

STATS300A - Lecture 4

Dominik Rothenhaeusler
Scribed by Michael Howes

09/29/21

Contents

1	Announcements	1
2	Recap	1
3	Conditional expectation	2
4	Estimation	2
4.1	Complete statistics	2
4.2	Risk reduction via conditioning	3
4.3	Optimal unbiased estimators	4
4.4	Consequences	5

1 Announcements

- HW1 due today.
- HW2 will be posted today.

2 Recap

We have seen

- Exponential families.
- Data reduction:
 - Sufficiency (all necessary information),
 - Minimal sufficiency (coarest sufficient statistic),
 - Ancillary (useless data),
 - Completeness (contains no useless data).

Today we will look at optimal unbiased estimation but first a refresher on conditional expectations.

3 Conditional expectation

(See Keener Chp 1.1) Let X and Z be random variables with density $p(x, z) = p(z|x)p(x)$. Let h be a function with finite expectation, that is

$$\int \int |h(x, z)|p(z|x)p(x)dzdx < \infty.$$

Then we define the conditional expectation of h given Z as

$$\mathbb{E}[h(X, Z)|X = x] = \int h(x, z)p(z|x)dx.$$

Note that $\mathbb{E}[h(X, Z)|X]$ is a function of x . Some properties of conditional expectation are

- (Pull out property) $\mathbb{E}[h_1(X)h_2(X, Y)|X = x] = h_1(x)\mathbb{E}[h_2(X, Y)|X = x]$.
- (Tower property) $\mathbb{E}[\mathbb{E}[h(X, Z)|X]] = \mathbb{E}[h(X, Z)]$.
- (Independence) If X and Z are independent ($p(z|x) = p(z)$ for all x, z), then

$$\mathbb{E}[h(Z)|X = x] = \mathbb{E}[h(Z)].$$

We will use these ideas when studying estimation.

4 Estimation

4.1 Complete statistics

See TSH Theorem 4.3.1. In a full rank exponential family, the statistic (T_1, \dots, T_s) is complete.

Theorem 1. [Basu's Theorem] *If T is complete and sufficient for $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$ and A is ancillary for \mathcal{P} , then $T(X)$ is independent of $A(X)$ which we write as $A(X) \perp\!\!\!\perp T(X)$.*

By independent we mean for all events C , $P_\theta(A(X) \in C | T(X) = t) = P_\theta(A(X) \in C)$. We will not prove this here but it can be done using the tower property. Here is an application of this theorem.

Example 1. Let $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. We wish to show that $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ and $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ are independent.

Proof. Fix σ^2 and consider the model where only μ is unknown. This is an exponential family and thus the statistic \bar{X} is complete and sufficient. Also since we are working with a location model the statistic

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

has the same distribution as

$$\frac{1}{n-1} \sum_{i=1}^n (Z_i - \bar{Z}),$$

where $Z_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$. Thus, by Basu's theorem $S^2 \perp\!\!\!\perp \bar{X}$. Although we fixed σ^2 in the submodel, σ^2 was arbitrary and thus $S^2 \perp\!\!\!\perp \bar{X}$ regardless of μ and σ^2 . \square

4.2 Risk reduction via conditioning

Definition 1. Let C be a convex subspace of a vector space. A function $f : C \rightarrow \mathbb{R}$ is *convex* if

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

for all $x, y \in C$ and all $\gamma \in [0, 1]$. If we have strict inequality for all $x, y \in C$ and $\gamma \in (0, 1)$, then we say that f is *strictly convex*.

Theorem 2. [Jensen's Inequality] Let $f : C \rightarrow \mathbb{R}$ be convex on an open subset C with $P(X \in C) = 1$. If $\mathbb{E}[X]$ exists, then

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

If f is strictly convex then

$$f(\mathbb{E}[X]) < \mathbb{E}[f(X)],$$

unless $P(X = \mathbb{E}[X]) = 1$.

Recall that $L(\theta, d)$ is the penalty when θ is the true parameter and decision d is made. Also recall that $R(\theta, \delta) = \mathbb{E}_\theta[L(\theta, \delta(X))]$ is the risk of the decision procedure δ .

Theorem 3. [Rao-Blackwell] Suppose that T is sufficient for $\{P_\theta : \theta \in \Omega\}$ and that $\delta(X)$ is an estimator with $\mathbb{E}_\theta|\delta(X)| < \infty$ and $R(\theta, \delta) < \infty$. Let $\eta(T) = \mathbb{E}[\delta(X)|T]$ (which is well-defined because T is sufficient). Then

(a) If $L(\theta, \cdot)$ is convex for a fixed θ , then

$$R(\theta, \eta) \leq R(\theta, \delta).$$

(b) If $L(\theta, \cdot)$ is strictly convex for a fixed θ , then

$$R(\theta, \eta) < R(\theta, \delta),$$

unless $\eta(T(X)) = \delta(X)$ with probability 1.

