STATS300A - Lecture 7

Dominik Rothenhaeusler
Scribed by Michael Howes

10/11/21

Contents

1 Overview of optimal estimation

- We have seen that uniformly best estimators do not exist.
- We can constrain our class of estimators eg unbiased, equivariant.
- We can also collapse the risk via Bayesian estimation or minimax estimation.

Today we will finish up on equivariance and discuss Bayesian estimation.

2 Equivariance

Recall that if δ_0 is any equivariant estimator, and $v^*(y)$ is defined to be the value v that minimises

$$\mathbb{E}_0[\rho(\delta_0(X) - v)|Y = y],$$

where $Y = (X_1 - X_n, \dots, X_{n-1} - X_n)$, then $\delta^*(X) = \delta_0(X) - v^*(Y)$ is MRE.

2.1 MREs vs UMRUES

- MREs depend on the loss function.
- UMRUES do not depend on the loss function provided the loss function is strictly convex. This is because the UMRUE is often the unique unbiased estimator that is a function of a complete sufficient statistic.
- UMRUES do not always exist.

- MREs usually do exist and we can find them via an optimisation procedure.
- UMRUES are often *inadmissible.*
- Pitman's estimator is admissible under weak regularity conditions (Stein 1959).
- MREs are often biased if something other squared error loss is being used.

2.2 Risk unbiasedness

Definition 1. An estimator δ is risk unbiased for the loss $L(\theta, d)$ if all θ, θ'

$$\mathbb{E}_\theta[L(\theta, \delta(X))] \leq \mathbb{E}_\theta[L(\theta', \delta(X))].$$

Intuitively this says the true parameter θ penalises less than the false parameter θ'. If L is squared error loss, then this is the same as regular unbiasedness.

Theorem 1. [TPE 3.127] *If δ if MRE for a location invariant decision problem, then δ is risk unbiased.*

Sketch. Prove the contrapositive. Show that if δ is not risk unbiased, then a shifted version of δ has strictly lower risk. □

3 Location-Scale Models

Suppose that $X = (X_1, \ldots, X_n)$ has the joint density

$$f_{\theta,\tau}(x_1, \ldots, x_n) = \frac{1}{\tau^n} f\left(\frac{x_1 - \theta}{\tau}, \ldots, \frac{x_n - \theta}{\tau}\right).$$

Our parameters are the *location* $\theta \in \mathbb{R}$ and *scale* $\tau > 0$.

Example 1. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \tau^2)$, then

$$\begin{aligned} f(x_1, \ldots, x_n) &= \frac{1}{(\sqrt{2\pi}\tau)^n} \exp\left\{-\frac{1}{2\tau^2}\sum_{i=1}^n (x_i - \theta)^2\right\} \\ &= \frac{1}{\tau^n}\frac{1}{\sqrt{2\pi}^n} \exp\left\{-\frac{1}{2}\sum_{i=1}^n \left(\frac{x_i - \theta}{\tau}\right)^2\right\} \\ &= \frac{1}{\tau^n} g\left(\frac{x_1 - \theta}{\tau}, \ldots, \frac{x_n - \theta}{\tau}\right), \end{aligned}$$

where $g \sim \mathcal{N}(0, I_n)$.

Note that if X_i has a location-scale distribution (θ, τ), then $X_i' = aX_i + b$ has a loc-scale distribution $(a\theta + b, a\tau)$ if $a > 0$ and $b \in \mathbb{R}$. Our goal is to estimate θ, we are not interested in estimating τ and we call τ a *nuissance parameter.*

Definition 2. A loss function is *loc-scale invariant* if

$$L((a\theta + b, a\tau), ad + b) = L((\theta, \tau), d),$$

for all $a > 0$ and $b \in \mathbb{R}$. This is equivalent to requiring that $L((\theta, \tau), d)$ is a function of $\frac{\theta - d}{\tau}$.

Definition 3. A model $\mathcal{P} = \{P_{(\theta,\tau)} : (\theta, \tau) \in \Omega\}$ is *loc-scale invariant* if

$$f_{(a\theta+b,a\tau)}(ax + b) = f_{(\theta,\tau)}(x),$$

for all $a > 0$ and $b \in \mathbb{R}$.

Definition 4. An estimator δ, is *loc-scale equivariant* if

$$\delta(aX+b) = a\delta(X)+b,$$

for all $a>0$ and $b \in \mathbb{R}$.

Theorem 2. *Suppose we have a loc-scale invariant loss and model. Let δ^*_τ be the MRE in the submodel where τ is fixed.*

If δ^*_τ does not depend on the scale τ, then $\delta^* = \delta^*_\tau$ is the MRE for the full mode. That is for any loc-scale equivariant estimator δ',

$$R((\theta,\tau),\delta^*) \leq R((\theta,\tau),\delta').$$

This is another example of the technique of restricting attention to a submodel and then concluding something about the full model (recall our semi-parametric example from earlier).

