Statistical Inference
Lecture Six

https://people.bath.ac.uk/masss/APTS/apts.html

Simon Shaw

University of Bath

APTS, 14-18 December 2020

Overview of Lecture Six

The key idea from Lecture Five is:

- Wald's Complete Class Theorem, CCT. A decision rule is admissible if and only if it is a Bayes rule for some prior distribution.

In this lecture we will consider use of loss functions for point estimation, set estimation and hypothesis testing.

- For quadratic loss, a point estimator for θ is admissible if and only if it is the conditional expectation with respect to some positive prior distribution $\pi(\theta)$.
- Level set property (LSP): a set $d \subset \Theta$ is a level set of the posterior distribution exactly when $d = \{\theta : \pi(\theta \,|\, x) \geq k\}$ for some k.
- If δ^* is a Bayes rule for $L(\theta, d) = |d| + \kappa(1 - \mathbb{1}_{\theta \in d})$ then it is a level set of the posterior distribution.

Point estimation

- We now look at possible choices of loss functions for different types of inference.
- For point estimation the decision space is $\mathcal{D} = \Theta$, and the loss function $L(\theta, d)$ represents the (negative) consequence of choosing d as a point estimate of θ.
- It will not be often that an obvious loss function $L : \Theta \times \Theta \to \mathbb{R}$ presents itself. There is a need for a generic loss function which is acceptable over a wide range of applications.

Suppose that Θ is a convex subset of $\mathbb{R}^p$. A natural choice is a convex loss function,

$$L(\theta, d) = h(d - \theta)$$

where $h : \mathbb{R}^p \to \mathbb{R}$ is a smooth non-negative convex function with $h(0) = 0$.

- This type of loss function asserts that small errors are much more tolerable than large ones.
- One possible further restriction is that h is an even function, $h(d-\theta) = h(\theta - d)$.
- In this case, $L(\theta, \theta+\epsilon) = L(\theta, \theta-\epsilon)$ so that under-estimation incurs the same loss as over-estimation.
- We saw previously, that for quadratic loss $\Theta \subset \mathbb{R}$, $L(\theta, d) = (\theta - d)^2$, the Bayes rule was the expectation of $\pi(\theta)$. As we will see, this attractive feature can be extended to more dimensions.
- There are many situations where this is not appropriate and the loss function should be asymmetric and a generic loss function should be replaced by a more specific one.

The bilinear loss function for $\Theta \subset \mathbb{R}$ is, for $\alpha, \beta > 0$,

$$L(\theta, d) \;=\; \begin{cases} \alpha(\theta - d) & \text{if } d \leq \theta, \\ \beta(d - \theta) & \text{if } d \geq \theta. \end{cases}$$

- The Bayes rule is a $\frac{\alpha}{\alpha+\beta}$-fractile of $\pi(\theta)$.
- If $\alpha = \beta = 1$ then $L(\theta, d) = |\theta - d|$, the absolute loss which gives a Bayes rule of the median of $\pi(\theta)$.
- $|\theta - d|$ is smaller that $(\theta - d)^2$ for $|\theta - d| > 1$ and so absolute loss is smaller than quadratic loss for large deviations. Thus, it takes less account of the tails of $\pi(\theta)$ leading to the choice of the median.
- If $\alpha > \beta$, so $\frac{\alpha}{\alpha+\beta} > 0.5$, then under-estimation is penalised more than over-estimation and so that Bayes rule is more likely to be an over-estimate.

Example

If $\Theta \in \mathbb{R}^p$, the Bayes rule δ^* associated with the distribution $\pi(\theta)$ and the quadratic loss

$$L(\theta, d) \;=\; (d-\theta)^T Q\,(d-\theta)$$

is the expectation $\mathbb{E}_{(\pi)}(\theta)$ for every positive-definite symmetric $p \times p$ matrix Q.

Example (Robert, 2007), $Q = \Sigma^{-1}$

Suppose $X \sim N_p(\theta, \Sigma)$ where the known variance matrix Σ is diagonal with elements σ_i^2 for each i. Then $\mathcal{D} = \mathbb{R}^p$. A possible loss function is

$$L(\theta, d) \;=\; \sum_{i=1}^{p} \left(\frac{d_i - \theta_i}{\sigma_i}\right)^2$$

so that the total loss is the sum of the squared component-wise errors.

- As the Bayes rule for $L(\theta, d) = (d - \theta)^T Q\,(d - \theta)$ does not depend upon Q, it is the same for an uncountably large class of loss functions.
- If we apply the Complete Class Theorem to this result we see that for quadratic loss, a point estimator for θ is admissible if and only if it is the conditional expectation with respect to some positive prior distribution $\pi(\theta)$.
- The value, and interpretability, of the quadratic loss can be further observed by noting that, from a Taylor series expansion, an even, differentiable and strictly convex loss function can be approximated by a quadratic loss function.

Set estimation

- For set estimation the decision space is a set of subsets of Θ so that each $d \subset \Theta$.
- There are two contradictory requirements for set estimators of Θ.
  1. We want the sets to be small.
  2. We also want them to contain θ.
- A simple way to represent these two requirements is to consider the loss function

$$L(\theta, d) \;=\; |d| + \kappa(1 - \mathbb{1}_{\theta \in d})$$

  for some $\kappa > 0$ where $|d|$ is the volume of d.
- The value of κ controls the trade-off between the two requirements.
  - If $\kappa \downarrow 0$ then minimising the expected loss will always produce the empty set.
  - If $\kappa \uparrow \infty$ then minimising the expected loss will always produce Θ.

