Introduction to Bayesian probability theory

AA20-21 364BB

Walter Del Pozzo

Problems discussion

Each student should choose one problem and attempt to solve it using any of the methods explored during the course plus any knowledge of physics he/she obtained during his/her studies. Each problem will involve independent research and a fair amount of thinking. The purpose of the exam will be to present and discuss the proposed solution, but mostly concentrating on the rationale behind any choice and assumption made in tackling the problem. The problems assigned are difficult and could be easily considered master thesis projects. Sometimes they are an open field of research, therefore I do not expect a full solution to any of them. Discussion is encouraged, but group solution is not. Each student is graded individually and thus I would expect each student to work on his/her own. For the exam itself, each student is encouraged to prepare a presentation in any format he/she prefers (slides, real time writing with a graphic tablet, etc) detailing his/her reasoning and solution. Students should aim for a 30 minutes presentation to be given in English. Also, note that some of the proposed readings in the previous section might contain useful information to solve the problems.

Determination of the orbit of an asteroid

A threshold for the completion of the exercise is to complete points 1), 2) and 3).
Several sky surveys constantly monitor the sky for potentially hazardous asteroids. Measurements consist of a time series of the angular position of the asteroid in the sky (right ascension α and declination δ) and magnitude M. The purpose of this project is to construct a model to be able to infer the orbit of the asteroid starting from the observables (α, δ, M). Apply the inference model developed to the data (from https://newton.spacedys.com/neodys/index.php?pc=0) available in https://newton.spacedys.com/~neodys2/mpcobs/2020JP.rwo.

Begin by considering the asteroid a material point (ignoring its rotational dynamics and consequent variations in albedo). Also, ignore three body interactions by considering the asteroid and the Earth an isolated system. In this case:

1. compute the posterior distribution on the relevant orbital parameters
2. discuss a way of visualising the evolution of orbit and relative error
3. how does the prediction accuracy evolve in time?

Include now the effect of the Earth orbital motion around the Sun. Can you apply the restricted three body model to the systems? If not, how would you tackle the problem of modelling the orbital evolution? Compare the predictions from your inference model to the previous case.

Generalise now to include also the Earth-Moon system dynamics in the picture. Can you develop an analytical model or do you have to resort to full numerical integration of the equations of motion? What about the effects of the remaining planets? Are they negligible?

Finally, we ignored the rotational dynamics of the asteroid itself. How would you include it in the model for the magnitude? Compare models with and without asteroid rotation in the isolated Earth-asteroid system.

Detection of transient signals in time series: application to gravitational wave data

A threshold for the completion of the exercise is to complete points 1), 2), 3) and 4).
We want to be able to deal with the problem of detecting a gravitational wave signal in the noisy data stream of an interferometric detector. We will begin with understanding and modelling the noise process. We will concentrate on the LIGO Hanford data corresponding to the first detection of a gravitational wave, GW150914. The data, strain as a function of time, can be found here: https://www.gw-openscience.org/eventapi/html/GWTC-1-confident/GW150914/v3/H-H1_GWOSC_4KHZ_R1-1126257415-4096.txt.gz .
Look at the data, model them as a stochastic process; for the entire 4096 seconds, determine

1. mean function and two-point autocorrelation; is the process stationary? If the answer is no, is there a timescale over which the process can be considered stationary? What about wide-sense stationarity?
2. Assuming wide-sense stationarity and knowledge of only the two-point autocorrelation function, write the expression of the noise (errors) distribution and assuming a 2PN waveform model (see the TaylorF2 frequency domain waveform model in https://arxiv.org/abs/0907.0700), write the likelihood function for the hypothesis "signal + noise".
3. Derive the so-called "matched filter" or Wiener filter by maximising the likelihood function.

The waveform model we are using depends on the physical parameters of the binary black hole system, in particular on the two masses and the two spins. Ignore the spin dependence (set them to zero), We want now to construct a template bank, by creating a grid of waveforms in the space of two mass parameters. It is convenient to parametrise the problem in terms of the chirp mass $\mathscr{M} = \frac{(m_1 m_2)^{3/5}}{(m_1+m_2)^{1/5}}$ and the mass ratio $q = m_1/m_2 \quad m_1 > m_2$. Assuming a uniform prior distribution for the component masses:

4. derive the prior distribution over chirp mass and mass ratio.
5. construct the template bank, generating waveforms in bins with a size of $0.1 M_\odot$ for the chirp mass and 0.05 for the mass ratio and, for simplicity, assume $\mathscr{M} \in (25,35)$ and $q \in (0.5,1.0)$.

- Given your template bank, can you determine the most probable value of the mass parameters for GW150914 and the time of the wave maximum in Hanford?

