Fist exam (A1)

Class: Bayesian Statistics
Instructor: Luiz Max Carvalho
TA: Isaque Pim

22 May 2024

- You have 4 (four) hours to complete the exam;
- Please read through the whole exam before you start giving your answers;
- Answer all questions briefly;
- Clealy mark your final answer with a square, circle or preferred geometric figure;
- The exam is worth $\min\{\text{your score}, 100\}$ marks.
- You can bring **one** **"cheat sheet"** A4 both sides, which must be turned in together with your answers.

1. I like 'em short.

For a prior distribution π, a set C_x is said to be an α-credible set if

$$P^\pi(\theta \in C_x | x) \geq 1 - \alpha.$$

This region is called an HPD α-credible region (for highest posterior density) if it can be written in the form:

$$\{\theta; \pi(\theta|x) > k_\alpha\} \subset C_x^\pi \subset \{\theta; \pi(\theta|x) \geq k_\alpha\},$$

where k_α is the largest bound such that $P^\pi(\theta \in C_x^\alpha | x) \geq 1 - \alpha$. This construction is motivated by the fact that they minimise the volume among α-credible regions. A special and important case are *HPD intervals*, when C_x is an interval (a, b).

a) (20 marks) Show that if the posterior density (i) is unimodal and (ii) never uniform for all intervals of $(1 - \alpha)$ probability mass of Ω, then the HPD region is an interval and it is unique.

   Hint: formulate a minimisation problem on two variables a and b with a probability restriction and solve for the Lagrangian.

b) (20 marks) We can also use decision-theoretical criteria to pick between credible intervals. A first idea is to balance between the volume of the region and coverage guarantees through the loss function

$$L(\theta, C) = \mathrm{vol}(C) + \mathbb{1}_{C^c}(\theta).$$

   Explain why the above loss is problematic.

c) * (20 bonus marks) Define the new loss function

$$L^*(\theta, C) = g\left(\mathrm{vol}(C)\right) + \mathbb{1}_{C^c}(\theta),$$

   where g is increasing and $0 \leq g(t) \leq 1$ for all t. Show that the Bayes estimator C_x^π for L^* is a HPD region.

Concepts: highest posterior density; interval estimation, loss function. **Difficulty**: intermediate.
Resolution:

a) Let b and a be the upper and lower bounds of our interval and let $\pi(\theta|x)$ be the posterior distribution. We seek to minimise the quantity $b - a$. Adding the probability restrictions we get:

$$\begin{aligned} \min_{b,a} \quad & b - a \\ \text{s.t.} \quad & \int_a^b \pi(\theta|x) d\theta = 1 - \alpha. \end{aligned}$$

   The Lagrangian can then be written as

$$\mathcal{L} = (b - a) + \lambda \left[\int_a^b \pi(\theta|x) d\theta - (1 - \alpha) \right].$$

Differentiate w.r.t b and a and set the results to zero to get

$$\frac{\partial \mathcal{L}}{\partial a} = -1 - \lambda \pi(a|x) = 0,$$
$$\frac{\partial \mathcal{L}}{\partial b} = 1 - \lambda \pi(b|x) = 0.$$

From this we get $\pi(a|x) = \pi(b|x) = -1/\lambda$. As π is a probability density, $\lambda < 0$. Note that the density on both ends of our interval must be equal, which makes sense according to our definition. The second order conditions gives us that

$$\frac{\partial^2 \mathcal{L}}{(\partial a)^2} = -\lambda \frac{\partial \pi(a|x)}{\partial a} \quad ; \quad \frac{\partial^2 \mathcal{L}}{(\partial a)^2} = -\lambda \frac{\partial \pi(b|x)}{\partial b} \quad ; \quad \frac{\partial^2 \mathcal{L}}{\partial a \partial b} = 0$$

Since the posterior density is unimodal and non-constant, we must have have that the derivative at a is positive and the derivative at b is negative. Then the Hessian matrix of second derivatives is positive definite, which implies we have achieved a minimum for the interval (a, b).

