Bayesian inference and model comparison problems

Roberto Trotta, Imperial College London
International elite PhD Course
Niels Bohr Institute, Copenhagen, 6-10th Oct 2014

September 26, 2014

Problems are divided in "introductory", "intermediate" and "advanced" level. I suggest you work on the introductory problems first unless you are already familiar with the techniques and concepts used.

You should not try and do everything in the time available! Focus on the subject/problem that interests you the most and try to get to the bottom of it during the tutorial session. Work through the rest at your own pace.

A dagger (†) denotes harder questions, for aficionados.

1 Bayesian reasoning

1.1 Introductory level

1. **Medical evidence.** A batch of chemistry undergraduates are screened for a dangerous medical condition called *Bacillum Bayesianum* (BB). The incidence of the condition in the population (i.e., the probability that a randomly selected person has the disease) is estimated at about 1%. If the person has BB, the test returns positive 95% of the time. There is also a known 5% rate of false positives, i.e. the test returning positive even if the person is free from BB. One of your friends takes the test and it comes back positive. Here we examine whether your friend should be worried about her health.

   (a) Translate the information above in suitably defined conditional probabilities. The two relevant propositions here are whether the test returns positive (denote this with a + symbol) and whether the person is actually sick (denote this with the symbol $BB = 1$. Denote the case when the person is healthy as $BB = 0$).

   (b) Compute the conditional probability that your friend is sick, knowing that she has tested positive, i.e., find $P(BB = 1|+)$.

   (c) Imagine screening the general population for a very rare desease, whose incidence in the population is 10^{-6} (i.e., one person in a million has the disease on average, i.e. $P(BB = 1) = 10^{-6}$). What should the reliability of the test (i.e., $P(+|BB = 1)$) be if we want to make sure that the probability of actually having the disease after testing positive is at least 99%? Assume first that the false positive rate $P(+|BB = 0)$ (i.e, the probability of testing positive while healthy), is 5% as in part (a). What can you conclude about the feasibility of such a test?

   (d) Now we write the false positive rate as $P(+|BB = 0) = 1 - P(-|BB = 0)$. It is reasonable to assume (although this is not true in general) that $P(-|BB = 0) = P(+|BB = 1)$, i.e. the probability of getting a positive result if you have the disease is the same as the probability of getting a negative result if you don't have it. Find the requested reliability of the test (i.e., $P(+|BB = 1)$) so that the probability of actually having the disease after testing positive is at least 99% in this case. Comment on whether you think a test with this reliability is practically feasible.

2. **The three doors problem.** In a game, you can pick one of three doors, labelled A, B and C. Behind one of the three doors lies a highly desirable price, such as for example a cricket bat. After you have picked one door (e.g., door A) the person who is presenting the game opens one of the remaining 2 doors so as to reveal that there is no prize behind it (e.g., door C might be opened). Notice that the gameshow presenter *knows* that the door he opens has no prize behind it. At this point you can either stick with your original choice (door A) or switch to the door which remains closed (door B). At the end, all doors are opened, at which point you will only win if the prize is behind your chosen door.

   (a) Given the above rules (and your full knowledge of them), should you stick with your choice or is it better to switch?

   (b) In a variation, you are given the choice to randomly pick one of doors B or C and to open it, after you have chosen door A. You pick door C, and upon opening it you discover there is nothing behind it. At this point you are again free to either stick with door A or to switch to door B. Are the probabilities different from the previous scenario? Justify your answers.

3. **Top Scientists on Twitter.** A Twitter survey of the "Top 50 Science Stars on Twitter" claims that "most high-performing scientists have not embraced Twitter"[1] . That article is debatable on other grounds, as well, in particular in terms of what defines a "Top Scientist" on Twitter. In fact, on closer inspection, the data on which this strong statement is based are fairly debatable, having been obtained by "sampl[ing] Twitter usage among 50 randomly chosen living scientists from the Scholarometer list", which is arguably not a great statistics.

   Even so, it is interesting to use some real maths to answer the question: Assuming it is true that top scientists shun Twitter, does being on Twitter make me (statistically) less of a good scientist? In other words, is it more probable for me to be a "mediocre" scientist if I'm a Twitter user?

