Lecture 5 — October 6

Lecturer: Lester Mackey *Scribe: Ahmed Bou-Rabee, Claire Donnat*

Warning: *These notes may contain factual and/or typographic errors.*

5.1 Optimal Unbiased Estimation

In the last lecture, we introduced three techniques for finding optimal unbiased estimators when the loss function is convex:

A. Conditioning/Rao-Blackwellization.

B. Solving directly for the unique δ satisfying $\mathbb{E}[\delta(T(X))] = g(\theta)$.

C. Stumbling upon an unbiased function of our complete sufficient statistic.

We will go through some examples of these strategies, starting with one of Strategy C.

Example 1 (Strategy C: Stumble)**.** Let $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$. First, note that if σ^2 is known, $\bar{X}$ is a complete sufficient statistic for μ and hence also the UMVUE. Consider the case when $\theta = (\mu, \sigma^2)$ is unknown.

(a) The UMVUE for μ is $\bar{X}$.

(b) The UMVUE for σ^2 is $\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{n-1}$.

(c) What is the UMVUE for σ? First, note that $X_i - \bar{X} \sim \mathcal{N}(0, \frac{n-1}{n}\sigma^2)$, and hence $\mathbb{E}[|X_i - \bar{X}|] = \sigma\sqrt{\frac{2}{\pi}}\sqrt{\frac{n-1}{n}}$. This implies

$$\frac{\sqrt{\pi n}}{\sqrt{2(n-1)}}|X_i - \bar{X}|$$

is unbiased for σ. At this point we could Rao-Blackwellize, but the math is messy. Instead, we will try to stumble upon the solution. Let

$$S^2 = \sum_{i=1}^n (X_i - \bar{X})^2.$$

We know that

$$S^2 \sim \sigma^2 \chi^2_{n-1}.$$

Thus,

$$\mathbb{E}(S) = \sigma \mathbb{E}(\chi_{n-1}).$$

Which in turn implies that

$$\frac{\mathbb{E}(S)}{\mathbb{E}(\chi_{n-1})} = \sigma,$$

meaning $\frac{S}{\mathbb{E}(\chi_{n-1})}$ is unbiased for σ and hence UMVU.

(d) What is the UMVUE for μ^2? Taking the expectation of the UMVUE for μ and squaring it yields

$$\mathbb{E}(\bar{X}^2) = \mu^2 + \sigma^2/n.$$

So,

$$\delta_n(X) = \bar{X}^2 - \frac{S^2}{n(n-1)}$$

is the UMVUE. Note that $\delta_n(X)$ may be negative even though it estimates a non-negative quantity. Indeed, δ_n is inadmissible and dominated by the biased estimator $\max(0, \delta_n(X))$.

Example 2 (Strategy B: Solve)**.** Let $X \sim \text{Poi}(\theta)$. Since this is a one-dimensional full-rank exponential family, X is a complete sufficient statistic. X is furthermore unbiased and therefore UMVU for θ. Suppose that our goal, however, is to estimate $g(\theta) = e^{-a\theta}$ for $a \in \mathbb{R}$ known.

To employ Strategy B we must find an estimator δ such that $\mathbb{E}[\delta(X)] = g(\theta)$ for all θ. Under our model, we may reexpress this system of equations as

$$\sum_{x=0}^{\infty} \delta(x) \frac{e^{-\theta}\theta^x}{x!} = e^{-a\theta} \quad \text{for all} \quad \theta \tag{5.1}$$

$$\implies \sum_{x=0}^{\infty} \frac{\delta(x)\theta^x}{x!} = e^{(1-a)\theta} = \sum_{x=0}^{\infty} \frac{(1-a)^x\theta^x}{x!} \tag{5.2}$$

$$\implies \delta(X) = (1-a)^X \text{ is the UMVUE of } g(\theta). \tag{5.3}$$

However, this estimator is somewhat unsatisfying: if $a = 2$, for instance, it will change its sign according to X, even though our estimand $e^{-a\theta}$ is nonnegative. The estimator is in fact inadmissible when $a > 1$ and dominated by $\max(\delta(X), 0)$.

So we have seen that although we may be able to compute an UMVUE, this may not be a desirable decision rule. The two examples above shows that, even in simple cases, the UMVUE may be inadmissible. The problems do not end here however; in some cases, an UMVUE may not even exist.

Definition 1 (U-estimable)**.** We say $g(\theta)$ is **U-estimable** if an unbiased estimate for $g(\theta)$ exists.