b) It is problematic because volume and coverage are not on the same scale. If the volume needs to be large to ensure coverage, is better to pick a region with null volume. For example, consider the case of finding a HPD for the mean of a normal distribution. Under Jeffrey's prior the HPD will be the classical t interval

$$C(\bar{x}, \bar{s}^2) = \left(\bar{x} - t_\alpha \sqrt{\frac{\bar{s}^2}{n}}, \bar{x} + t_\alpha \sqrt{\frac{\bar{s}^2}{n}} \right).$$

The volume of the HPD above is twice the standard deviation term. If this volume is larger than 1, it is better to collapse the interval to a point if we are trying to minimise the loss. So the interval under this loss becomes

$$C'(\bar{x}, \bar{s}^2) = \begin{cases} C(\bar{x}, \bar{s}^2), & \sqrt{\bar{s}^2} > \sqrt{n}/2t_\alpha, \\ \{\bar{x}\}, & \text{otherwise}. \end{cases}$$

This makes little sense, as one deposits essentially infinite certainty on a single point. See Section 5.5.3 in Robert (2007) for more details.

c) Let C^π be the Bayes estimator under the given loss. By definition C^π minimises the posterior expected loss

$$R(C|x) = \mathbb{E}\left[L^*(C, \theta|x)\right] = g(\text{vol(C)}) + \int_{C^c} \pi(\theta|x) d\theta,$$

which is equivalent to finding C that minimises

$$g(\text{vol(C)}) - \int_C \pi(\theta|x) d\theta.$$

If the Bayes estimator is not an HPD, there exists $k \geq 0$ such that

$$C^\pi \cap \{\theta : \pi(\theta|x) < k\} \neq \emptyset \quad \text{and} \quad (C^\pi)^c \cap \{\theta : \pi(\theta|x) \geq k\} \neq \emptyset,$$

the intersections being different from zero (we are working with sets defined only up to sets of Lebesgue measure zero). Thus, there exists sets A and B such that

$$A \subset C^\pi \cap \{\theta : \pi(\theta|x) < k\} \quad \text{and} \quad B \subset (C^\pi)^c \cap \{\theta : \pi(\theta|x) \geq k\},$$

and $\text{vol}(A) = \text{vol}(B) > 0$. If we now define $C^* = (C^\pi - A) \cup B$, it follows that

$$R(C^\pi|x) > R(C^*|x),$$

as $\text{vol}(C^\pi) = \text{vol}(C^*)$ and $\int_A \pi(\theta|x)d\theta < \int_B \pi(\theta|x)d\theta$. Therefore we have a contradiction, so C^π must be an HPD.

■

Comment: Here we saw how to frame the problem of interval inference – from a unimodal posterior – as an optimisation problem, which under regularity conditions yields a well-behaved solution. Moreover, we saw that a loss function that makes intuitive sense might lead to strange conclusions. Finally, we proved a little result that characterises the HPD as the solution of a particular class of problems, where the volume of the resulting estimate (interval) is transformed through an increasing function, generalising the previous finding.

2. Savage!

We will now study the case of point hypothesis testing as a case of two nested models. Let $\theta_0 \in \Omega_0 \subset \Omega$. We want to compare model $M_0 : \theta = \theta_0$ to $M_1 : \theta \in \Omega$. That is, under model M_1, θ can vary freely. Assume further that the models are *properly nested*, that is,

$$P(x|\theta, M_0) = P(x|\theta = \theta_0, M_1).$$

a) (25 marks) Given observed data x, show that the Bayes Factor BF_{01} can be written as

$$\text{BF}_{01} = \frac{p(\theta_0|x, M_1)}{p(\theta_0|M_1)},$$

where the numerator is the posterior under M_1 and the denominator the prior probability under M_1.

b) (25 marks) Apply the result from part (a) to the problem of testing whether a coin is fair. Specifically, we want to compare $H_0 : \theta = 0.5$ against $H_1 : \theta \neq 0.5$, where theta is the probability of the coin landing heads. Given $n = 24$ trials and $x = 3$ heads and employing a uniform prior on θ, calculate the Bayes factor BF_{01}. Based on the Bayes factor, would you prefer H_0 over H_1? How strong should the prior be for a change in this preference?

