Astrostatistics problems

Roberto Trotta, Imperial College London

November 14, 2011

Some problems have been contributed by Daniel Mortlock and Andrew Jaffe, to whom go my thanks.

1 Probabilistic reasoning

1. A batch of chemistry undergraduates are screened for a dangerous medical condition called *Bacillum Bayesianum* (BB). The incidence of the condition in the population is estimated at about 1%. If the person has BB, the test returns positive 90% 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. Should your friend be worried?

2. A pan contains 10 ravioli, of which 9 are filled with pesto and one with ricotta. You put in the pan a further raviolo filled with pesto and cover with an opaque lid.hen you randomly draw a raviolo, eat it and discover that it is filled with pesto. After this procedure, the pan is again in the same state as before. What is now the probability that the next raviolo drawn will be filled with pesto?

   On a different night, you cook a pan of mixed pesto and ricotta ravioli (in equal proportions). One last raviolo remains in your plate, which could be either pesto or ricotta. Your friend tosses into your plate her last raviolo, which she tells you is a pesto-filled one. Then you mix the two ravioli randomly, pick one and realize it's pesto. What is now the probability that the last raviolo in your plate is pesto?

3. In a TV debate, a politician named Barack affirms that climate change is caused by human activities. You trust Barack to tell the truth with probability 2/3. Another politician, Silvio, then agrees that what Barack has said is indeed true. Your trust in Silvio is much weaker, and you estimate that he lies with probability 3/4. After you have heard Silvio, what is your probability that climate change is indeed caused by humans? How can this help explain the polarization of political positions, *e.g.*, in the US?

   (You may assume that you have no other information or prior opinion on the origin of climate change other than what you heard from the two politicians)

4. You are playing poker and the dealer gives himself a Three of a Kind, for which the odds are about 46:1. What is the posterior probability that the dealer is cheating, given that he is: (a) St Augustin (b) your older brother (c) Al Capone?

5. A body has been found on the Baltimore West Side, with no apparent wounds, although it transpires that the deceased, a Mr Fuzzy Dunlop, was a heavy drug user. The detective in charge suggests to close the case and to attribute the death to drugs overdose, rather than murder.

   Knowing that, of all murders in Baltimore, about 30% of the victims were drug addicts, and that the probability of a dead person having died of overdose is 50% (without further evidence apart from the body) estimate the probability that the detective's hunch is correct.

2 Single parameter inference

1. A coin is tossed N times and heads come up H times. Determine the bias of the coin, i.e., the probability p_H that a single flip will give heads and its uncertainty. What is the probability that the $(N+1)$-th flip will give heads? Plot your results as a function of H for $N = 10, 100, 1000$. Discuss your choice of priors.

2. An astronomer wishes to know the (mono-chromatic) flux of a particular source and makes a photometric measurement which registers $N_{\rm src}$ photons. Assume that all the photons have come from the source itself (*i.e.*, there is no background or or source confusion) and that the known calibration constant, C, is such that a source of true flux $F_{\rm src}$ would, on average, yield $F_{\rm src}/C$ photons in such a measurement (*i.e.*, a generic estimate of the source's flux would be $\hat{F}_{\rm src} \simeq C N_{\rm src}$).

   (a) What is the model parameter that the astronomer is trying to infer?

   (b) What is/are the datum/data?

   (c) What is the likelihood [*i.e.*, the probability $\Pr(N_{\rm src}|F_{\rm src})$]?

   (d) What prior information might the astronomer have *before* making (or at least making use of) the measurement?

   (e) If the astronomer had access to a catalogue of sources of similar fluxes from a different part of the sky, how might this catalogue be used to generate an appropriate, if approximate, prior distribution for the source's true flux, $F_{\rm src}$?

   (f) If the distribution of source fluxes was known to increase as $\Pr(F_{\rm src}) \propto F_{\rm src}^{-5/2}$, what would the resultant posterior information on the source's flux be upon combining this knowledge about the source population and the data on the particular source of interest? Is this prior normaliseable (*i.e.*, proper)?

