Advanced methods in statistical data analysis

Examples for Friday

Part A: Model selection

1. You watch me toss a coin a number of times. Being a trusting sort of person, you start out thinking that there is only a one-in-a-hundred chance that my coin is two-headed rather than a normal fair coin. How many consecutive 'heads' have to come up before you think it is more likely that my coin is two-headed (assuming you are a dedicated Bayesian)?

2. On the course WWW page on indico for this session, find two files each containing some data points (x_i, y_i) with uncertainties σ_i on the y_i values. Carry out the following procedure for each dataset.

Assume that the likelihood of a model can be expressed as

$$\mathcal{L} = \exp\left(-\frac{1}{2}\sum_i \frac{(y_{\rm model} - y_i)^2}{\sigma_i^2}\right)$$

Consider as possible models a straight-line fit $y = a + bx$ and a quadratic fit $y = a + bx + cx^2$. Use the Metropolis–Hastings sampler you wrote yesterday (or another method of your choice[1]) to find the maximum likelihood of each model. Use the Bayesian Information Criterion

$$\mathrm{BIC} = -2\ln\mathcal{L}_{\rm max} + k\ln N$$

(k is number of parameters, N is number of datapoints) to decide which of these two models is the preferred fit to the data.

[If you have spare time in this session, you could consider analyzing other polynomial models.]

3. Dark energy models are often classified according to the dark energy equation of state w. In this question, we will consider two models, the cosmological constant where w is fixed at -1 and a model where w is an unknown constant drawn from a uniform probability distribution $-1 \leq w \leq -1/3$.

According to arXiv:0803.0547, when WMAP5 data is combined with other data the likelihood for w is well approximated by a gaussian of mean -1.06 and standard deviation 0.13 (i.e. $w = -1.06 \pm 0.13$ at 68% confidence). Compute the Bayesian evidence for each model and determine their relative probabilities, under the assumption that we thought the models equally likely before we saw the data.

[Note: a properly normalized gaussian distribution of mean x_0 and standard deviation σ is $p(x) = (\sqrt{2\pi}\,\sigma)^{-1}\exp\left(-(x - x_0)^2/\sigma^2\right)$.]

[1] I assume that ROOT can do this, though I am not a ROOT user. However it may be more instructive to use your own code.

Part B: Forecasting and optimization

4. This question follows on from Q3 above. A future satellite mission is proposed which is advertized as being able to improve the uncertainty on w by a factor of a hundred. Estimate as simply as possible the probability that the satellite will rule out the cosmological constant model (under the assumption that these two models are the complete set of relevant models).