Note: The ratio above is called the *Savage-Dickey* ratio. It provides a straightforward way to compute Bayes factors, which can be more intuitive and less computationally intensive than other methods.

Concepts: Bayes factors, priors for testing, Savage-Dickey. **Difficulty**: intermediate.

Resolution:

a) The Bayes Factor is given by

$$\mathrm{BF}_{01} = \frac{p(x|M_0)}{p(x|M_1)}.$$

We can expand the numerator and use the nesting condition to get

$$\begin{aligned} p(x|M_0) &= \int p(x|\theta, M_0)p(\theta|M_0)d\theta \\ &= \int p(x|\theta = \theta_0, M_1)p(\theta|M_0)d\theta \\ &= p(x|\theta = \theta_0, M_1). \end{aligned}$$

Now, using Bayes theorem we get

$$p(x|\theta = \theta_0, M_1) = \frac{p(\theta_0|x, M_1)p(x|M_1)}{p(\theta_0|M_1)}.$$

Substitute $p(x|M_0)$ back into our first expression and we get the result

$$\mathrm{BF}_{01} = \frac{p(x|M_0)}{p(x|M_1)} = \frac{p(\theta_0|x, M_1)p(x|M_1)}{p(x|M_1)p(\theta_0|M_1)} = \frac{p(\theta_0|x, M_1)}{p(\theta_0|M_1)}$$

Now we can test point hypothesis by just evaluating the ratio of the prior and the posterior under M_1 on the point representing the null set.

b) The uniform prior is a $\mathrm{Beta}(1, 1)$ distribution, conjugate to the binomial. The posterior is then a $\mathrm{Beta}(x+1, n-x+1)$ distribution. Evaluating the posterior/prior ratio at 1/2 we get

$$\mathrm{BF}_{01} = \frac{\frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)}\frac{1}{2^n}}{\frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}\frac{1}{2}} = \frac{\Gamma(26)}{\Gamma(5)\Gamma(21)}\frac{1}{2^{24}} = \frac{26 \cdot 23 \cdot 22 \cdot 5}{2^{24}}.$$

That is approximately 0.004, so it is 255 times more likely for the coin to be biased than not – which makes perfect sense, since there were only 3 heads out of 24 throws. If we wanted to change this decision we could put aside the idea of nested models and place a point hypothesis for H_1, such as $\theta = 1$. We could keep the nested model and try to concentrate prior density on a point to the right of 1/2. If we use a prior $\mathrm{Beta}(\alpha, \alpha)$ and take α to infinity, it is easy to show that the Bayes Factor converges to 1 – basically prior and posterior will be a point mass at 1/2. Both cases are super strong prior choices.

■

Comment: See Dickey (1971) for more details.

3. Hey, you're biased!

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from an Exponential(θ) distribution with $\theta > 0$ and common density $f(x \mid \theta) = \theta^{-1}\exp(-x/\theta)\mathbb{I}(x > 0)$ w.r.t. the Lebesgue measure on $\mathbb{R}$.

a) (10 marks) Find a conjugate prior for θ;

b) (20 marks) Exhibit the Bayes estimator under quadratic loss for θ, $\delta_B(\boldsymbol{X})$;

c) (10 marks) Show that the bias $\delta_B(\boldsymbol{X})$ is $O(n^{-1})$.

d) $*$ (10 bonus marks) Show how to obtain the uniformly minimum variance unbiased estimator (UMVUE) from $\delta_B(\boldsymbol{X})$ by taking limits of the hyperparameters.

Concepts: Bayes estimator, conjugacy, connections with frequentist/orthodox theory. **Difficulty**: easy.