   (g) Assuming, for simplicity, that $C = 1$, plot both the likelihood, $\Pr(N_{\rm src}|F_{\rm src})$, and the posterior distribution, $\Pr(F_{\rm src}|N_{\rm src})$, as a function of $F_{\rm src}$ in i) the case that $N_{\rm src} = 5$ (plausible for an X-ray observation) and ii) the case that $N_{\rm src} = 10^4$ (plausible for an optical observation). Are any of these functions approximately Gaussian? What is the probability that the source has $F_{\rm src} = 0$? What is the probability that the source has $F_{\rm src} < 0$? How did utilising the photometric measurement of the source affect these probabilities?

   (h) What would be a reasonable "best estimate" of the source's flux? (There are several plausible answers.) How do these best estimates relate to the naive estimate $\hat{F}_{\rm src} = C N_{\rm src}$? Does this make sense?

3. This problem takes you through the steps to derive the posterior distribution for a quantity of interest θ, in the case of a Gaussian prior and Gaussian likelihood.

   Let us assume that we have made N independent measurements, $\hat{x} = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_N\}$ of a quantity of interest θ (this could be the temperature of an object, the distance of a galaxy, the mass of a planet, etc). We assume that each of the measurements in independently Gaussian distributed with known experimental standard deviation σ. Let us denote the sample mean by $\bar{x}$, i.e.

   $$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_i. \tag{1}$$

   Before we do the experiment, our state of knowledge about the quantity of interest θ is described by a Gaussian distribution on θ (i.e., the prior in Bayes theorem), centered around 0 (we can always choose the units in such a way that this is the case). Such a prior might come e.g. from a previous experiment we have performed. The new experiment is however much more precise, i.e. $\Sigma \gg \sigma$. Our prior state of knowledge be written in mathematical form as the following Gaussian pdf:

$$p(\theta) \sim \mathcal{N}(0, \Sigma^2). \tag{2}$$

(a) Write down the likelihood function for the measurements and show that it can be recast in the form:

$$\mathcal{L}(\theta) = L_0 \exp\left(-\frac{1}{2} \frac{(\theta - \bar{x})^2}{\sigma^2/N} \right), \tag{3}$$

where L_0 is a constant that does not depend on θ.

(b) By using Bayes theorem, compute the posterior probability for θ after the data have been taken into account, i.e. compute $p(\theta|\hat{x})$. Show that it is given by a Gaussian of mean $\bar{x}\frac{\Sigma^2}{\Sigma^2+\sigma^2/N}$ and variance $\left[\frac{1}{\Sigma^2} + \frac{N}{\sigma^2}\right]^{-1}$.
Hint: you may drop the normalization constant from Bayes theorem, as it does not depend on θ

(c) Show that as $N \to \infty$ the posterior distribution becomes independent of the prior.

(d) Show that as $N \to \infty$ the mean of the posterior distribution converges to the MLE of the mean for θ. This means that for a large number of measurements, the Bayesian result matches the frequentist MLE result.

3 Model selection

1. 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.

2. 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. The astronomer from Section 2, Question 2 makes another photometric observation of a different source, again measuring $N_{\rm src}$ photon counts. (And, again, all the photons are assumed to have come from the source and the calibration constant, C, to be known perfectly.) This time, however, the source is known to be either of Type A, in which case its flux would be $F_{\rm A}$, or Type B, in which case its flux would be $F_{\rm B}$. Sources of Type A are ten times more common than sources of Type B. The astronomer's main aim is to classify the source as Type A or Type B.

   (For those who prefer reasoning in a less abstract context, you could think of Type A objects as, say, high-redshift quasars which are very faint at the wavelength of the observation, and Type B objects as stars, which are relatively much brighter. The situation described above might arise if a source was detected in an observation at longer wavelengths and the astronomer was trying to determine whether the source might be an "interesting" quasar or a "mundane" star. A full spectrum would give the answer, but it would be much cheaper observationally to determine the source's nature using just a photometric measurement if possible.)

   (a) This problem can still be treated using the parameter estimation formalism described above. The data and likelihood are the same as before, but now the prior is different; write down this new flux prior [*i.e.*, $\Pr(F_{\rm src})$]. Having done so, then write down the posterior for the source's true flux, $\Pr(F_{\rm src}|N_{\rm src})$. How can this statement about the astronomer's knowledge of the source's flux be simply converted to a probabilistic classification that the source is of Type A or Type B?

