ICIC Data Analysis Workshop

Day 1 problems

11 September 2013

1 Probabilistic reasoning and parameter estimation

1. Imagine you have made it to the Wimbledon singles final and the umpire is about to flip a (specially minted) coin to see who will serve first.

   What probability do you assign that the coin will, when tossed, land heads? (Try and justify your answer.)

   The coin is tossed and the umpire catches it, immediately covering it with his other hand. What is your probability the coin has landed heads? What is the umpire's probability the coin has landed heads?

   The umpire lifts his hand a little to check the coin, which has landed heads, but you can't see it. What is your probability the coin has landed heads? What is the umpire's probability the coin has landed heads?

   You are now shown the coin, and also see that it has landed heads. What is your probability the coin has landed heads? What is the umpire's probability the coin has landed heads?

   Now think through the above sequence in full. At what point did the answers change? Was it related to physical events? Was it related to the information other people had? Was it related to the information you had?

   Now imagine that someone had, before the toss, passed you a note saying that the coin is biased, and will land one way up 99% of the time (but, unhelpfully, that the note doesn't say which way the coin is biased). Answer the above four questions again. How has this new piece of inside information changed your answers? What if you'd noticed the umpire tossing the coin beforehand while you were signing autographs and you'd noticed it had landed heads that time?

2. Solve the 'Monty Hall' problem given in the lectures, using Bayes' theorem. To recall: you have three doors in front of you, two of which have bad prizes and one with a good prize. You select one door, and the host opens another and shows that it has one of the bad prizes. Should you switch?

3. 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. Then

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, *before your friend threw hers in*, is pesto? What is the probability that the *very last* raviolo is pesto?

Compare this to the Monty Hall problem.

4. 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. You may assume that in crime-ridden Baltimore all deaths (at least those investigated by the detective) are either by overdose or murder. Do you have to make any other assumptions?

5. The distribution of flux densities of extragalactic radio sources is a power-law with slope $-\alpha$, say, so the likelihood to measure a source flux S is $p(S) \propto S^{-\alpha}$, above some (known) instrumental limiting flux density of S_0. In a non-evolving Euclidean universe $\alpha = 3/2$ and departure of α from the value 3/2 is evidence for cosmological evolution of radio sources (we assume measurement errors are negligible). This was the most telling argument against the steady-state cosmology in the early 1960's (even though they got the value of a wrong by quite a long way).

   - Given observations of radio sources with flux densities S, what is the most probable value of α, assuming a uniform prior? (Hint: in this case you will have to normalise $p(S)$).
   - Show that if a single source is observed, and the flux is $2S_0$, that the most probable value of α is 2.44.
   - By examining the second derivative of the posterior, estimate the error on α to be 1.44.

2 Optional problems

1. 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 $\hat{F}_{\rm src} \simeq CN_{\rm src}$).

(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 $\hat{F}_{\rm src} = CN_{\rm src}$? Does this make sense?

2. 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 above – 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 rest of the problem requires techniques to be discussed on day 2 of the workshop.

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.