Suggestion: in all calculations, just parametrise the wave with a global amplitude parameter A, so that the gravitational wave in the frequency domain is $A \exp[i\Phi(f)]$ with the phase taken from https://arxiv.org/abs/0907.0700

Fitting an isochrone

A threshold for the completion of the exercise is to complete points 1) and 2).
Among one of the most interesting problems in astronomy, is the determination of stealer populations and stellar ages that are a cornerstone to test our understanding of nuclear physics and plasma physics as well having repercussions on large scale astrophysics. One of the problems is, however, isolating homogeneous populations. This relates to the problem of determining, from an astrometric and photometric survey, whether a star belongs to a cluster, and thus it is a sample of the same initial conditions, or if it belongs to a contaminating back/foreground.

1. For each star, determine the probability that it belongs to a stellar cluster or not.

We are going to use data from a combination of GAIA and 2MASS observations. Using the data linked here (https://elearning.df.unipi.it/mod/folder/view.php?id=5019)

2. determine the likelihood function for the absolute magnitude of each star in the cluster. How do you convert between observed magnitudes and distances?

Empty space, is not empty. Clouds of dust and gas absorb part of the radiation coming from each star and thus changes the reconstructed luminosity. This is the so-called reddening and the effect is to make stars look dimmer than they are.

3. Include the effect of reddening E(B-V) in the likelihood function determined above. Use the following values for the extinction corrections in the bands H, J and K:

cor_J = 8.845323e-01
cor_H = 5.584712e-01
cor_Ks = 3.609361e-01

so that ext(i) = E(B-V) * cor_i.

Having set up the likelihood for the magnitude, and for the belonging to the cluster or not, we are in the position to be able to fit a isochrone model. An isochrone is a theoretical prediction constructed from the ground up using state-of-the-art stellar physics models to predict the distribution in the color-magnitude (HR) diagram of stars. Using the precomputed models in (https://elearning.df.unipi.it/mod/folder/view.php?id=5019), determine the mode of the posterior distribution over age, reddening and distance modulus and estimate the 90% credible region.

The black hole mass - velocity dispersion relation

A threshold for the completion of the exercise is to complete points 1), 2) and 3).

The discovery of a relation between bulge velocity dispersion and central black hole mass is a founding piece of evidence of the co-evolution of super-massive black holes and their hosts. In this exercise, we will infer the relation from data coming from https://academic.oup.com/mnrasl/article/426/1/L51/988535.

1. Take the measurements of M_{BH} and σ_{GC} given in https://academic.oup.com/mnrasl/article/426/1/L51/988535, Table 1 and construct an inference model, considering the uncertainties both on the black hole mass and on the bulge velocity dispersion. Assume the distribution of the errors on both quantities to be a Gaussian with known variance — average the errors reported in the table — and fixed for each data point. Assume a linear relation between $\log(M_{BH}/M_{\odot})$ and $\log\sigma$. Compute the posterior distributions for all relevant parameters.
2. Compare the evidences for a linear model and a quadratic one. Which model is favoured?
3. Let's now relax the hypothesis that the uncertainties are gaussian-distributed with known variance fixed for all data point. Model asymmetric errors by assigning an appropriate error distribution and infer the posterior distribution for all relevant parameters for a linear relation and a quadratic one. Which model is favoured in the case?
4. Let us know introduce the possibility of random scatter in the relation between $\log(M_{BH}/M_{\odot})$ and $\log\sigma$:

$$\log(M_{BH}/M_{\odot}) = a\log\sigma + b + \epsilon$$

with $\epsilon \sim N(0,\Sigma)$. Repeat the analysis in point 3) under this new model. How do the results change?

On polling and election prediction

A threshold for the completion of the exercise is to complete points 1), 2) and 3).
Consider the following scenario: we randomly select a subset of n individuals from a population of N, $n < N$; we find that at an upcoming election, n_A individuals will vote for party A, n_B will vote for party B and n_C will vote for party C, with $n_A + n_B + n_C = n$. All polled individuals say the truth about their vote.

1. Compute the posterior on the probability that each party will win the first round of elections.

Assume now that the people interviewed are not guaranteed to say the truth about their vote: each individual has a known probability p of saying the truth about his vote.

2. Compute the posterior on the probability that each party will win the first round of elections. How do the results change as a function of p?