   Use Bayes theorem to estimate the probability of being a Top Scientists (TS) given that one is a Twitter User (TU), i.e. the quantity $P(\mathrm{TS}|\mathrm{TU})$.

   For the sake of definiteness, define "Top Scientists" as somebody in the top 10% of their discipline. Also, use the fact that (according to the same source) "only a fifth of scientists have an identifiable Twitter profile."

   Make reasonable assumptions about the other probabilities you need, and evaluate the sensitivity of your answer on those assumptions.

1.2 Intermediate level

Bayes in politics. In a TV debate, politician A affirms that a certain proposition S is true. You trust politician A to tell the truth with probability $4/5$. Politician B then agrees that what politician A has said is indeed true. Your trust in politician B is much weaker, and you estimate that he lies with probability $3/4$.

After you have heard politician B, what is the probability that statement S is indeed true?

(You may assume that you have no other information on the truth of proposition S other than what you heard from politicians A and B)

Hint: Start by denoting by A_T the statement "politician A tells the truth", and by B_T the statement "politician B tells the truth". What you are after is the probability of the statement "proposition S is true" after you have heard politician B say so.

[1] http://news.sciencemag.org/scientific-community/2014/09/top-50-science-stars-twitter, last accessed on Sept 26th 2014.

2 Parameter inference

2.1 Introductory level

Coin tossing. This is a traditional example, but this time you'll do it in the Bayesian way. A coin is tossed N times and heads come up H times.

1. What is the likelihood function? Identify clearly the parameter, θ, and the data.
2. What is a reasonable, non-informative prior on θ?
3. Compute the posterior probability for θ. Recall that θ is the probability that a single flip will give heads. This integral will prove useful:
$$\int_0^1 d\theta\theta^N(1-\theta)^M = \frac{\Gamma(N+1)\Gamma(M+1)}{\Gamma(N+M+2)}. \tag{1}$$
4. Determine the posterior mean and standard deviation of θ.
5. Plot your results as a function of H for $N = 10, 100, 1000$.
6. † Generalize your prior to the Beta distribution,
$$p(\theta|\nu_1,\nu_2) = \frac{1}{B(\nu_1,\nu_2)}\theta^{\nu_1-1}(1-\theta)^{\nu_2-1} \tag{2}$$
where $B(\nu_1,\nu_2) = \Gamma(\nu_1)\Gamma(\nu_2)/\Gamma(\nu_1+\nu_2)$ is the beta function and the "hyperparameters" $\nu_1,\nu_2 > 0$. Clearly, a uniform prior is given by the choice $(\nu_1,\nu_2) = (1,1)$. Evaluate the dependency of your result to the choice of hyperparameters.
7. † What is the probability that the $(N+1)$-th flip will give heads?

2.2 Intermediate level

The Gaussian linear model. As idealised a case as it is, the Gaussian linear model is a great tool to hone your computational skills and intuition. Furthermore, it applies in an approximate way to many cases of interest. We first solve analytically the general problem in n dimensions, and then specialise to the 2-dimensional case for a numerical application.

We consider the following *linear model*
$$y = F\theta + \epsilon \tag{3}$$
where the dependent variable y is a d-dimensional vector of observations (the *data*), $\theta = \{\theta_1,\theta_2,\dots,\theta_n\}$ is a vector of dimension n of unknown parameters that we wish to determine and F is a $d \times n$ matrix of known constants which specify the relation between the input variables θ and the dependent variables y (so-called "design matrix").

In the following, we will specialize to the case where observations $y_i(x)$ are fitted with a linear model of the form $f(x) = \sum_{j=1}^n \theta_j X^j(x)$. Then the matrix F is given by the basis functions X^j evaluated at the locations x_i of the observations, $F_{ij} = X^j(x_i)$.

Furthermore, ϵ is a d-dimensional vector of random variables with zero mean (the *noise*). If we assume that ϵ follows a multivariate Gaussian distribution with uncorrelated covariance matrix $C \equiv \mathrm{diag}(\tau_1^2,\tau_2^2,\dots,\tau_d^2)$, then the likelihood function takes the form
$$p(y|\theta) = \frac{1}{(2\pi)^{d/2}\prod_j \tau_j}\exp\left[-\frac{1}{2}(b-A\theta)^t(b-A\theta)\right], \tag{4}$$
where we have defined $A_{ij} = F_{ij}/\tau_i$ and $b_i = y_i/\tau_i$ where A is a $d \times n$ matrix and b is a d-dimensional vector.

