(\mathbf{X})$ is an estimator (in the form of a $k \times 1$ random variable) of the k-dimensional parameter $\boldsymbol{\theta}$. Then the variance-covariance matrix of $T(\mathbf{X})$ is such that

$$\text{Var}(T(\mathbf{X})) \geq \mathcal{I}^{-1}(\boldsymbol{\theta}).$$

Here, the matrix inequality $\geq$ is such that $A \geq B$ implies that $A - B$ is positive semi-definite for square matrices A and B of the same dimension.

$k-$dimensional Parameter: Cramér-Rao lower bound

Moreover, if $T(\mathbf{X})$ is such that $T_j(\mathbf{X})$ is an unbiased estimator for θ_j (for some $j \in \{1, \ldots, k\}$) then the Cramér-Rao lower bound for the variance of $T_j(\mathbf{X})$ is

$$\mathsf{Var}(T_j(\mathbf{X})) \geq \left[\mathcal{I}^{-1}(\boldsymbol{\theta})\right]_{jj}.$$

Here, $T_j(\mathbf{X})$ is the j^{th} component of $T(\mathbf{X}) = (T_1(\mathbf{X}), \ldots, T_k(\mathbf{X}))^T$ and θ_j is the j^{th} component of $\boldsymbol{\theta}$.

$\left[\mathcal{I}^{-1}(\boldsymbol{\theta})\right]_{jj}$ is the $(j,j)^{\text{th}}$ element of the inverse of the Fisher information matrix.

The condition for equality is that $T_j(\mathbf{X})$ is a linear function of $U_1(\boldsymbol{\theta}; \mathbf{X}), \ldots, U_k(\boldsymbol{\theta}; \mathbf{X})$ where $U_j(\boldsymbol{\theta}; \mathbf{X})$ denotes the j^{th} component of the $k \times 1$ score function $U(\boldsymbol{\theta}; \mathbf{X})$ ($j \in \{1, \ldots k\}$).

$k-$dimensional Parameter: Example

Suppose that $X_1, \dots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$ random variables.

(a) Compute the Fisher information matrix for the parameters $(\mu, \sigma^2)^\mathsf{T}$.

(b) Hence

  (i) Determine the Cramér-Rao lower bound for the variance of unbiased estimators of μ

  (ii) Determine the Cramér-Rao lower bound for the variance of unbiased estimators of σ^2

$k-$dimensional Parameter: Example

$k-$dimensional Parameter: Example

$k-$dimensional Parameter: Example

Exponential Family

We have seen that, if an estimator $T(\mathbf{X})$ attains the Cramér-Rao lower bound, then the score function is of the form

$$\begin{aligned} U(\theta; \mathbf{X}) &= A(\theta)(T(\mathbf{X}) - m(\theta)) \\ &= A(\theta)T(\mathbf{X}) + B(\theta) \ \text{ where } \ B(\theta) = -A(\theta)m(\theta). \end{aligned}$$

Then

$$\frac{\partial}{\partial\theta} \log \mathcal{L}(\theta \mid \mathbf{X}) = A(\theta)T(\mathbf{X}) + B(\theta)$$

and integrating the above with respect to θ, we obtain

$$\log \mathcal{L}(\theta \mid \mathbf{X}) = a(\theta)T(\mathbf{X}) + b(\theta) + c(\mathbf{X})$$

with

$$a(\theta) = \int_\Theta A(\theta)\mathrm{d}\theta, \ \ b(\theta) = \int_\Theta B(\theta)\mathrm{d}\theta.$$

Exponential Family

Hence, the likelihood function is

$$\begin{aligned}\mathcal{L}(\theta \mid \mathbf{X}) &= \exp\left\{a(\theta)T(\mathbf{X}) + b(\theta) + c(\mathbf{X})\right\} \\ &= \left[\exp\left\{a(\theta)T(\mathbf{X}) + b(\theta)\right\}\right]\left[\exp\left\{c(\mathbf{X})\right\}\right]\end{aligned}$$

and, by the factorisation criterion, we see that $T(\mathbf{X})$ is sufficient for θ.

Recall that the likelihood function, $\mathcal{L}(\theta \mid \mathbf{X})$ is the joint probability density/mass function of $\mathbf{X}$ where θ is a distributional parameter.

Thus, the joint density/mass $f(\mathbf{x};\theta)$ is such that

$$f(\mathbf{x};\theta) = \exp\left\{a(\theta)T(\mathbf{x}) + b(\theta) + c(\mathbf{x})\right\}.$$

Exponential Family

A family of probability distributions whose density/mass function has the form

$$f(x;\theta) = \exp\left\{a(\theta)T(x) + b(\theta) + c(x)\right\}$$

for θ in some parameter space Θ and $X \sim \mathcal{D}(\theta)$, is said to be an **exponential family** of distributions.

We note that the support of $f(x;\theta)$ (i.e. the values of x for which $f(x;\theta)$ is defined) must not depend on θ.

In addition, $c(x)$ is not arbitrary because the pdf/pmf, $f(x;\theta)$, must integrate/sum to 1.

