

STATS300A - Lecture 5

Dominik Rothenhaeusler
Scribed by Michael Howes

10/04/21

Contents

1	Recap	1
2	An example from semi-parametrisation	1
3	Orthogonality	2
4	Cramer-Rao lower bound (CRLB)	2
5	Equivariance	3

1 Recap

We have the following techniques for finding optimal estimators.

- (a) Conditioning and using Rao-Blackwellisation.
- (b) Solving $\mathbb{E}_\theta \delta(T) = g(\theta)$ for δ .
- (c) Guessing.
- (d) Orthogonality constraints (to be discussed today).

2 An example from semi-parametrisation

Semi-parametrisation refers to the set up where θ is a finite dimensional parameter of interest but our set of measures \mathcal{P} is infinite dimensional. The following example is semi-parametric.

$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F \in \mathcal{F}$ where \mathcal{F} is the collection all cdfs which are symmetric around some $\theta \in \mathbb{R}$ and have finite second moment. The parameter θ is a function of $F \in \mathcal{F}$ and $\theta = \mathbb{E}_F[X_i]$. We wish to estimate θ from X_1, \dots, X_n . One can ask, does a UMVUE exist for θ ? Suppose one does and call it T .

- (a) Consider the submodel $\{N(\theta, 1) : \theta \in \mathbb{R}\}$, then we know that \bar{X}_n is the UMVUE for θ . This estimator is also unbiased on the full model \mathcal{F} .
- (b) The risk of T and \bar{X}_n must be equal on the submodel since they are both UMVUE on the submodel.
- (c) Since \bar{X}_n is the unique UMVUE on the submodel we must have $T = \bar{X}_n$.

- (d) Repeat (a)-(c) for the new submodel $\{\text{Unif}[\theta - 1, \theta + 1] : \theta \in \mathbb{R}\}$. The UMVUE for this model is also unique by completeness and it does not equal \bar{X}_n (see homework for a calculation of this estimator). This estimator is again unbiased on the whole model.
- (e) This gives us a contradiction since T cannot be equal to the two different UMVUEs.

3 Orthogonality

Suppose δ_i is a UMVUE for $g_i(\theta)$. Can we conclude that $\sum_i \delta_i$ is UMVUE for $\sum_i g_i(\theta)$?

Definition 1. Define the set Δ as follows $\Delta = \{\delta(X) : \mathbb{E}_\theta(\delta(X)^2) < \infty, \text{ for all } \theta\}$.

Theorem 1. [TPE 2.17] $\delta_0 \in \Delta$ is the UMVUE for $g(\theta) = \mathbb{E}_\theta \delta_0(X)$ if and only if $\mathbb{E}_\theta \delta_0(X)U = 0$ for all θ and all $U \in \Delta$ such that $\mathbb{E}_\theta U = 0$.

Proof. See scribed notes. □

We can now answer our question with a yes! If each δ_i is the UMVUE for $g_i(\theta)$, then $\mathbb{E}_\theta[\delta_i(X)U] = 0$ for all first order ancillary U . Furthermore $\mathbb{E}_\theta[\sum_i \delta_i(X)] = \sum_i g_i(\theta)$ and

$$\mathbb{E}_\theta[\sum_i \delta_i(X)U] = \sum_i \mathbb{E}_\theta[\delta_i(X)U] = 0.$$

Thus $\sum_i \delta_i(X)$ is the UMVUE for $\sum_i g_i(\theta)$.

4 Cramer-Rao lower bound (CRLB)

Definition 2. We define the *log likelihood* of a density $p(x; \theta)$ to be

$$l(x; \theta) = \log p(x; \theta).$$

For this definition we require $p(x; \theta) > 0$ for all x and θ . We also define the *score* or *score function* to be

$$S(x, \theta) = \partial_\theta l(x; \theta).$$

Note that

$$p(x; \theta_0 + \varepsilon) = p(x; \theta_0) \exp \{ \varepsilon S(x, \theta_0) + o(\varepsilon) \}.$$

Thus $p(x; \theta_0 + \varepsilon)$ “looks like” an exponential family with parameter ε and sufficient statistic $S(x, \theta_0)$.

Theorem 2. [CRLB - Keener Thrm 4.9] Let $p(x; \theta)$ be densities with $p(x; \theta) > 0$ for all x, θ and such that $p(x; \theta)$ is differentiable in θ . Suppose furthermore that for some function g

$$(a) \quad \mathbb{E}_\theta[S(X, \theta)] = 0.$$

$$(b) \quad \mathbb{E}_\theta[S(X, \theta)\delta(X)] = g'(\theta).$$

Then

$$\text{Var}_\theta(\delta) \geq \frac{g'(\theta)^2}{I(\theta)},$$

where $I(\theta)$ is equal to

$$I(\theta) = \mathbb{E}_\theta[S(X, \theta)^2],$$

and called the Fisher information.