1. Show that the likelihood function can be cast in the form

$$p(y|\theta) = \mathcal{L}_0 \exp\left[-\frac{1}{2}(\theta-\theta_0)^t L(\theta-\theta_0)\right], \tag{5}$$

with the likelihood Fisher matrix L (a $n \times n$ matrix) given by

$$L \equiv A^t A \tag{6}$$

and a normalization constant

$$\mathcal{L}_0 \equiv \frac{1}{(2\pi)^{d/2}\prod_j \tau_j} \exp\left[-\frac{1}{2}(b-A\theta_0)^t(b-A\theta_0)\right]. \tag{7}$$

Here θ_0 denotes the parameter value which maximises the likelihood, given by

$$\theta_0 = L^{-1}A^t b. \tag{8}$$

2. Assume as a prior pdf a multinormal Gaussian distribution with zero mean and the $n \times n$ dimensional prior Fisher information matrix P (recall that that the Fisher information matrix is the inverse of the covariance matrix), *i.e.*

$$p(\theta) = \frac{|P|^{1/2}}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\theta^t P\theta\right], \tag{9}$$

where $|P|$ denotes the determinant of the matrix P.

Show that the posterior distribution for θ is given by multinormal Gaussian with Fisher information matrix $\mathcal{F}$

$$\mathcal{F} = L + P \tag{10}$$

and mean $\bar{\theta}$ given by

$$\bar{\theta} = \mathcal{F}^{-1}L\theta_0. \tag{11}$$

3. Show that the model likelihood (or "Bayesian evidence", *i.e.*, the normalizing constant in Bayes theorem) is given by

$$\begin{aligned} p(y) &= \mathcal{L}_0 \frac{|\mathcal{F}|^{-1/2}}{|P|^{-1/2}} \exp\left[-\frac{1}{2}\theta_0^t(L - L\mathcal{F}^{-1}L)\theta_0\right] \\ &= \mathcal{L}_0 \frac{|\mathcal{F}|^{-1/2}}{|P|^{-1/2}} \exp\left[-\frac{1}{2}(\theta_0^t L\theta_0 - \bar{\theta}^t\mathcal{F}\bar{\theta})\right]. \end{aligned} \tag{12}$$

Hints: recall this standard result for Gaussian integrals:

$$\int \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m})^t\Sigma^{-1}(\mathbf{x}-\mathbf{m})\right]d\mathbf{x} = \sqrt{\det(2\pi\Sigma)} \tag{13}$$

4. Now we specialize to the case $n = 2$, *i.e.*we have two parameters of interest, $\theta = \{\theta_1, \theta_2\}$ and the linear function we want to fit is given by

$$y = \theta_1 + \theta_2 x. \tag{14}$$

(In the formalism above, the basis vectors are $X^1 = 1, X^2 = x$).

The file `LinearModel.txt` contains an array of $d = 10$ measurements $y = \{y_1, y_2, \ldots, y_{10}\}$, together with the values of the independent variable x_i. Assume that the uncertainty in the same for all measurements, i.e. $\tau_i = 0.1$ $(i = 1, \ldots, 10)$. You may further assume that measurements are uncorrelated. The data set is shown in the left panel of Fig. 1

