Bayesian statistics problem set 2

Florent Leclercq[a)]

CNRS & Sorbonne Université, UMR 7095, Institut d'Astrophysique de Paris, 98 bis boulevard Arago, F-75014 Paris, France

(Dated: 6 April 2025)

I. THE SLOPE OF THE NUMBER COUNTS OF RADIO SOURCES

The flux densities of extragalactic radio sources is distributed as a power-law with slope $-\alpha$. In a non-evolving Euclidean universe $\alpha = \frac{3}{2}$ (can you prove this?) and a deviation of α from $\frac{3}{2}$ is evidence for cosmological evolution of radio sources. This was the most telling argument against the steady-state cosmology in the early 1960s (even though they got the value of α wrong by quite a long way).

1. Derivation of the probability distribution. Assuming a known, fixed detection limit S_0, and $\alpha > 1$, find the expression for the normalised probability density function (pdf) of the flux density of radio sources.

2. Derivation of the posterior distribution for α. Assume that we have n independent and error-free observations $\{S_i\}_{i=1}^n$ with each $S_i \geq S_0$. By treating the flux density distribution as a pdf and adopting a uniform prior for α, derive the likelihood and posterior functions based on the observations. What is the MAP (maximum a posteriori) value of α (i.e. the value of α that maximises the posterior)?

3. Inference of α. If a single source is observed with flux $S_1 = 2S_0$, what is the most probable value of α? By evaluating the curvature (i.e., the second derivative) of the log-posterior at its maximum, derive the approximate uncertainty (standard deviation) on α. Explain why it is possible to infer the slope from only one object.

II. BAYESIAN DECISION THEORY

You've completed a Bayesian analysis of a problem, obtaining the posterior probability density function (pdf) $p(\theta|d)$. However, your boss wants a single number, and you need to decide which value to report. Your boss has offered a bonus of 1000 euros for your analysis, but any error in your reported value will result in a penalty. The penalty will be the square of the error, converted to euros, and deducted from your next month's paycheck. The true value of θ will be revealed next month.

You need to determine the optimal value to report, $\theta_\star$, given the uncertainty represented by the posterior pdf. This is a classic problem of decision-making under uncertainty.

1. Find the expected net gain as a function of the reported value, $\theta_\star$.

2. Determine the value of $\theta_\star$ that maximises the expected net gain.

3. Check that the optimal reported value is equal to the posterior mean.

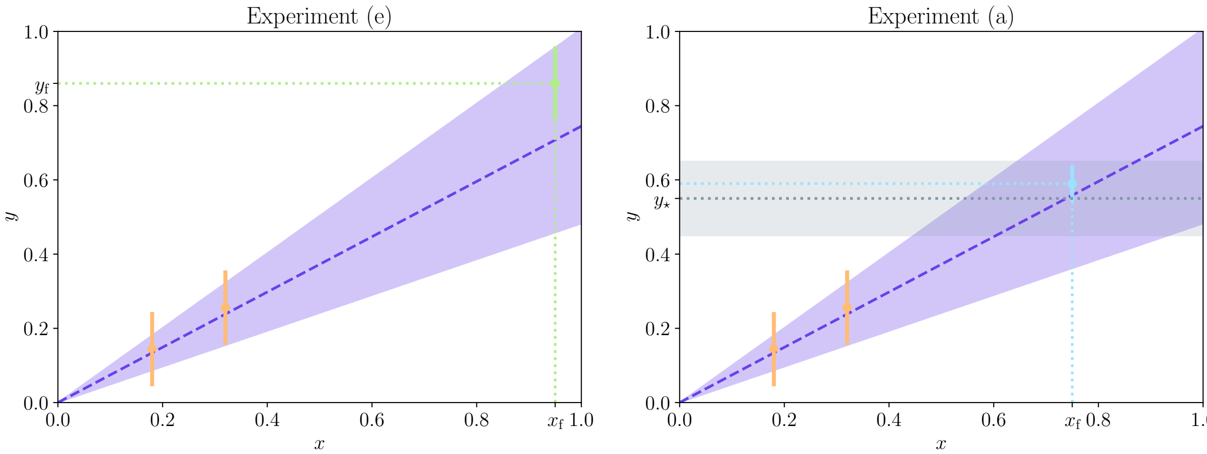
III. BAYESIAN LINEAR MODEL: EXPERIMENTAL DESIGN, HIERARCHY, AND PREDICTION

We are interested in a model $y = mx$. We want to measure the slope m of this relationship.

1. Bayesian experimental design (exercise inspired by **?**). Suppose that we have measured two points y_0 and y_1 with error σ at two locations x_0 and x_1. We now have the choice between two (equally expensive) experiments:

   - **Instrument (e):** As accurate as today's instrument, it will measure y_f at a much larger value x_f (so as to increase the lever arm in the measurement of the slope).
   - **Instrument (a):** Much more accurate instrument, but built so as to have a "sweet spot" at a certain value of y, called $y_\star$, and much less accurate elsewhere.