   (b) This problem could also be thought of in terms of model comparison, with two models A and B. Do either of these models have any parameters? What are the evidences for Type A and Type B? Including the model priors, what is the final probabilistic classification? Does this differ from the result obtained above using parameter estimation techniques?

   (c) Before heading off to the telescope, the astronomer decides to check if the source has been observed previously and finds an "upper limit", $F_{\rm lim}$, in the literature. For simplicity this is taken to imply that the photon counts of the source in that observation were less than $F_{\rm lim}/C$. What is the likelihood, $\Pr(N_{\rm src} < F_{\rm lim}/C|F_{\rm src})$? What is the resultant posterior classification if $F_{\rm lim}/C = 10.0$? If $F_{\rm A} = 0.1F_{\rm lim}$ and $F_{\rm B} = 1.2F_{\rm lim}$ (*i.e.*, the observation is of an "interesting" depth that the source might be detected if it was of Type A but almost certainly wouldn't have been if it was of Type B), what is the resultant probabilistic classification?

(d) The astronomer then sends off a polite e-mail to the author of the “upper limit” paper requesting the raw data in which the source was not detected. After reanalysing the data the astronomer is now able to replace the constraint $N_{\rm src} < F_{\rm lim}/C$ with an actual measurement $N_{\rm src} = 0.9F_{\rm lim}/C$. Assuming $F_{\rm A} = 0.1F_{\rm lim}$ and $F_{\rm B} = 1.2F_{\rm lim}$ again, and $N_{\rm src} = 9$ (consistent with the figures for the above limit), what are the probabilistic inferences now? How has the inclusion of the full data changed this inference relative to the use of a limit only (and relative to the situation when only the prior information was available)?

4. 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 $0 \le \Omega_\Lambda \le 2$. Comment on the strength of evidence in favour of ΛCDM.

5. 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

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

where $k_{\rm max}$ is a cutoff scale for the maximum wavenumber contributing to the energy density[1]. Adopting the Planck mass as a plausible cutoff scale (i.e., $k_{\rm max} = c/\hbar M_{\rm Pl}$) leads to “the cosmological constant problem”, i.e., the fact that the predicted energy density

$$\rho_\Lambda \sim 10^{76}\ {\rm GeV}^4 \tag{5}$$

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

Repeat the above estimation of the evidence in favour of a non-zero cosmological constant, adopting this time a flat prior in the range $0 \le \Omega_\Lambda/\Omega_\Lambda^{\rm obs} < 10^{120}$. 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?

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 effect our estimate of Ω_Λ and the model selection problem?

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

4 Multiple parameter inference

1. The astronomer, having become disillusioned with the lazy data-reporting practices in optical astronomy, has moved into X-ray astronomy. Having found a source of interest, the astronomer falls back on old habits and can't resist trying to do some photometry, just for old time's sake. This is something of a shock, however, both because the expected number of photons from the source is very small (*i.e.*, single figures) and also because there is now an appreciable background (*i.e.*, maybe a third of the photons registered might not have been emitted from the target source). The basic task is the same as in Section 2 – to infer the flux of the source – but now there is the additional complication of a background which must be included in the model. Just as a source of (true) flux $F_{\rm src}$ would provide an average of $\bar{N}_{\rm src} = F_{\rm src}/C$ photons in this measurement, the background flux (in the measurement aperture), $F_{\rm bkg}$, would be expected to contribute $\bar{N}_{\rm bkg} = F_{\rm bkg}/C$ photons in such a measurement.

   (a) It is quite possible that the background rate is known precisely (or with so much more accuracy than the measurement that it is effectively exact), so that $F_{\rm bkg}$ can be treated as a known constant. Given a single on-source measurement of $N_{\rm on}$ photons, what is the likelihood and the posterior for the source flux [again assuming the prior $\Pr(F_{\rm src}) \propto F_{\rm src}^{-5/2}$]?

   (b) Unfortunately, the uncertainty in the background is often significant; in such cases it must also be measured and its level inferred. The astronomer now makes two measurements: one with the telescope aperture centred on the source, which yields $N_{\rm on}$ photons, and one with the telescope aperture pointed at a "blank" patch of sky, which yields $N_{\rm off}$ photons. Although the astronomer is only really interested in $F_{\rm src}$, it is also necessary to include the unknown $F_{\rm bkg}$ in the modelling (to be marginalized over later).