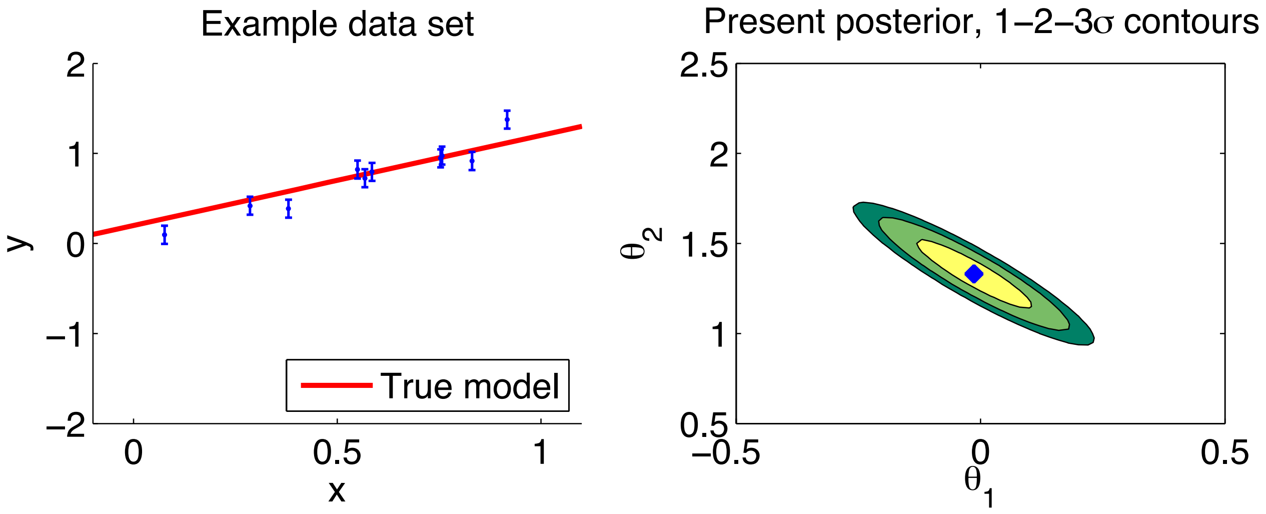

Figure 1: Left panel: data set for the Gaussian linear problem. The solid line shows the true value of the linear model from which the data have been generated, subject to Gaussian noise. Right panel: 2D credible intervals from the posterior distribution for the parameters. The the blue diamond is the Maximum Likelihood Estimator, from Eq. (8), whose value for this data set is $x = -0.0136, y = 1.3312$.

(a) Assume a Gaussian prior with Fisher matrix $P = \mathrm{diag}\left(10^{-2}, 10^{-2}\right)$ for θ.

Find the posterior distribution for θ given the data, and plot it in 2 dimensions in the (θ_1, θ_2) plane (see right panel of Fig. 1 for an example).

Use the appropriate contour levels to demarcate 1, 2 and 3 sigma joint credible intervals of the posterior.

(b) In a language of your choice, write an implementation of the Metropolis-Hastings Markov Chain Monte Carlo algorithm explained in the lecture, and use it to obtain samples from the posterior distribution.

Plot *equal weight* samples in the (θ_1, θ_2) space, as well as marginalized 1-dimensional posterior distributions for each parameter.

(c) Compare the credible intervals that you obtained from the MCMC with the analytical solution obtained above.

2.3 Advanced level

Toy supernovae type Ia cosmology. Supernovae type Ia can be used as standardizable candles to measure distances in the Universe. This series of problems explores the extraction of cosmological information from a simplified SNIa toy model.

The cosmological parameters we are interested in constraining are

$$\mathscr{C} = \{\Omega_m, \Omega_\Lambda, h\} \tag{15}$$

where Ω_m is the matter density (in units of the critical energy density) and Ω_Λ is the dark energy density, assumed here to be in the form of a cosmological constant, i.e. $w = -1$ at all redshifts. In the following, we will fix $h = 0.72$ for simplicity, where the Hubble constant today is given by $H_0 = 100h\mathrm{km/s/Mpc}$.

In an FRW cosmology defined by the parameters $\mathscr{C}$, the distance modulus μ (i.e., the difference between the apparent and absolute magnitudes, $\mu = m - M$) to a SN at redshift z is given by

$$\mu(z, \mathscr{C}) = 5\log\left[\frac{D_L(z, \Omega_m, \Omega_\Lambda, h)}{\mathrm{Mpc}}\right] + 25, \tag{16}$$

where D_L denotes the luminosity distance to the SN. Recalling that $D_L = cd_L/H_0$, We can rewrite this as