Some remarks on our two conditions. If $\delta(X)$ is unbiased for $g(\theta)$, then under some regularity conditions

$$\begin{aligned} g'(\theta) &= \frac{d}{d\theta} \mathbb{E}_\theta[\delta(X)] \\ &= \frac{d}{d\theta} \int p(x; \theta) \delta(x) d\mu(x) \\ &= \int \frac{d}{d\theta} p(x; \theta) \delta(x) d\mu(x) \\ &= \int \frac{\frac{d}{d\theta} p(x; \theta)}{p(x; \theta)} \delta(x) p(x; \theta) d\mu(x) \\ &= \int S(x, \theta) \delta(x) p(x; \theta) d\mu(x) \\ &= \mathbb{E}_\theta[S(X, \theta) \delta(X)]. \end{aligned}$$

Thus condition (b) is equivalent to regularity plus unbiased. Condition (a) is equivalent to a regularity condition on $p(x; \theta)$ and can be seen by taking $\delta(X) = 1$ and applying what we have done above.

We will now prove the CRLB.

Proof. By Cauchy-Schwarz

$$\begin{aligned} |g'(\theta)| &= |\mathbb{E}_\theta[\delta(X) S(X; \theta)]| \\ &= |\text{Cov}_\theta(S(X; \theta), \delta(X))| \quad \text{since } \mathbb{E}_\theta[S(X; \theta)] = 0. \\ &\leq \sqrt{\text{Var}_\theta(S(X; \theta)) \text{Var}_\theta(\delta(X))}. \end{aligned}$$

Squaring and dividing by $I(\theta) = \text{Var}_\theta(S(X; \theta))$ gives $\text{Var}_\theta(\delta(X)) \geq \frac{g'(\theta)^2}{I(\theta)}$. □

Another remark, if $\int \partial_\theta^2 p(x; \theta) d\mu(x) = \partial_\theta^2 \int p(x; \theta) d\mu(x) = 0$, then

$$I(\theta) = -\mathbb{E}_\theta[\partial_\theta^2 l(x; \theta)].$$

Thus we can think of $I(\theta)$ as a measure of curvature. Consider two cases

- (a) Small changes in θ result in large changes in $l(x; \theta)$ (high Fisher information).
- (b) Small changes in θ result in small changes in $l(x; \theta)$ (low Fisher information).

We want (a) when we are making inferences about θ . Small changes in θ will result in large changes in the distribution of our data. Thus we can make precise statements about θ based on our data.

Example 1. Suppose $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} p(x; \theta)$. Then $p(x; \theta) = \prod_{i=1}^n p(x_i; \theta)$ and so $S(x_1, \dots, x_n; \theta) = \sum_{i=1}^n S(x_i; \theta)$ and furthermore, by our iid assumption,

$$I_n(\theta) = \text{Var}_\theta(S(X_1, \dots, X_n); \theta) = n \text{Var}_\theta(S(X_1); \theta) = nI(\theta).$$

This is one indication for why lower bounds scale at a rate of $\frac{1}{n}$ (under our regularity assumptions).

There is another example in the scribed notes that relates to a Gaussian model.

5 Equivariance

We are done with unbiasedness and now we will look at restricting our estimators to respect certain symmetries. Consider the location model $X_1, \dots, X_n \sim f_\theta(x)$ where f is a known pdf, $\theta \in \mathbb{R}$ is unknown and $f_\theta(x) = f(x_1 - \theta, \dots, x_n - \theta)$. A special case of this is when X_i are iid and thus $f_\theta(x) = \prod_{i=1}^n g(x_i - \theta)$ for some g .

Definition 3. A model is called *location invariant* if

$$f_{\theta+c}(x+c) = f_{\theta}(x),$$

for all θ, x and c .

Definition 4. A loss function is called *location invariant* if

$$L(\theta + c, d + c) = L(\theta, d),$$

for all θ, d and c .

Note that squared error loss $L(\theta, d) = (\theta - d)^2$ is location invariant as is any other loss that is a function of $\theta - d$. In fact these are the only location invariant losses. Since if L is location invariant then $L(\theta, d) = L(\theta - d, 0) =: \rho(\theta - d)$.

Definition 5. A decision problem is *location invariant* if the model and the loss function are both location invariant.

Definition 6. An estimator δ is *location equivariant* if

$$\delta(X_1 + c, \dots, X_n + c) = \delta(X) + c.$$

The sample mean, sample median and sample quartiles are all examples of location equivariant estimators.

Theorem 3. [TPE 3.1.4] *If δ is a location equivariant estimator for a location invariant decision problem, then the risk, variance and bias of δ all are constant as functions of θ .*

Proof. We will prove that the risk is constant.

$$\begin{aligned} R(\theta, \delta) &= \mathbb{E}_{\theta}[L(\theta, \delta(X))] \\ &= \mathbb{E}_{\theta}[L(0, \delta(X) - \theta)] \\ &= \mathbb{E}_{\theta}[\rho(\delta(X) - \theta)] \\ &= \int \rho(\delta(x) - \theta) p(x; \theta) d\mu(x) \\ &= \int \rho(\delta(x_1 - \theta, \dots, x_n - \theta)) p(x; \theta) d\mu(x) \\ &= \int \rho(\delta(x_1 - \theta, \dots, x_n - \theta)) p(x_1 - \theta, \dots, x_n - \theta; 0) d\mu(x) \\ &= \int \rho(\delta(x)) p(x; 0) d\mu(x) \\ &= \mathbb{E}_0[\rho(\delta(X))] \\ &= R(0, \delta). \end{aligned}$$

which does not depend on θ . The bias and variance are similar. □

The upshot of this theorem is that we can always compare equivariant estimators. We have restricted our class of estimators in such a way so that the risk is just a number. It is no longer a function of θ .