Proof. For (a),

$$\begin{aligned} R(\theta, \eta) &= \mathbb{E}_\theta[L(\theta, \eta(X))] \\ &= \mathbb{E}_\theta[L(\theta, \mathbb{E}[\delta(X)|T])] \\ &\leq \mathbb{E}_\theta[\mathbb{E}[L(\theta, \delta(X))|T]] \quad (\text{Jensen's inequality}) \\ &= \mathbb{E}_\theta[L(\theta, \delta(X))] \quad (\text{Tower property}) \\ &= R(\theta, \delta). \end{aligned}$$

For (b) we note that this inequality is strict if $P(\eta(T) \neq \delta(X)) > 0$. □

The take away

- Under convex loss, a deterministic estimator based on sufficient statistics is as good or better than any other estimator.
- If our loss is strictly convex, then additional randomness deteriorates performance.

Example 2. $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Bern}(\theta)$, $L(\theta, d) = (\theta - d)^2$ and $\delta(X) = X_1$. We know that $T(X) = \sum_{i=1}^n X_i$ is sufficient for this model. Thus Rao-Blackwell states

$$\eta(T) = \mathbb{E}[X_1|T(X)],$$

is at least as good as δ . Note that $\eta(T) = \mathbb{E}[X_i|T]$ by the iid assumption. Thus

$$\begin{aligned}\eta(T) &= \frac{1}{n}(n\eta(T)) \\ &= \frac{1}{n} \left(\sum_{i=1}^n \mathbb{E}[X_i|T] \right) \\ &= \frac{1}{n} \mathbb{E}[T|T] \\ &= \frac{1}{n} T \\ &= \frac{1}{n} \sum_{i=1}^n X_i \\ &= \bar{X}.\end{aligned}$$

Thus Rao-Blackwell recovers the sample average. The risk of δ was $\theta(1-\theta)$ and the risk of η is $\frac{\theta(1-\theta)}{n}$.

Consider also the example $\delta_{\text{goofy}}(X) = 0.5$. Then $\eta(T) = \mathbb{E}[0.5|T] = 0.5 = \delta_{\text{goofy}}(X)$ so Rao-Blackwell does not improve our estimator (it also doesn't make it worse). Conditioning reduces variance but it does not reduce bias.

4.3 Optimal unbiased estimators

Goal: Find the uniformly minimum risk unbiased estimator (UMRUE). That is we want for a fixed function g

- (a) $\mathbb{E}_\theta \delta(X) = g(\theta)$, for all $\theta \in \Omega$.
- (b) $R(\theta, \delta) \leq R(\theta, \delta')$ for all $\theta \in \Omega$ and all decision procedures δ satisfying (a).

A special case is when $L(\theta, d) = (g(\theta) - d)^2$, then the UMRUE is also called a UMVUE where the V stands for variance. This is because in this case $R(\theta, \delta) = \text{bias}^2 + \text{variance} = \text{variance}$.

Theorem 4. [Lehmann-Scheffe] *If T is complete and sufficient and $\mathbb{E}_\theta h(T) = g(\theta)$ for all θ , then*

- (a) $h(T)$ is the only function of T that is unbiased for $g(\theta)$.
- (b) $h(T)$ is the UMRUE if $L(\theta, \cdot)$ is convex for all $\theta \in \Omega$.
- (c) $h(T)$ is the unique UMRUE if $L(\theta, \cdot)$ is convex for all $\theta \in \Omega$ and $L(\theta_0, \cdot)$ is strictly convex for some $\theta_0 \in \Omega$.
- (d) $h(T)$ is the unique UMVUE.

Proof. Suppose $\mathbb{E}_\theta \tilde{h}(T) = g(\theta)$ for all $\theta \in \Omega$, then $E_\theta[(h - \tilde{h})(T)] = 0$ for all $\theta \in \Omega$. Thus $h - \tilde{h}$ is first order ancillary for T and since T is complete, this implies $h - \tilde{h} = 0$. Thus $\tilde{h} = h$ and we have proved part (a).

Suppose δ is an unbiased estimator for $g(\theta)$. Define $\eta(T) = \mathbb{E}[\delta(X)|T]$, then $\eta(T)$ is unbiased by the tower property and well defined by sufficiency. Since $\eta(T)$ is a function of T we have $\eta(T) = h(T)$ by part (a). Finally by Rao-Blackwell we have for all $\theta \in \Omega$.

$$R(\theta, h(t)) = R(\theta, \eta(T)) \leq R(\theta, \delta(X)), \quad (1)$$

thus h is UMRUE.

Suppose $L(\theta_0, \cdot)$ is strictly convex, then the inequality (1) is strictly convex unless $h(T) = \delta(X)$. This shows that $h(T)$ is the unique UMRUE.

Finally the case of mean square error is a special case of a strictly convex loss. Thus part (c) implies part (d). \square

4.4 Consequences

As a consequence of this theorem we can do optimal estimation for full rank exponential families as we know their complete sufficient statistics. This theorem also gives us strategies for finding UMRUEs.

- (a)
 - i. Find a complete sufficient statistic T .
 - ii. Find an unbiased estimator δ .
 - iii. Compute $\mathbb{E}[\delta(X)|T]$.

Step (iii) may be hard.

- (b)
 - i. Find a complete sufficient statistic T .
 - ii. Recall that there is at most one unbiased estimator that is a function of T and that this function is the UMRUE.
 - iii. Solve for δ in the equation $\mathbb{E}_\theta[\delta(T)] = g(\theta)$
- (c) Guess the UMRUE.