Proof. Assume δ' is strictly better at (θ_0,τ_0) and that δ' is loc-scale equivariant. Then δ' has strictly better risk on the submodel $\{(\theta,\tau) : \tau = \tau_0\}$ and δ' is loc equivaraint on the submodel. This is contradiction to δ^* being the MRE on the submodel. □

Example 2. In the model $\mathcal{N}(\theta,\tau^2)$, $\bar{X}$ is the MRE under squared error loss for fixed τ. It does not depend on τ and thus $\bar{X}$ is the MRE in the full model.

Example 3. If $f_{\theta,\tau} \sim \frac{1}{\tau}\exp\left\{-\frac{1}{\tau}(x-\theta)\right\}\mathbb{I}(x \geq \theta)$, then the loc MRE under squared error loss if $X_{(1)} - \frac{\tau}{n}$ depends on τ. We will study this example more on the next assignment.

4 Bayes Estimators

As before we have our data $X \in \mathcal{X}$ and model $\mathcal{P} = \{P_\theta : \theta \in \Omega\}$. Let Λ be a measure on Ω. We wish to find an estimator δ that minimizes the *average risk*

$$r(\Lambda,\delta) = \int_\Omega R(\theta,\delta)d\Lambda(\theta).$$

If Λ is a probability distribution, then Λ is called a prior distribution. Is δ_Λ minimises $r(\Lambda,\delta)$, then δ_Λ is called a *Bayes estimator* and we call $r(\Lambda,\delta_\Lambda)$ the *Bayes risk*.

Proposition 1. *If Λ is a probability distribution, then we have*

$$r(\Lambda,\delta) = \mathbb{E}L(\Theta,\delta(X)),$$

where $\Theta \sim \Lambda$ and $X|\Theta=\theta \sim P_\theta$.

Note that the above expectation is with respect to both Θ and X. This is different to the regular risk which is only an expectation with respect to X.

Proof. Note that

$$\begin{aligned}\mathbb{E}(L(\Theta,\delta(X))) &= \mathbb{E}[\mathbb{E}[L(\Theta,\delta(X))|\Theta]]\\ &= \mathbb{E}[R(\Theta,\delta)]\\ &= \int_\Omega R(\theta,\delta)d\Lambda(\theta)\\ &= r(\Lambda,\delta).\end{aligned}$$

□

The usual interpretation is that Λ encodes prior beliefs about θ that we have before seeing the data X. Our main result is the following:

Theorem 3. *Suppose $X \sim \Lambda$ and $X|\Theta = \theta \sim P_\theta$. If there exists an estimator δ_0 with finite risk and if for almost every x, there exists a value $\delta(x)$ that minimises*

$$\mathbb{E}[L(\Theta, d)|X = x],$$

over d, then $\delta(X)$ is the Bayes estimator.

Proof. For a.e. x and every estimator δ' we have

$$\mathbb{E}[L(\Theta, \delta'(X))|X = x] \geq \mathbb{E}[L(\Theta, \delta(X))|X = x].$$

Thus,

$$\begin{aligned} r(\Lambda, \delta') &= \mathbb{E}\left[\mathbb{E}\left[L(\theta, \delta'(X))|X = x\right]\right] \\ &\geq \mathbb{E}\left[\mathbb{E}\left[L(\theta, \delta(X))|X = x\right]\right] \\ &= r(\Lambda, \delta). \end{aligned}$$

Thus δ is the Bayes estimator. □

Note that under squared error loss the minimizer of $\mathbb{E}[(g(\Theta) - \delta(x))^2|X = x]$ is the conditional expectation $\mathbb{E}[g(\Theta)|X = x]$ which we call the posterior mean.

4.1 A binomial example

Suppose that $X \sim \text{Bin}(n, \theta)$ where $\theta \sim \Lambda = \text{Beta}(a, b)$. That is the prior for θ has density

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}.$$

The likelihood of θ is

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

Recall Bayes rule

$$\begin{aligned} \text{posterior} &= \frac{\text{joint density}}{\text{marginal of } x} \\ &= \frac{\text{prior} \times \text{likelihood}}{\text{marginal of } x} \\ \therefore p(\theta|x) &= \frac{p(x,\theta)}{p(x)} \\ &= \frac{\pi(\theta)f(x|\theta)}{p(x)} \\ &= \frac{\pi(\theta)f(x|\theta)}{\int \pi(\theta')f(x|\theta')d\theta'} \end{aligned}$$

Thus

$$\begin{aligned} \pi(\theta|x) &\propto \text{likelihood} \times \text{prior} \\ &\propto \theta^{x+a-1}(1-\theta)^{n-x+b} \\ &\propto \text{Beta}(x+a, n-x+b). \end{aligned}$$

Thus the Bayes estimator under squared error loss is the mean of Beta$(x+a, n-x+b)$ which is $\frac{x+a}{n+a+b}$ (exercise). Note that

$$\frac{x+a}{n+a+b} = \frac{n}{n+a+b}\frac{x}{n} + \frac{a+b}{n+a+b}\frac{a}{a+b} = \lambda_n \cdot \text{UMVUE} + (1-\lambda_n) \cdot \text{prior mean.}$$

Thus the Bayes estimator is a convex combination of the UMVUE and the prior. Also $\lambda_n \to 1$ and $n \to \infty$ and so with enough data we approach the UMVUE. If n is small compared to $a+b$, then the Bayes estimator is closer to the mean of the Bayes's prior.