Statistics 210A Lecture 6 Notes

Daniel Raban

September 14, 2021

1 Basu's Theorem, Rao-Blackwell, and Unbiased Estimation

1.1 Recap: Minimal sufficient, complete, and ancillary statistics

Last time we discussed **minimal sufficient** statistics, which are

1) $T(X)$ is sufficient.

2) For any other sufficient statistic $S(X)$, $T(X) = f(S(X))$ for some f (a.s. in $\mathcal{P}$).

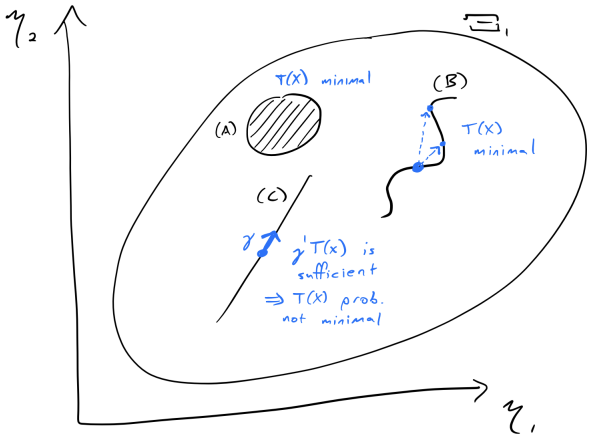
For an s-parameter exponential family with $p_\theta(x) = e^{\eta(\theta)^\top T(x) - B(\theta)} h(x)$, $T(X)$ is minimal if

$$\text{span}\{\eta(\theta_1) - \eta(\theta_2) : \theta_1, \theta_2 \in \Theta\} = \mathbb{R}^s.$$

We also discussed **complete statistics**, which have the property that

$$\mathbb{E}_\theta[f(T(x))] = \theta \quad \forall 0 \implies f(T(x)) \overset{\text{a.s.}}{=} 0.$$

We saw that in an exponential family, $T(X)$ is complete if Ξ contains an open set, but this is not a necessary condition. In the following picture, $T(X)$ will be complete in exponential families A and B but not necessarily in family C.



Family B is usually called a **curved exponential family** since the parameter space is a lower-dimensional space within the natural parameter space.

We saw that completeness is an upgrading of minimality for sufficient statistics:

Theorem 1.1. *Complete sufficient statistics are minimal.*

We also introduced **ancillary statistics** $V(X)$, where the distribution of V doesn't depend on θ.

1.2 Basu's theorem

Theorem 1.2 (Basu)**.** *If $T(X)$ is complete sufficient and $V(X)$ is ancillary for $\mathcal{P}$, then $V(X) \amalg T(X)$ for all $\theta \in \Theta$.*

Proof. We want to show that for all sets A, B and for all θ,

$$\mathbb{P}_\theta(V \in A, T \in B) = \mathbb{P}_\theta(V \in A)\mathbb{P}_\theta(T \in B).$$

This is equivalent to showing

$$\mathbb{P}_\theta(V \in A \mid T \in B) = \mathbb{P}_\theta(V \in A)$$

whenever $\mathbb{P}_\theta(T \in B) > 0$. Let

$$q_A(T(X)) = \mathbb{P}(V \in A \mid T(X)), \qquad p_A = \mathbb{P}(V \in A).$$

Note that q_A, p_A are independent of θ. We have

$$\mathbb{E}_\theta[q_A(T(X)) - p_A] = p_A - p_A = 0,$$

so by completeness of $T(X)$, $q_A(T(X)) \overset{\text{a.s.}}{=} p_A$. □

Remark 1.1. The hypotheses of Basu's theorem apply to a model, whereas the conclusions apply to each distribution. So sometimes, to prove that statistics are independent, we can apply Basu's theorem to submodels of the original model.