$$\mu(z,\mathscr{C}) = \eta + 5\log d_L(z,\Omega_m,\Omega_\Lambda), \tag{17}$$

where

$$\eta = -5\log\frac{100h}{c} + 25 \tag{18}$$

and c is the speed of light in km/s. We have defined the dimensionless luminosity distance

$$d_L(z,\Omega_m,\Omega_\Lambda) = \frac{(1+z)}{\sqrt{|\Omega_\kappa|}}\mathrm{sinn}\{\sqrt{|\Omega_\kappa|}\int_0^z \mathrm{dz'}[(1+\mathrm{z'})^3\Omega_\mathrm{m} + \Omega_\Lambda + (1+\mathrm{z'})^2\Omega_\kappa]^{-1/2}\}. \tag{19}$$

The curvature parameter is given by the constraint equation

$$\Omega_\kappa = 1 - \Omega_m - \Omega_\Lambda \tag{20}$$

and the function

$$\mathrm{sinn}(x) = \begin{cases} x & \text{for a flat Universe } (\Omega_\kappa = 0); \\ \sin(x) & \text{for a closed Universe } (\Omega_\kappa < 0); \\ \sinh(x) & \text{for an open Universe } (\Omega_\kappa > 0). \end{cases} \tag{21}$$

We now assume that from each SNIa in our sample we get a measurement of the distance modulus with Gaussian noise[2], i.e., that the likelihood function for each SN i ($i = 1,\dots,N$) is of the form

$$\mathscr{L}_i(z_i,\mathscr{C},M) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{1}{2}\frac{(\hat{\mu}_i - \mu(z_i,\mathscr{C}))^2}{\sigma_i^2}\right). \tag{22}$$

The observed distance modulus is given by $\hat{\mu}_i = \hat{m}_i - M$, where $\hat{m}_i$ is the observed apparent magnitude and M is the intrinsic magnitude of the SNIa. We assume that each SN observation is independent of all the others.

The provided data file[3] (`SNe_simulated.dat`) contains simulated observations from the above simplified model of $N = 300$ SNIa. The two columns give the redsfhit z_i and the observed apparent magnitude $\hat{m}_i$. The observational error is the same for all SNe, $\sigma_i = \sigma = 0.4$ mag for $i = 1,\dots,N$.

A plot of the data set is shown in the left panel of Fig. 2. The characteristics of the simulated SNe are designed to mimic currently available datasets[4].

1. We assume that the intrinsic magnitude[5] is known and fix $M = M_0 = -19.3$ and that $h = 0.72$. We also assume that the observational error is known, given by the value above.

   Using a language of your choice, write a code to carry out an MCMC sampling of the posterior probability for $(\Omega_m,\Omega_\Lambda)$ and plot the resulting 68% and 95% posterior regions, both in 2D and marginalized to 1D, using uniform priors on $(\Omega_m,\Omega_\Lambda)$ (be careful to define them explicitly).

   You should obtain a result similar to the 2D plot shown in the right panel of Fig. 2.

2. † Add the quantity σ (the observational error) to the set of unknown parameters and estimate it from the data along with $\mathscr{C}$. Notice that since σ is a "scale parameter", the appropriate (improper) prior is $p(\sigma) \propto 1/\sigma$.

[2] We neglect the important issue of applying the empirical corrections known as Phillip's relations to the observed light curve. This is of fundamental important in order to reduce the scatter of SNIa within useful limits for cosmological distance measurements, but it would introduce a technical complication here without adding to the fundamental scope of this exercice.

[3] Thanks to Marisa March for help with the simulation.

[4] See Kowalski et al, *Astrophys. J.*, 686:749-778, 2008 (arXiv:0804.4142) and Amanullah et al, 2010 (arXiv:1004.1711). More recently, Rest et al (2013), arXiv: 1310.3828.

[5] In reality the SNe intrinsic magnitude is not fixed, but there is an "intrinsic dispersion" (even after Phillips' corrections) reflecting perhaps intrinsic variability in the explosion mechanism, or environmental parameters which are currently poorly understood.