Resolution: This is a very straightforward question and we shall proceed accordingly. First we note that the assumption that the X_i are conditionally i.i.d given θ leads to

$$\begin{aligned} f(\boldsymbol{X} \mid \theta) &= \prod_{i=1}^{n} \frac{\exp(-X_i/\theta)}{\theta} \mathbb{I}(X_i > 0), \\ &= \theta^{-n} \exp\left(S_n/\theta\right) \mathbb{I}\left(\prod_{i=1}^{n} X_i > 0\right), \end{aligned}$$

where $S_n := \sum_{i=1}^{n} X_i$. From here, there is no point in pretending that we don't know what a good guess for a conjugate family to this likelihood is: an inverse gamma distribution with parameters $\alpha, \beta > 0$ would lead to a posterior

$$\begin{aligned} p(\theta \mid \boldsymbol{X}) &\propto (\boldsymbol{X} \mid \theta)\, \pi(\theta \mid \alpha, \beta), \\ &= \theta^{-n-(\alpha+1)} \exp\left(S_n/\theta + \beta/\theta\right) \mathbb{I}\left(\prod_{i=1}^{n} X_i > 0\right), \end{aligned}$$

which, after re-arranging, can be recognised as the kernel of an inverse gamma distribution with parameters $\alpha_n = n + \alpha$ and $\beta_n = S_n + \beta$. To answer b), we need to remember that the Bayes estimator under quadratic loss is the posterior mean. Thus,

$$\begin{aligned} \delta_B(\boldsymbol{X}_n) &= \frac{\beta_n}{\alpha_n - 1}, \\ &= \frac{n\bar{X}_n + \beta}{n + \alpha - 1}, \end{aligned}$$

where the last line comes from noticing we can write $S_n = n\bar{X}_n$ where $\bar{X}_n$ is the sample mean. To compute the bias, we will take

$$\mathbb{E}_\theta\left[\delta_B(\boldsymbol{X}_n) - \theta\right] = \frac{n + \theta\beta - (n + \alpha - 1)\theta^2}{\theta(n + \alpha - 1)},$$

which is $O(1/n)$, as requested. From orthodox[1] theory we know[2] that the UMVUE for θ is $\bar{X}_n$. So the way to get it from $\delta_B(\boldsymbol{X}_n)$ is to take $\alpha, \beta \to 0$, i.e.,

[1] Frequentist

[2] If you need a refresher, consider: (i) showing that $\bar{X}_n$ is unbiased, computing the Cramér-Rao lower bound for unbiased estimators and showing that its variance matches the bound or (ii) noticing that S_n is complete suficient and using Lehmann-Scheffé or, yet, (iii) noticing that the exponential distribution belongs to the exponential family – in canonical form – and thus the sample mean is UMVUE.

to "flatten" out the prior so it approaches the (improper) uniform on $\mathbb{R}_+$. ■

Comment: This is a very straightforward question just to make sure we know our basics. There is some interesting discussion about the relationship with frequentist estimation if we consider other estimands. Consider estimating $\eta_t := \exp(t/\theta)$ for some $t > 0$, for instance. In this case we can show[3] that the Bayes estimator under quadratic loss is

$$\tilde{\delta}_B(\boldsymbol{X}_n) = \left(1 + \frac{t}{n\bar{X}_n + \beta}\right)^{-(n+\alpha)},$$

which is biased but consistent. The UMVUE is ,

$$\tilde{\delta}_{\text{UMVUE}}(\boldsymbol{X}_n) = \left(1 - \frac{t}{n\bar{X}_n}\right)^{-n},$$

however, so it is not a limit of Bayes estimators of the sort we considered – or any for that matter. See example 4.7 (page 242) in Shao (2003).

[3] Just consider the moment-generating function of an inverse-gamma distribution.

Bibliography

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, pages 204–223.

Robert, C. P. et al. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer.

Shao, J. (2003). *Mathematical Statistics*. Springer Science & Business Media.