Example 1.1. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. We want to show that $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i \amalg S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$. Let $\mathcal{Q}_{\sigma^2} = \{N(\mu, \sigma^2)^n : \mu \in \mathbb{R}\}$. In this model, $\overline{X}$ is complete sufficient (which we can verify by writing this as an exponential family). To show that S^2 is ancillary, let $Z_i = X_i - \mu \overset{\text{iid}}{\sim} N(0, \sigma^2)$ (not that these are not statistics, since they suppose the value of μ). Then

$$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 = \underbrace{\frac{1}{n-1}\sum_{i=1}^n (Z_i - \overline{Z})^2}_{\sim \frac{\sigma^2}{n-1}\chi^2_{n-1}}$$

has distribution not depending on θ. So by Basu's theorem, $\overline{X} \amalg S^2$ for all μ, σ^2. Take note that we split the model into submodels where σ^2 was fixed.

1.3 The Rao-Blackwell Theorem

Why should we use sufficient statistics or complete statistics $T(X)$? The idea is that if θ only depends on $T(X)$, then using anything else would be adding extra randomness, or "noise," that obscures our result. To talk about what effect this has on our loss functions, let's introduce a condition on our loss functions.

Definition 1.1. A function $f(x)$ is **convex** if for any $\gamma \in (0,1)$,

$$f(\gamma x + (1-\gamma)y) \leq \gamma f(x) + (1-\gamma)f(y).$$

The function f is **strictly convex** if the inequality is strict ($<$).

Jensen's inequality says that this extends to general averages, not just the average of two points.

Theorem 1.3 (Jensen's inequality)**.** *If f is convex, then*

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)].$$

If f is strictly convex, we get $<$, unless $X \overset{\text{a.s.}}{=} c$.

Here, X could be a random vector.

If the loss $L(\theta; d)$ is convex in d, then we lose by adding extra noise. Jensen's inequality tells us that more the distribution spreads out, the more the average risk increases.

Theorem 1.4 (Rao-Blackwell)**.** *Assume $T(X)$ sufficient and $\delta(X)$ is an estimator for $g(\theta)$. Let $\overline{\delta}(T(X)) = \mathbb{E}[\delta(X) \mid T(X)]$. If $L(\theta; d)$ is convex, then*

$$R(\theta; \overline{\delta}) \leq R(\theta; \delta) \quad \forall \delta.$$

If $L(\theta; d)$ is strictly convex, then the inequality is strict, unless $\overline{\delta} \overset{\text{a.s.}}{=} \delta$.

Proof. The risk is

$$R(\theta; \overline{\delta}) = \mathbb{E}_\theta[L(\theta; \mathbb{E}(\delta \mid T))]$$

By Jensen's inequality (applied to the conditional expectation given T),

$$\begin{aligned} &\leq \mathbb{E}_\theta[\mathbb{E}[L(\theta; \delta) \mid T]] \\ &= \mathbb{E}_\theta[L(\theta; \delta)] \end{aligned}$$

with strict inequality for strict convexity unless $\overline{\delta} \overset{\text{a.s.}}{=} \delta$. □

Remark 1.2. Where did we use sufficiency in the proof? We used it when defining $\overline{\delta}$, where the conditional expectation should not depend on θ.

Turning δ into $\overline{\delta}$ is called **Rao-Blackwellization**.

1.4 Unbiased estimation

Definition 1.2. The **bias** of an estimator $\delta(X)$ for $g(\theta)$ is

$$\text{Bias}_\theta(\delta(X)) = \mathbb{E}_\theta\, \delta(X) - g(\theta).$$

The statistic $\delta(X)$ is **unbiased** for $g(\theta)$ if $\mathbb{E}_\theta[\delta(X)] = g(\theta)$ for all θ.

An unbiased estimator may not always exist.

Definition 1.3. We say $g(\theta)$ is U**-estimable** if there is an estimator $\delta(X)$ that is unbiased for $g(\theta)$.

Definition 1.4. An estimator $\delta(X)$ is **uniform minimum variance unbiased (UMVU)** if for any other unbiased $\widetilde{\delta}$, $\text{Var}_\theta(\delta(X)) \le \text{Var}_\theta(\widetilde{\delta}(X))$.

We could equivalently say $\text{MSE}(\theta;\delta) \le \text{MSE}(\theta;\widetilde{\delta})$.