Example 3 (Unbiased estimators of binomial distribution)**.** For $X \sim \text{Bin}(n, \theta)$ the only U-estimable functions of θ are polynomials of degree $\leq n$.

It is not uncommon for an UMVUE to be inadmissible, and it is often easy to construct a dominating (biased) estimator. Due to these and other limitations, the constraint of unbiasedness can be difficult to justify.

Example 4 (UMVUE for normal population variance)**.** We revisit the first example. Let $X_1..X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, where both μ, σ^2 are unknown. Let

$$S^2 = \sum_1^n (X_i - \overline{X})^2.$$

In this setting, $\frac{S^2}{n-1}$ is the UMVUE for σ^2. However, the biased Maximum Likelihood Estimator (MLE) $\frac{S^2}{n}$ has lower mean squared error. Furthermore, the shrunk estimator $\frac{S^2}{n+1}$ has lower mean squared error still. In this case, both the UMVUE and MLE are inadmissible!

We now present an explicit example of an UMVUE not existing.

Example 5 (Semiparametric unbiased estimators)**.** For $n > 2$, let $X_1, \ldots, X_n \overset{iid}{\sim} F$, where F is some unknown distribution on $\mathbb{R}$. Suppose that F is symmetric about some unknown point $\theta \in \mathbb{R}$. That is, suppose for all $X \sim F$, we have

$$X =_d 2\theta - X.$$

Consider the model

$$\mathcal{F} = \{\text{all distributions on } \mathbb{R} \text{ with finite variance symmetric about } \theta \in \mathbb{R}\}^1.$$

Then there is no UMVUE for the point of symmetry θ.

Proof. Suppose for sake of contradiction that the UMVUE $T(X)$ exists. Since $\overline{X}$ is unbiased for the full model $\mathcal{F}$, $T(X)$ must have variance no larger than $\overline{X}$. However, we know that $\overline{X}$ is the *unique* UMVUE for the Gaussian submodel, $\{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}$, and so $T(X)$ must equal $\overline{X}$ a.s. in the Gaussian submodel. This implies that $T(X) = \overline{X}$ a.s. under any continuous distribution on $\mathbb{R}$. In particular, $T(X) = \overline{X}$ a.s. under the uniform submodel $\{\text{Unif}(\theta - 1, \theta + 1) : \theta \in \mathbb{R}\}$. However, we learned in Homework 2, Problem 3 (b), that the distinct estimator $\frac{1}{2}(X_{(1)} + X_{(n)})$ has strictly better variance than $\overline{X}$ in the uniform submodel. Since $\frac{1}{2}(X_{(1)} + X_{(n)})$ is also unbiased for the full model $\mathcal{F}$, $T(X)$ cannot be the UMVUE after all. We are forced to conclude that no UMVUE exists over the whole family. □

Remark: In the previous proof, we argued that if the UMVUE existed, it must correspond to $\bar{X}$, which is the UMVUE of a Gaussian submodel. A question was raised in class concerning the choice of that UMVUE and that sub-model: can we make the same argument with any arbitrary submodel?

If we consider for instance the submodel with a single distribution $\mathcal{P} = \{N(\theta, 1)\}$ with $\theta = 2$, $\tilde{\eta}(X) = 2$ is an unbiased estimator for $\mathcal{P}$. However, this estimator does not put any constraints on the UMVUE for our model $\mathcal{F}$. Indeed, $\bar{X}$ is unbiased for every model in $\mathcal{F}$, while $\tilde{\eta}(X) = 2$ is only unbiased on a very specific submodel of $\mathcal{F}$, but not on the entire model $\mathcal{F}$. This distinction is important as any putative UMVUE $\delta(x)$ for $\mathcal{F}$ is required to have a variance at least as small as $\bar{X}$ in every model of F, while $\delta(x)$ is not required to have a variance at least as small as 2, as 2 is not unbiased for every model in $\mathcal{F}$.

[1]Such a model is called *semiparametric* because there is a finite dimensional unknown of interest, θ (the parametric part), as well as an infinite-dimensional unknown, F (the nonparametric part).

5.1.1 Sums of UMVUEs

We will end our exploration of unbiased estimation with an alternative characterization of UMVUEs inspired by the following question: Suppose we have δ_i UMVU for $g_i(\theta)$ for $i \in \{1, 2\}$. Is $\delta_1 + \delta_2$ then UMVU for $g_1(\theta) + g_2(\theta)$?