   Write down the likelihood [*i.e.*, $\Pr(N_{\rm on}, N_{\rm off}|F_{\rm src}, F_{\rm bkg})$] and the prior [*i.e.*, $\Pr(F_{\rm src}, F_{\rm bkg})$]. (Think carefully about the prior for the background flux. If you have a strongly motivated choice of prior make sure it is justified; if you are less certain try working through the problem with different plausible priors that you think might span the possibilities.)

   (c) The full posterior $\Pr(F_{\rm src}, F_{\rm bkg}|N_{\rm on}, N_{\rm off})$ is fairly complicated (whatever prior for the background level was chosen). To explore this distribution without the need for any significant additional programming (or algebgra), generate 10^5 samples from the full posterior using MCMC in the case that $N_{\rm on} = 9$ and $N_{\rm off} = 3$ (and, again for convenience, that $C = 1$). Make a scatter plot showing the range of plausible $F_{\rm src}$ and $F_{\rm bkg}$ values. Are they independent or correlated? Can you explain this intuitively? Are these two parameters linked physically at all (*i.e.*, does the flux of a particular source have anything to do with the background)?

   (d) It is only the marginal posterior of the source flux, $\Pr(F_{\rm src}|N_{\rm on}, N_{\rm off})$, that is really of interest. To obtain this marginalized distribution, post-process the MCMC output by making a histogram of the $F_{\rm src}$ values, ignoring the $F_{\rm bkg}$ values.

2. 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

$$\mathscr{C} = \{\Omega_m, \Omega_\Lambda, h\} \qquad (6)$$

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. Also, we will fix $h = 0.72$, where the Hubble constant today is given by $H_0 = 100h$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

$$\mu(z, \mathscr{C}) = 5 \log \left[\frac{D_L(z, \Omega_m, \Omega_\Lambda, h)}{\mathrm{Mpc}} \right] + 25, \tag{7}$$

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

$$\mu(z, \mathscr{C}) = \eta + 5 \log d_L(z, \Omega_m, \Omega_\Lambda), \tag{8}$$

where

$$\eta = -5 \log \frac{100h}{c} + 25 \tag{9}$$

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

$$d_L(z, \Omega_m, \Omega_\Lambda) = \frac{(1+z)}{\sqrt{|\Omega_\kappa|}} \mathrm{sinn} \{ \sqrt{|\Omega_\kappa|} \int_0^z \mathrm{dz'} [(1+\mathrm{z'})^3 \Omega_\mathrm{m} + \Omega_\Lambda + (1+\mathrm{z'})^2 \Omega_\kappa]^{-1/2} \}. \tag{10}$$

The curvature parameter is given by the constraint equation

$$\Omega_\kappa = 1 - \Omega_m - \Omega_\Lambda \tag{11}$$

and the function $\mathrm{sinn}(x) = x, \sin(x), \sinh(x)$ for a flat Universe ($\Omega_\kappa = 0$), a closed Universe ($\Omega_\kappa < 0$) or an open Universe, respectively.

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 ($i = 1, \ldots, N$) is of the form

$$\mathcal{L}_i(z_i, \mathscr{C}, M) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{1}{2} \frac{(\hat{\mu}_i - \mu(z_i, \mathscr{C}))^2}{\sigma_i^2} \right). \tag{12}$$

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 $i = 1, \ldots, N$. A plot of the data set is shown in the left panel of Fig. 1. The characteristics of the simulated SNe are designed to mimic currently available datasets[4].

(a) We begin by assuming that the intrinsic magnitude is known and fix $M = M_0 = -19.3$. 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. Be careful in defining and discussing your priors on $(\Omega_m, \Omega_\Lambda)$ explicitly. You should get something similar to the 2D plot shown in the right panel of Fig. 1.

[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). For a fully Bayesian treatment, see March et al (arXiv: 1102.3237).