Let's now assume that the electoral system is a two stage one: at the first step votes are cast, counted and the two parties with the most votes pass to the next stage. Each elector has a definite ranking for casting their votes:

a) group of people that vote for A on the first round will vote for B if A loses the first round
b) group of people that vote for B on the first round will vote for C if B loses the first round
c) group of people that vote for C on the first round will vote for A if C loses the first round

3. What is the posterior probability that each party will win the election?

Assume the following rankings now:

a) group of people that vote for A on the first round will vote for B with probability p_1 if A loses the first round
b) group of people that vote for B on the first round will vote for C with probability p_2 if B loses the first round
c) group of people that vote for C on the first round will vote for A with probability p_3 if C loses the first round

4. What is the posterior probability that each party will win the election in this case?

Spatio-temporal coincidences

A threshold for the completion of the exercise is to complete points 1), 2), 3) and 4).
Let's begin by considering a spatio-temporal point process in the sphere. Consider the following proposition:

E_1: "an event happened in $\in [t, t+dt]$ and $\in [V, V+dV]$"

So that we can write $p(E_1 \mid r_1(V,t)\, I) = r_1(V,t)dtdV$.

1. Derive the probability to observe n events in a time interval T within a volume V.
2. Assume a Euclidean volume in spherical coordinates, what is the probability of observing 1 event within a given area ΔA? Is the distribution still a Poisson distribution?

Consider now a second spatio-temporal point process, such that

E_2: "an event happened in $\in [t, t+dt]$ and $\in [V, V+dV]$"

So that we can write $p(E_2 \mid r(V,t)\, I) = r_2(V,t)dtdV$.

3. Calculate the probability of observing 1 event from class E_1 and 1 event from class E_2 within a given surface area ΔA, assuming that the two classes are independent. Assume constant and homogeneous rate parameters.

Assume now that the two classes are correlated, so that the occurrence of 1 event in one class affects the probability of 1 event occurring in the other class, e.g. if

$$p(1_{E_1} \mid r_1(V,t)\, I) = p_1 dVdt$$
$$p(1_{E_2} \mid r_2(V,t)\, I) = p_2 dVdt$$

then the conditional probabilities become

$$p(1_{E_2} \mid 1_{E_1} r_2(V,t)\, I) = (p_2 + \delta)dVdt$$
$$p(1_{E_1} \mid 1_{E_2} r_1(V,t)\, I) = (p_1 + \epsilon)dVdt$$

4. Compute the probability of observing 1 event from class E_1 and 1 event from class E_2 within a given area ΔA.

5. Generalise the results in points 1), 2), 3) and 4) assuming that each event class can be in an ON or OFF state (hint: assume that each of the rate parameters is given by a sum of an active rate $s_i(V,t)$ and an inactive rate $n_i(V,t)$)

6. Generalise all the above results for a Friedmann-Robertson-Walker-LeMaitre universe.

Gamma-ray burst classification

A threshold for the completion of the exercise is to complete points 1a), 2), 3) and 4).

Gamma-ray bursts (GRBs) are usually classified into two different categories, according to their duration: *short* GRBs and *long* GRBs.

The distribution of the burst duration, T_{90}, is usually modeled as the weighted sum of two log-normal distributions:

$$p(T_{90}) = w_1 N(\log(T_{90})|\mu_1, \sigma_1) + w_2 N(\log(T_{90})\,|\mu_2, \sigma_2)\,.$$

We will make use of Fermi/GBM data available here (LINK). Unless stated, we will neglect measurement uncertainties.

1. Determine:
    a. The parameters of the distribution.
    b. As above, assuming Gaussian uncertainties on each $\log(T_{90})$.

Assuming the parameters inferred in the previous point, we now turn our attention to the problem of classifying each GRB.

2. Compute the probability of GRB170817A ($T_{90} = 2.0$ s) of being a short or long GRB.
3. Decide a figure of merit and determine the threshold value for T_{90} to discriminate between short and long GRBs.

Some authors propose a third class of GRBs, *intermediate*.

4. Which of the two hypothesis is favored according to the available data?

Another possibility is to classify GRBs in *soft* and *hard*, according to the hardness ratio HR,

$$HR = F_{50-100\,keV}/F_{20-50\,keV}\,,$$

where F_i is the measured flux in a certain energy interval.

5. Study the bimodal distribution in the $\log(T_{90}) - \log\,(HR)$ space. Suggestion: specialize https://dp.tdhopper.com/collapsed-gibbs/ to a fixed number of components.

Differential diagnosis

A threshold for the completion of this exercise is to complete points 1) and 2). Differential diagnosis is the medical practice of diagnose a disease by using evidence such as symptoms or the patient's health history.

Let's assume of having access to a medical database that contains all the information on the patients treated by a hospital up to a certain date.
In particular, this database contains the symptoms and the disease of each patient. For simplicity, we will assume that all these diagnoses are correct and the symptoms are uncorrelated with each other.

A new patient shows up with a list of symptoms.

1. Assuming that the hospital treated at least once every existing disease and observed all the possible symptoms, give the posterior probability for each disease conditioned on the symptoms the patient has.
2. Same as point 1, but accounting for the possibility of testing the patient for certain disease. Keep in mind that tests might give false positive/negatives.
3. Same as point 1, but accounting for the possibility of a new, previously unobserved, disease.