If our underlying family of distributions has a complete sufficient statistic, then Lehmann-Scheffe tells us this is definitely the case. However, we would like to say something about this even when no complete sufficient statistic exists, for which we will find the following theorem useful:

Theorem 1 (Characterization of UMVUEs; TPE 2.1.7)**.** Let $\Delta = \{\delta : \mathbb{E}_\theta[\delta^2] < \infty\}$. Then $\delta_0 \in \Delta$ is UMVU for $g(\theta) = \mathbb{E}[\delta_0]$ if and only if $\mathbb{E}[\delta_0(\theta)U] = 0$ for every $U \in \mathcal{U}$, where $\mathcal{U} = \{\text{unbiased estimators of } 0\}$.

Proof. If δ_0 is UMVUE let us consider $\delta_\lambda = \delta_0 + \lambda U$ for $\lambda \in \mathbb{R}, U \in \mathcal{U}$. Since δ_0 has minimal variance,

$$\text{Var}(\delta_\lambda) = \text{Var}(\delta_0) + \lambda^2 \text{Var}(U) + 2\lambda \text{Cov}(\delta_0, U) \geq \text{Var}(\delta_0). \tag{5.4}$$

Now consider the quadratic form $q(\lambda) = \lambda^2 \text{Var}(U) + 2\lambda \text{Cov}(\delta_0, U)$. The form q has the roots 0 and $-2\text{Cov}(\delta_0, u)/\text{Var}(u)$. If the roots are distinct, then the form must be negative at some point, which would violate the inequality (5.4). Hence, $-2\text{Cov}(\delta_0, U)/\text{Var}(U) = 0$, and thus $\mathbb{E}[U\delta_0] = \text{Cov}(\delta_0, U) = 0$.

For the converse result, we assume $\mathbb{E}[\delta_0 U] = 0,\ \forall U \in \mathcal{U}$, and consider any δ unbiased for $g(\theta)$. Then $\delta - \delta_0 \in \mathcal{U}$, so $\mathbb{E}[\delta_0(\delta - \delta_0)] = 0$. This implies that $\mathbb{E}[\delta_0 \delta] = \mathbb{E}[\delta_0^2]$, and subtracting $\mathbb{E}[\delta_0]\mathbb{E}[\delta]$ on both sides we get

$$\text{Var}(\delta_0) = \text{Cov}(\delta_0, \delta) \leq \sqrt{\text{Var}(\delta_0)\text{Var}(\delta)}$$

by Cauchy-Schwarz. Hence $\text{Var}(\delta_0) \leq \text{Var}(\delta)$ for any arbitrary unbiased estimator δ, and δ_0 is thus UMVU. □

Note that Theorem 1 provides a way to check for the existence of an UMVUE and to check whether a given estimator is UMVU, even when no complete sufficient statistic is known.

Turning back to our original question, we find that $\delta_1 + \delta_2$ is UMVU for $g_1(\theta) + g_2(\theta)$ simply by noting that

$$\forall U \in \mathcal{U},\ \ \mathbb{E}[(\delta_1 + \delta_2)U] = \mathbb{E}[\delta_1 U] + \mathbb{E}[\delta_2 U] = 0.$$

5.2 Optimal Equivariant Estimation

We now depart from the classical constraint of unbiasedness and turn our attention to the sorts of symmetries that naturally arise in decision problems. We begin by studying **location families** and their behavior under translation.

Let us consider a **location model** in which $X = (X_1, \ldots, X_n)$ follows a joint probability density of the form $f_\theta(X) \equiv f(x_1 - \theta, x_2 - \theta, \ldots, x_n - \theta)$, where f is fixed and known, and

$\theta \in \mathbb{R}$, the **location parameter**, is our unknown statistic. We denote this situation by $(X_1, \ldots, X_n) \sim \text{LocModel}(\theta)$. For example, this occurs when $X_i = U_i + \theta$ for $(U_1, \ldots, U_n) \sim f_0$.

Note that if X_i's are i.i.d., then the joint density factorizes as

$$f_\theta(X) \equiv \prod_{i=1}^{n} g_\theta(x_i) = \prod_{i=1}^{n} g(x_i - \theta).$$

Now, if $(X_1, \ldots, X_n) \sim \text{LocModel}(\theta)$ then let $X_i' = X_i + c$ for fixed c and all i, so that $(X_1', \ldots, X_n') \sim \text{LocModel}(\theta + c)$. Notice that this means that the shift in the data by a constant results in a shift in the model by a constant. Let's formalize these notions with some definitions to be able to do optimal inference in this setting.

Definition 2. (Location invariant model). A family of densities $\mathcal{P} = \{f_\theta | \theta \in \Omega\}$ is location invariant if $f_{\theta+c}(x + c) = f_\theta(x)$.