Exponential Family

If $\mathbf{X} = \{X_1, \ldots, X_n\}$ are independent random variables from the same univariate exponential family then the likelihood function is given by

$$\mathcal{L}(\theta \mid \mathbf{x}) = \exp\left\{a(\theta)T(\mathbf{x}) + nb(\theta) + c(\mathbf{x})\right\}$$

where

$$T(\mathbf{x}) = \sum_{i=1}^{n} T(x_i) \quad \text{and} \quad c(\mathbf{x}) = \sum_{i=1}^{n} c(x_i).$$

Notice that the score function of the above implies that $\frac{T(\mathbf{X})}{n}$ is a MVBUE for $-\frac{b'(\theta)}{a'(\theta)}$.

Exponential Family: k-dimensional parameter

Suppose now that $\boldsymbol{\theta}$ is a k-dimensional parameter $(k > 1)$ where

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\mathsf{T} .$$

A family of probability distributions $\{f(x; \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ belongs to the exponential family of distributions if $f(x; \boldsymbol{\theta})$ has the form

$$f(x; \boldsymbol{\theta}) = \exp\left\{\sum_{r=1}^{k} a_r(\boldsymbol{\theta}) T_r(x) + b(\boldsymbol{\theta}) + c(x)\right\}.$$

As in the one-dimensional parameter case, the support of $f(x; \boldsymbol{\theta})$ must not depend on $\boldsymbol{\theta}$.

Exponential Family: k-dimensional parameter

If $\mathbf{X} = \{X_1, \ldots, X_n\}$ are independent random variables from the an exponential family parameterised by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\mathsf{T} \in \Theta$ then the joint density of $\mathbf{X}$ is given by

$$f(\mathbf{x}; \boldsymbol{\theta}) = \exp\left\{\sum_{r=1}^{k} a_r(\boldsymbol{\theta}) T_r(\mathbf{x}) + n b(\boldsymbol{\theta}) + c(\mathbf{x})\right\}$$

where

$$T_r(\mathbf{x}) = \sum_{i=1}^{n} T_r(x_i) \quad \text{and} \quad c(\mathbf{x}) = \sum_{i=1}^{n} c(x_i).$$

It follows from the factorisation criterion that

$$\{T_1(\mathbf{X}), \ldots, T_k(\mathbf{X})\}$$

are jointly sufficient for $\boldsymbol{\theta}$.

Exponential Family: k-dimensional parameter

Here, where the number jointly sufficient statistics is equal to the dimension of $\boldsymbol{\theta}$, the jointly sufficient statistics $T_1(\mathbf{X}), \ldots, T_k(\mathbf{X})$ may be expressed as functions of

$$U_j(\boldsymbol{\theta}; \mathbf{X}) = \frac{\partial \ell(\boldsymbol{\theta}; \mathbf{X})}{\partial \theta_j} \quad j = 1, \ldots, k.$$

Hence, estimators that are linear functions of $T_1(\mathbf{X}), \ldots, T_k(\mathbf{X})$ will have variances that attain the corresponding Cramér-Rao lower bound.

Exponential Family

At this point, we might ask '*why is an exponential family important?*'

In answer

- If a distribution is from an exponential family, then sufficient statistics for the distributional parameters are guaranteed to exist.
- In a Bayesian framework, if a likelihood function/joint density belongs to an exponential family, then there exists a **conjugate prior** distribution for the distributional parameters.

Conjugate Prior

A conjugate prior is a prior whereby, when combined with a given likelihood function, the posterior distribution belongs to the same family as the prior.

For example, if we have a likelihood function from a Bin(θ) distribution and we choose a Beta prior for θ then we will obtain a Beta distribution for the posterior of θ.

This is a situation where the prior is a conjugate prior.

Exponential Families

Many familiar classes of distributions belong to the exponential family of distributions. For example

- Binomial, Poisson;
- Exponential, gamma;
- Normal, Beta.

Exponential Family: Example 1

Suppose that $X \sim \text{Poi}(\theta)$. Show that X belongs to the exponential family of distributions and identify the functions $T(x)$, $a(\theta)$, $b(\theta)$ and $c(x)$.

Exponential Family: Example 2

Suppose that $X \sim \text{Beta}(\alpha, \beta)$. Show that X belongs to the exponential family of distributions and identify the functions $T(x)$, $a(\boldsymbol{\theta})$, $b(\boldsymbol{\theta})$ and $c(x)$, where $\boldsymbol{\theta} = (\alpha, \beta)^{\mathsf{T}}$. (Note: $T(x)$ and $a(\boldsymbol{\theta})$ have dimension 2)

Exponential Family: Example 2

Learning Outcomes

- Understand how to identify the Fisher information and CRLB when the score function has been written in a linear form.
- Understand what is meant by Minimum variance bound estimators (MVBEs) and minimum variance bound unbiased estimators (MVBUEs), and how to determine such estimators.
- Learn how to calculate the Fisher information matrix and Cramér lower bounds for $k-$dimensional parameters.
- Understand what is meant by the exponential family of distributions and learn how to determine whether or not a given distribution belongs to the exponential family.