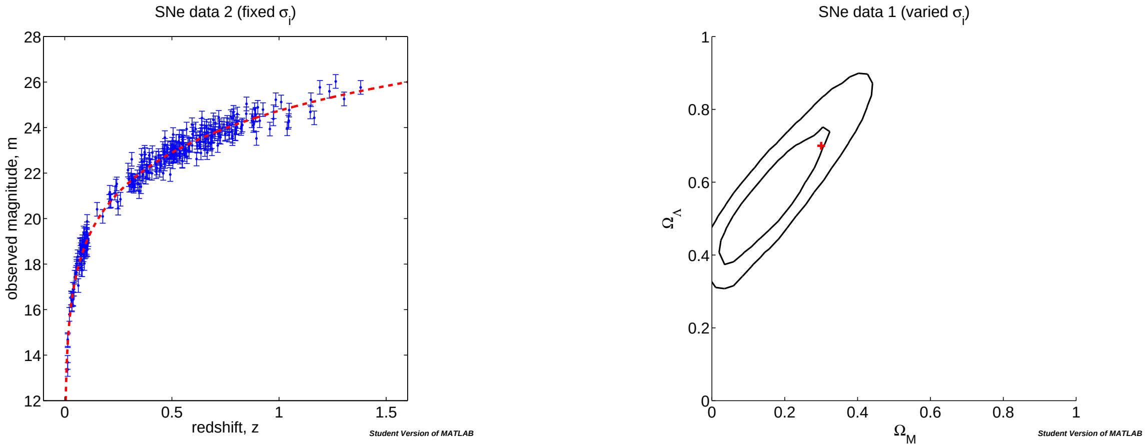

Figure 2: Left: Simulated SNIa dataset, `SNe_simulated.dat`. The solid line is the true underlying cosmology. Right: constraints on Ω_m, Ω_Λ from this dataset, with contours delimiting 2D joint 68% and 95% credible regions (uniform priors on the variables Ω_m, Ω_Λ, assuming $M = M_0$ fixed and $h = 0.72$). The red cross denotes the true value.

3. The location of the peaks in the CMB power spectrum gives a precise measurement of the angular diameter distance to the last scattering surface, divided by the sound horizon at decoupling. This approximately translates into an effective constraint[6] on the following degenerate combination of Ω_m and Ω_Λ:
$$1.41\Omega_\Lambda + \Omega_m = 1.30 \pm 0.04. \tag{23}$$
Add this constraint (assuming a Gaussian likelihood, with the above mean and standard deviation) to the SNIa likelihood and plot the ensuing combined 2D and 1D limits on $(\Omega_m, \Omega_\Lambda)$.

4. The measurement of the baryonic acoustic oscillation scale in the galaxy power spectrum at small redshift gives an effective constraint on the angular diameter distance D_A out to $z \sim 0.3$. This measurement can be summarized[7] (simplifying somewhat) by the constraint:
$$D_A(z = 0.3) = (893 \pm 27)\ \text{Mpc}. \tag{24}$$
Add this constraints (again assuming a Gaussian likelihood) to the above CMB+SNIa limits and plot the resulting combined 2D and 1D limits on $(\Omega_m, \Omega_\Lambda)$.
Hint: recall that $D_L(z) = (1+z)^2 D_A(z)$.

3 Model comparison

3.1 Introductory level

1. **Coin bias.** A coin is tossed $N = 250$ times and it returns $H = 140$ heads. Evaluate the evidence that the coin is biased using Bayesian model comparison and contrast your findings with the usual (frequentist) hypothesis testing procedure (*i.e.*, testing the null hypothesis that $p_H = 0.5$). Discuss the dependency on the choice of priors.

[6] For full details, see Spergel et al, *Astrophys. J. Suppl.*, 170:377, 2007 (astro-ph/0603449), Fig. 20.

[7] For details, see Percival et al, *Mon. Not. Roy. Astron. Soc.*, 401:2148-2168, 2010 (arXiv:0907.1660).

2. **Light deflection and GR.** In 1919 two expeditions sailed from Britain to measure the light deflection from stars behind the Sun's rim during the solar eclipse of May 29th. Einstein's General Relativity predicts a deflection angle
$$\alpha = \frac{4GM}{c^2 R},$$
where G is Newton's constant, c is the speed of light, M is the mass of the gravitational lens and R is the impact parameter. It is well known that this result it exaclty twice the value obtained using Newtonian gravity. For $M = M_\odot$ and $R = R_\odot$ one gets from Einstein's theory that $\alpha = 1.74$ arc seconds.