- For loss functions of the form $L(\theta, d) = |d| + \kappa(1 - \mathbb{1}_{\theta \in d})$ we'll show there is a a simple necessary condition for a rule to be a Bayes rule.

Definition (Level set)

A set $d \subset \Theta$ is a level set of the posterior distribution exactly when $d = \{\theta : \pi(\theta \,|\, x) \geq k\}$ for some k.

Theorem (Level set property, LSP)

If δ^* is a Bayes rule for $L(\theta, d) = |d| + \kappa(1 - \mathbb{1}_{\theta \in d})$ then it is a level set of the posterior distribution.

Proof

Note that

$$\begin{aligned} \mathbb{E}\{L(\theta, d) \,|\, X\} &= |d| + \kappa(1 - \mathbb{E}(\mathbb{1}_{\theta \in d} \,|\, X)) \\ &= |d| + \kappa \mathbb{P}(\theta \notin d \,|\, X). \end{aligned}$$

Proof continued

- For fixed x, we show that if d is not a level set of the posterior distribution then there is a $d' \neq d$ which has a smaller expected loss so that $\delta^*(x) \neq d$.
- Suppose that d is not a level set of $\pi(\theta \,|\, x)$. Then there is a $\theta \in d$ and $\theta' \notin d$ for which $\pi(\theta' \,|\, x) > \pi(\theta \,|\, x)$.
- Let $d' = d \cup d\theta' \setminus d\theta$ where $d\theta$ is the tiny region of Θ around θ and $d\theta'$ is the tiny region of Θ around θ' for which $|d\theta| = |d\theta'|$.
- Then $|d'| = |d|$ but

$$\mathbb{P}(\theta \notin d' \,|\, X) < \mathbb{P}(\theta \notin d \,|\, X)$$

Thus, $\mathbb{E}\{L(\theta, d') \,|\, X\} < \mathbb{E}\{L(\theta, d) \,|\, X\}$ showing that $\delta^*(x) \neq d$. □

- The Level Set Property Theorem states that δ having the level set property is necessary for δ to be a Bayes rule for loss functions of the form $L(\theta, d) = |d| + \kappa(1 - \mathbb{1}_{\theta \in d})$.
- The Complete Class Theorem states that being a Bayes rule is a necessary condition for δ to be admissible.
- Being a level set of a posterior distribution for some prior distribution $\pi(\theta)$ is a necessary condition for being admissible for loss functions of this form.
- Bayesian HPD regions satisfy the necessary condition for being a set estimator.
- Classical set estimators achieve a similar outcome if they are level sets of the likelihood function, because the posterior is proportional to the likelihood under a uniform prior distribution.[1]

[1] In the case where Θ is unbounded, this prior distribution may have to be truncated to be proper.

Hypothesis tests

- For hypothesis tests, the decision space is a partition of Θ, denoted

$$\mathcal{H} \ := \ \{H_0, H_1, \ldots, H_d\}.$$

- Each element of $\mathcal{H}$ is termed a hypothesis.
- The loss function $L(\theta, H_i)$ represents the (negative) consequences of choosing element H_i, when the true value of the parameter is θ.
- It would be usual for the loss function to satisfy

$$\theta \in H_i \implies L(\theta, H_i) \ = \ \min_j L(\theta, H_j)$$

on the grounds that an incorrect choice of element should never incur a smaller loss than the correct choice.

- Consider the test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ where $\Theta_1 = \Theta \setminus \Theta_0$. Let $\mathcal{D} = \{d_0, d_1\}$ where d_i corresponds to accepting H_i. A generic loss function is the 0-1 ('zero-one') loss function

$$L(\theta, d_i) = \begin{cases} 0 & \text{if } \theta \in \Theta_i, \\ 1 & \text{if } \theta \notin \Theta_i. \end{cases}$$

- The classical risk is the probability of making a wrong decision,

$$R(\theta, \delta) = \begin{cases} \mathbb{P}(\delta(X) = d_1 \,|\, \theta) & \text{if } \theta \in \Theta_0, \\ \mathbb{P}(\delta(X) = d_0 \,|\, \theta) & \text{if } \theta \in \Theta_1, \end{cases}$$

  which correspond to the familiar Type I and Type II errors.
- The Bayes rule is to choose H_0 if $\mathbb{P}_\pi(\theta \in \Theta_0) > \mathbb{P}_\pi(\theta \in \Theta_1)$ and H_1 otherwise, where $\mathbb{P}_\pi(\cdot)$ is the probability when $\theta \sim \pi(\theta)$.
- Hence, if $\pi(\theta) = f(\theta \,|\, x)$, the Bayes rule is to choose the hypothesis with the largest posterior probability.

- This approach can be naturally extended to multiple hypotheses $\mathcal{H} = \{H_0, H_1, \ldots, H_d\}$ which partition Θ by taking

$$L(\theta, H_i) \;=\; 1 - \mathbb{1}_{\{\theta \in H_i\}}.$$

  i.e., zero if $\theta \in H_i$, and one if it is not.
- For the posterior decision, the Bayes rule is to select the hypothesis with the largest posterior probability.
- However, this loss function is hard to defend as being realistic.
- If we choose H_i and it turns out that $\theta \notin H_i$ then the inference is wrong and the loss is the same irrespective of where θ lies.
- An alternative approach is to co-opt the theory of set estimators.
- The statistician can use her set estimator δ to make at least some distinctions between the members of $\mathcal{H}$:
  - Accept H_i exactly when $\delta(x) \subset H_i$,
  - Reject H_i exactly when $\delta(x) \cap H_i = \emptyset$,
  - Undecided about H_i otherwise.