Definition 3. (Location invariant loss function). A loss function L is location invariant if $L(\theta, d) = L(\theta + c, d + c)\ \forall \theta, d, c$. This implies that the the loss function is of the form $L(\theta, d) = \rho(\theta - d)$, since $L(\theta, d) = L(\theta - d, 0)$.

Definition 4. (Location invariant decision problem). A decision problem is location invariant if the family of distribution $\mathcal{P}$ and the loss function L both are.

When our decision problem displays this sort of invariance to transformation, it is reasonable to constrain our estimator to respect these symmetries as well. This gives rise to the following definition of **equivariance**.

Definition 5. (Location equivariant estimator). An estimator δ is location equivariant if $\delta(X_1 + c, \ldots, X_n + c) = \delta(X_1, \ldots, X_n) + c$.

Many simple estimators (e.g., the arithmetic mean, the median, and any weighted average of order statistics) are location equivariant. And, indeed, it seems a reasonable constraint: if you are measuring heights and arbitrarily add a constant to all of them, you would expect that your estimator changes by that same constant.

Armed with these definitions we develop a general result, which will greatly simplify our search for an optimal equivariant estimator.

Theorem 2. (TPE 3.1.4) If δ is a location equivariant estimator for a location invariant decision problem $(\mathcal{P}, L)$ then the bias, risk and variance of δ have no θ dependence.

Proof. We only show the proof for the bias of the estimator. The calculations are analogous for the risk and the variance. The bias of δ is

$$\mathbb{E}_\theta[\delta(X) - \theta] = \mathbb{E}_\theta[\delta(X_1 - \theta, \ldots, X_n - \theta)] = \tag{5.5}$$
$$= \mathbb{E}_0[\delta(U_1, \ldots, U_n)] \text{ where } (U_1, \ldots, U_n) \sim f_0. \tag{5.6}$$

Thus the bias has no dependence on θ. □

The key idea here is that $\delta(X) - \theta$ has no θ dependence. This same lack of dependence occurs with risk and variance.

Our optimality goal in this constrained setting is to find a Minimum Risk Equivariant (MRE) estimator. If we wish to find an MRE estimator, we may leverage the fact that the risk is constant for all θ, which means that an MRE is necessarily the best estimator in its class for all θ. We will see soon that such an MRE typically exists. First, we will develop a characterization of all location equivariant estimators. For that, we need two lemmas (both proved in TPE).

Lemma 1. (TPE 3.1.6) If δ_0 is a location equivariant estimator then any other estimator δ is location equivariant if and only if:

$$\delta(X_1, \ldots, X_n) = \delta_0(X_1, \ldots, X_n) - U(X_1, \ldots, X_n) \tag{5.7}$$

where the statistic U is location invariant, i.e. $U(X_1 + c, \ldots, X_n, +c) = U(X_1, \ldots, X_n),\ c \in \mathbb{R}$.

Lemma 2. (TPE 3.1.7) A statistic U is location invariant if and only if U is a function $v(Y_1, \ldots, Y_n)$ of the differences $Y_i = X_i - X_n$, for $i = 1, \ldots, n-1$.

The difference above is taken between i and n simply for convenience. Employing these two lemmas we now that every location equivariant estimator has the form

$$\delta(X_1, \ldots, X_n) = \delta_0(X_1, \ldots, X_n) - v(Y_1, \ldots, Y_{n-1}) \tag{5.8}$$

where δ_0 is a reference location equivariant estimator, v is an arbitrary function, and $Y_i = X_i - X_n$.

Example 6. In a sample of a single observation ($n = 1$) all location equivariant estimators have the form $\delta(X_1) = X_1 + c$ for some $c \in \mathbb{R}$, since v in this case is a function of no arguments. Note that X_1 is itself location equivariant, so $\delta_0(X_1) = X_1$. These estimators are more interesting when $n > 1$.

The following theorem helps us find the best location equivariant estimator.

Theorem 3. (TPE 3.1.10) Given a location invariant decision problem, if δ_0 is location invariant (with finite risk) and for each Y, $v^*(Y)$ minimizes the expected conditional loss

$$\mathbb{E}_{\theta=0}[\rho(\delta_0(X) - v)|Y = y]$$

as a function of v, where $Y = (X_1 - X_n, \ldots, X_{n-1} - X_n)$, then an MRE estimator is $\delta_0(X) - v^*(Y)$.

We will prove this and present applications of this theorem next time.