   The team led by Eddington reported 1.61 ± 0.40 arc seconds (based on the position of 5 stars), while the team headed by Crommelin reported 1.98 ± 0.16 arc seconds (based on 7 stars).

   What is the Bayes factor between Einstein and Newton gravity from those data? Comment on the strength of evidence.

3. **Evidence for a cosmological constant.** Assume that the combined constraints from CMB, BAO and SNIa on the density parameter for the cosmological constant can be expressed as a Gaussian posterior distribution on Ω_Λ with mean 0.7 and standard deviation 0.05. Use the Savage-Dickey density ratio to estimate the Bayes factor between a model with $\Omega_\Lambda = 0$ (i.e., no cosmological constant) and the ΛCDM model, with a flat prior on Ω_Λ in the range $0 \leq \Omega_\Lambda \leq 2$. Comment on the strength of evidence in favour of ΛCDM.

3.2 Intermediate level

The anthropic principle. If the cosmological constant is a manifestation of quantum fluctuations of the vacuum, QFT arguments lead to the result that the vacuum energy density ρ_Λ scales as

$$\rho_\Lambda \sim \frac{c\hbar}{16\pi} k_{\max}^4 \tag{25}$$

where $k_{\max}$ is a cutoff scale for the maximum wavenumber contributing to the energy density[8]. Adopting the Planck mass as a plausible cutoff scale (i.e., $k_{\max} = c/\hbar M_{\rm Pl}$) leads to "the cosmological constant problem", i.e., the fact that the predicted energy density

$$\rho_\Lambda \sim 10^{76}\ \mathrm{GeV}^4 \tag{26}$$

is about 120 orders of magnitude larger than the observed value, $\rho_{\rm obs} \sim 10^{-48}$ GeV^4.

1. Repeat the above estimation of the evidence in favour of a non-zero cosmological constant, adopting this time a flat prior in the range $0 \leq \Omega_\Lambda/\Omega_\Lambda^{\rm obs} < 10^{120}$. What is the meaning of this result? What is the required observational accuracy (as measured by the posterior standard deviation) required to override the Occam's razor penalty in this case?

2. It seems that it would be very difficult to create structure in a universe with $\Omega_\Lambda \gg 100$, and so life (at least life like our own) would be unlikely to evolve. How can you translate this "anthropic" argument into a quantitative statement, and how would it affect our estimate of Ω_Λ and the model selection problem?

[8] See e.g. Carroll & Press, *Ann. Rev. Astron. Astrophys.* 30:499-542, 1992.

3.3 Advanced level

Flat Universe from SNIa and other data. This problem follows up the cosmological parameter estimation problem from supernovae type Ia[9] in section 2.3.

1. Adopt uniform priors $\Omega_\mathrm{m} \sim U(0,2)$ and $\Omega_\Lambda \sim U(0,2)$. Produce a 2D marginalised posterior pdf in the $(\Omega_\mathrm{m}, \Omega_\Lambda)$ plane.
2. Produce a 1D marginalised posterior pdf for the curvature parameter, $\Omega_\kappa = 1 - \Omega_\Lambda - \Omega_\mathrm{m}$, paying attention to normalising it to unity probability content.

   What is the shape of the prior on Ω_κ implied by your choice of a uniform prior on $\Omega_\mathrm{m}, \Omega_\Lambda$?
3. Use the Savage-Dickey density ratio formula to estimate from the above 1D posterior the evidence in favour of a flat Universe, $\Omega_\kappa = 0$, compared with a non-flat Universe, $\Omega_\kappa \neq 0$, with prior $P(\Omega_\kappa) = U(-1,1)$.

   Discuss the dependency of your result on the choice of the above prior range.

[9] For a more thorough treatment, see M. Vardanyan, R. Trotta and J. Silk (2009) Mon. Not. R. Astron. Soc. 397 , 431-444 (2009) and M. Vardanyan, R. Trotta and J. Silk (2011) MNRASLett 413, 1, 2011, L91ÐL95.