[a)] Electronic mail: florent.leclercq@iap.fr; https://www.florent-leclercq.eu/; ORCID: 0000-0002-9339-1404



Which instrument should we go for? The answer should probably depend on how good our current knowledge of m is. Is the current uncertainty on m small enough to target accurately enough $x = x_\star$ so that we get to the sweet spot $y_\star = mx_\star$? We can use for the utility of experiments (e) and (a) the inverse variance of the future posterior on m.

(a) Write down the inverse variance of the future posterior on m.

(b) Denoting by o the current experiment, write down the utility for experiment (e), $U(e|o)$.

(c) Assuming for the noise levels of instrument (a) the toy model:

$$\tau_a^2 = \tau_\star^2 \exp\left[\frac{(y-y_\star)^2}{2\Delta^2}\right], \tag{1}$$

where Δ is the width of the sweet spot, show that the utility for experiment (a) is:

$$U(a|o) = \frac{x_\mathrm{f}^2}{\tau_\star^2} \exp\left[-\frac{1}{2}\frac{(\bar{m}x_\mathrm{f} - y_\star)^2}{\Delta^2 + \Delta_y^2}\right] + \text{constant}, \tag{2}$$

where $\bar{m} \equiv \langle m \rangle_{p(m|o)}$ and Δ_y is the uncertainty at x_f.

(d) Discuss the best choice of experiment in the two cases where $\Delta \gg \Delta_y$ and $\Delta \ll \Delta_y$.

(e) Conclude on the opportunity of designating an experiment that exploits a "sweet spot".

2. Bayesian hierarchical model. Suppose that we measure X and Y, but they both have Gaussian measurement errors, σ_x and σ_y, respectively. We want to infer the slope m given X and Y.

(a) Build a Bayesian hierarchical model (BHM) for the problem, with two latent variables x and y for the "true" values of the measured quantities.

(b) Assuming a uniform prior on m and x, write down an expression for the (unnormalised) marginal posterior on m, $p(m|X,Y)$.

(c) Write the joint posterior on x and m, $p(x,m|X,Y)$. What can be said about the conditionals $p(x|X,Y,m)$ and $p(m|X,Y,x)$? Exploit this property to build an appropriate sampler for the joint distribution. Write a piece of code to check your results.

3. Posterior predictive test. Let us now use the BHM for making predictions of new data.

(a) Write down the formal form for the posterior predictive distribution for a new data point $(\tilde{X},\tilde{Y})$, $p(\tilde{X},\tilde{Y}|X,Y)$. Using the previous result, how can we draw samples from $p(\tilde{X},\tilde{Y}|X,Y)$?

(b) We now want to acquire new data at another location $\tilde{x}$. What is the posterior predictive distribution $p(\tilde{X},\tilde{Y}|\tilde{x},X,Y)$? Evaluate it on a grid and plot its marginals for a few values of $\tilde{x}$.

IV. INFORMATION AND ENTROPY

We wish to compare, on a logarithmic scale (in "thermodynamic units") the amount of information (or equivalently, the logarithm of the number of microstates) in various systems. (In what follows information is measured in units of entropy, i.e. using the Boltzmann relation

$$S = k_\mathrm{B} \ln \Omega, \tag{3}$$

so that one "bit" corresponds to an entropy of $k_\mathrm{B} \ln 2$.)

1. Assuming that the macrostate describing 1 mg of water at ordinary temperature corresponds to a certain number of equally probable quantum microstates, evaluate this number. The entropy of water is given as $70\,\mathrm{J}\cdot\mathrm{mol}^{-1}\cdot\mathrm{K}^{-1}$, the molar mass of water $M_{\mathrm{H2O}} \approx 18\,\mathrm{g}\cdot\mathrm{mol}^{-1}$, and the Boltzmann constant $k_{\mathrm{B}} = 1.38 \times 10^{-23}\ \mathrm{J}\cdot\mathrm{K}^{-1}$.

2. Estimate the order of magnitude of the information content in the Bibliothèque Nationale de France in thermodynamic units. Assume the library contains approximately 7×10^{6} volumes, with an average of 500 pages per volume and 1000 characters per page. Compare this with the entropy of 1 mg of water.

3. Compare the above with the order of magnitude of genetic information per individual. Human DNA contains 3×10^{9} base pairs, with 4 possible choices for each base pair.

4. The brain contains 10^{10} neurons, each forming synapses with 1000 neighboring neurons. A synapse, the connection between two neurons, can be in one of two states: inhibitory or excitatory. Determine the amount of information required to describe the state of all synapses in the brain. Now suppose that for each neuron, we must choose the 1000 neurons to which it is connected. What additional amount of information would this require?

5. Plot the quantities of information obtained in the previous questions on a logarithmic scale. Comment on your findings. What do you think about the possibility that the entire wiring of the brain is contained within the genetic material?

ACKNOWLEDGEMENTS

I thank Benjamin Wandelt, Alan Heavens and colleagues of the SOC of the ICIC Data Analysis workshop 2021 for their own lectures, from which this document is inspired.