Theorem 1.5 (Lehmann-Scheffé)**.** *Suppose $T(X)$ is complete sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Then for any U-estimable function $g(\theta)$, there is an a.s. unique UMVU estimator of the form $\delta(T(X))$.*

Proof. Assume $\delta_0(X)$ is unbiased for $g(\theta)$. Then define

$$\delta(T) = \mathbb{E}[\delta_0 \mid T].$$

This is unbiased because

$$\mathbb{E}_\theta[\delta(T)] = \mathbb{E}_\theta[\mathbb{E}[\delta_0 \mid T]] = \mathbb{E}[\delta_0] = g(\theta).$$

If $\widetilde{\delta}(T)$ is unbiased, then $\mathbb{E}[\delta(T) - \widetilde{\delta}(T)] = 0$ for all θ. So by completeness, $\delta(T) \overset{\text{a.s.}}{=} \widetilde{\delta}(T)$.

Now suppose $\delta^*(X)$ is unbiased. By Rao-Blackwell,

$$\text{MSE}(\theta; \underbrace{\mathbb{E}[\delta^* \mid T]}_{=\delta}) \le \text{MSE}(\theta;\delta^*). \qquad \square$$

Remark 1.3. The picture is the same for any convex loss, not just the mean squared error. For strictly convex loss, the unique UMVU has strictly less loss than any other unbiased estimator.

Remark 1.4. Unbiased estimators are not always the best, but this shows that there is at least a best one.

How do we find an unbiased estimator? Assume T is complete sufficient. We now have two options:

1. Find an unbiased estimator $\delta(T)$ which is a function of T.

2. Find any unbiased estimator $\delta_0(X)$ and Rao-Blackwellize it.

Example 1.2 (German tank problem[1])**.** Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} U[0, \theta]$ with $\theta > 0$.

$$\begin{aligned} p(x) &= \prod_{i=1}^{n} p^{(1)}(x_i) \\ &= \prod_{i=1}^{n} \mathbb{1}_{\{0 \le x_i \le \theta\}} \frac{1}{\theta} \\ &= \frac{1}{\theta^n} \mathbb{1}_{\{0 \le X_{(n)} \le \theta\}}. \end{aligned}$$

Is the maximum complete sufficient?

$$\begin{aligned} \mathbb{P}_\theta(X_{(n)} \le t) &= \left(\frac{t}{\theta} \wedge 1 \right)^n \\ &= \left(\frac{t}{\theta} \right)^n \wedge 1, \end{aligned}$$

so the density is

$$\begin{aligned} p_\theta(t) &= \frac{d}{dt} \mathbb{P}_\theta(X_{(n)} \le t) \\ &= n \frac{t^{n-1}}{\theta^n} \mathbb{1}_{\{t \le \theta\}}. \end{aligned}$$

Suppose that for all $\theta > 0$,

$$0 = \mathbb{E}_\theta[f(T)] = \frac{n}{\theta^n} \int_0^\theta f(t) t^{n-1} \, dt.$$

Then

$$0 = \int_0^\theta f(t) t^{n-1} \, dt,$$

so differentiating with respect to θ tells us that

$$f(\theta) \theta^{n-1} = 0$$

for all $\theta > 0$.

Let's calculate

$$\mathbb{E}_\theta[X_{(n)}] = \frac{n}{\theta^n} \int_0^\theta t \cdot t^{n-1} \, dt$$

[1] Imagine you're hiding in the bushes in World War II, and you count the serial numbers. You observe the largest serial number to try to determine the number of German tanks.

$$= \frac{n}{\theta^n(n+1)}[t^{n+1}]_0^\theta$$
$$= \frac{n}{n+1}\theta.$$

So we can just get an unbiased estimator via

$$\mathbb{E}_\theta\left[\frac{n+1}{n}X_{(n)}\right] = \theta.$$

Another way to get an unbiased estimator is to use $\mathbb{E}_\theta[2X_1] = \theta$. Then you can show that

$$\mathbb{E}[2X_i \mid X_{(n)}] = \frac{n+1}{n}X_{(n)}.$$