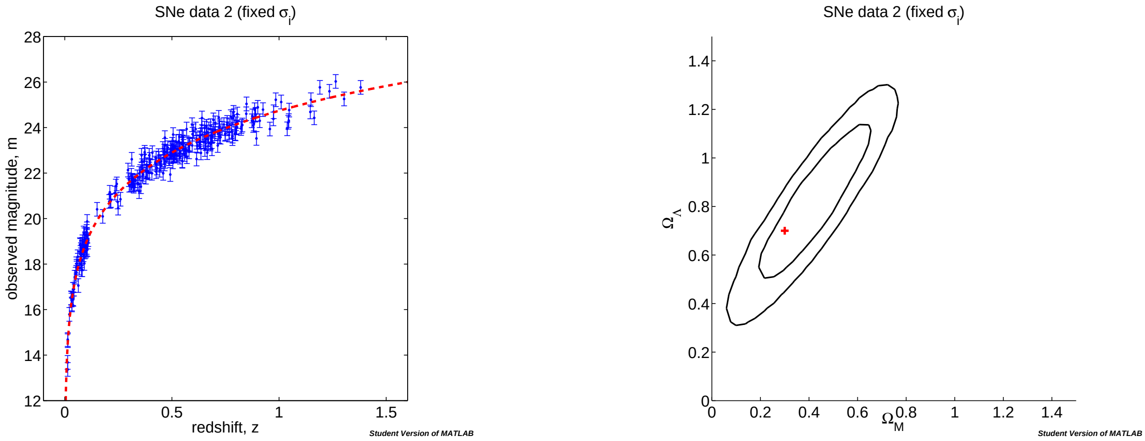

Figure 1: Left: Simulated SNIa dataset, `SNe_simulated.dat`. The solid line is the true underlying cosmology. Right: constraints on Ω_m, Ω_Λ from this dataset, with contours delimiting 2D joint 68% and 95% regions (flat priors on the variables Ω_m, Ω_Λ, assuming $M = M_0$ fixed and $h = 0.72$). The red cross denotes the true value.

(b) Add the quantity σ 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 $p(\sigma) \propto 1/\sigma$.

(c) 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. Suppose that the intrinsic magnitude of each SNIa, M_i, is probabilistically drawn from a Gaussian distribution with mean M_0 and standard deviation Δ, i.e.

$$M_i \sim \mathcal{N}(M_0, \Delta^2), \quad i = 1, \ldots, N. \tag{13}$$

Derive the effective likelihood accounting for intrinsic dispersion by introducing a vector $\mathbf{M} = \{M_1, M_2, \ldots, M_N\}$ of unobserved intrinsic magnitudes and marginalizing over it analytically (assume for this that the observational error is known, as in part (a)).

Hints: recall that $\mathcal{N}(\mathbf{m}, \Sigma)$ denotes a Normal distribution of mean $\mathbf{m}$ and covariance matrix Σ, i.e.

$$p(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m})^t \Sigma^{-1} (\mathbf{x}-\mathbf{m})\right] \tag{14}$$

It is useful to rewrite the joint likelihood in matrix form and carry out all of the Gaussian integrals at once by using this standard result:

$$\int \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m})^t \Sigma^{-1} (\mathbf{x}-\mathbf{m})\right] d\mathbf{x} = \sqrt{\det(2\pi\Sigma)} \tag{15}$$

(d) 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[5] on the following degenerate combination of Ω_m and Ω_Λ:

$$1.41\Omega_\Lambda + \Omega_m = 1.30 \pm 0.04. \tag{16}$$

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

Add this constraint (assuming a Gaussian distribution) to the SNIa likelihood and plot the ensuing combined 2D and 1D limits on $(\Omega_m, \Omega_\Lambda)$.

(e) 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 $z \sim 0.3$ (see lectures by Daniel Eisenstein and Bruce Bassett). This measurement can be summarized[6] (simplifying somewhat) by the constraint:

$$D_A(z = 0.3) = (893 \pm 27)\ \text{Mpc}. \tag{17}$$

Add this constraints (again assuming it is Gaussian distributed) to the above CMB+SNIa limits and plot the resulting combined 2D and 1D limits on $(\Omega_m, \Omega_\Lambda)$.
Hint: recall that $D_L(z) = (1+z)^2 D_A(z)$.

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