ICIC Astrostatistics Summer School

Posterior Sampling Project

The discovery of the Higgs boson was a fantastic example of both experimental science (in its grandest form, the Large Hadron Collider, the LHC) and theoretical physics (as the subsequent Nobel Prize went to the theoreticians who predicted the existence of such a particle). However, this discovery was also fundamentally statistical in nature, as the evidence for the Higgs boson took the form of a small bump in an otherwise smooth but noisy spectrum of measured collision energies.

This project is to implement a simplified version of the LHC Higgs analysis, applying Bayesian methods to a simulated data-set. This is given in the file `spectrum_lhc.dat`, and consists of the number of recorded collisions, $N_{1:B}$, in each of $B = 60$ bins with central energies $E_{1:B}$ (given in GeV). The bins all have an energy width of $\Delta E = 1$ GeV and the data corresponds to LHC runs of total duration $T = 3$ yr.

The overall energy-dependent rate of collisions is given by the sum of the two contributions as

$$R(E) = R_{\rm bkg}(E) + R_{\rm H}(E),$$

where $R_{\rm bkg}(E)$ is the smooth background and $R_{\rm H}(E)$ is the (possible) Higgs contribution. These rate $R(E)$ is defined such that the average/expected number of collisions in an infinitessimal time period $\mathrm{d}T$ and an infinitessimal energy range $\mathrm{d}E$ is $R(E)\,\mathrm{d}T\,\mathrm{d}E$.

The background spectrum can be taken to be of the form

$$R_{\rm bkg}(E) = \frac{\Gamma_{\rm bkg}}{E_0}\,\left(\frac{E}{E_0}\right)^{\alpha},$$

where $\Gamma_{\rm bkg}$ is the overall background rate (with units of inverse time), $E_0 = 100$ GeV is an (arbitrary) reference energy, and $\alpha < -1$ is the logarithmic slope of the background spectrum.

If the Higgs boson exists, a proposition denoted as H, it would producce a contribution of the form

$$R_{\rm H}(E) = \Gamma_{\rm H}\,\mathrm{N}(E; E_{\rm H}, \sigma_{\rm H}^2) = \frac{\Gamma_{\rm H}}{(2\pi)^{1/2}\,\sigma_{\rm H}}\,\exp\left[-\frac{1}{2}\,\left(\frac{E - E_{\rm H}}{\sigma_{\rm H}}\right)^2\right],$$

where $E_{\rm H} = m_{\rm H}\,c^2$ is the rest-mass energy of the Higgs boson, $\Gamma_{\rm H}$ is the associated rate of collisions, and $\sigma_{\rm H}$ characterises the spread in energy of the Higgs collision products.

1. Read in and make a plot of the LHC data, *i.e.*, number of collisions vs. energy.

   Add a plausible smooth background-only power-law spectrum to the plot, and hence make an informal assessment of i) whether there is evidence for a Higgs boson in the data and ii) what energy range is plausible for the Higgs boson (under the assumption that the Higgs is present).

2. Show that it is reasonable to assume the probability of recording N_b collisions in bin b, with central energy E_b, is

   $$\mathbb{P}(N_b|\lambda_b) = \Theta(N_b)\,\frac{{\lambda_b}^{N_b}\,e^{-\lambda_b}}{N_b!},$$

where $\lambda_b = R(E_b)\,T\,\Delta E$ and $\Theta(\cdot)$ is the Heaviside step function.

Also give an argument why it is reasonable to assume that the bins can be treated independently, and hence use this to write down an expression for the log-likelihood, $\log[\mathbb{P}(N_{1:B}|E_{\rm H}, \Gamma_{\rm H}, \sigma, \Gamma_{\rm bkg}, \hat{\alpha})]$. (This will be most easily done by again using λ_b as a shorthand.)

3. Fit a background-only model to the data by calculating the posterior $\mathbb{P}(\Gamma_{\rm bkg}, \alpha|N_{1:B}, \neg\mathsf{H}, \mathsf{I})$. State explicitly whatever prior information, I, is being assumed.

   Plot the joint posterior distribution in $\Gamma_{\rm bkg}$ and α, showing 30%, 60%, 90% highest posterior density credible regions.

   Comment on whether the resultant best-fit model can explain the data-set.

4. Adopting the fixed background model that $\hat{\Gamma}_{\rm bkg} = 2000$ yr^{-1} and $\hat{\alpha} = -4$, use the Metropolis algorithm to draw samples from the posterior distribution of the Higgs parameters, $\mathbb{P}(E_{\rm H}, \Gamma_{\rm H}, \sigma|N_{1:B}, \hat{\Gamma}_{\rm bkg}, \hat{\alpha}, \mathsf{H}, \mathsf{I})$. Provide evidence that the posterior has been sampled effectively, using graphical methods (*e.g.*, trace plots, auto-correlation functions, corner plots) and/or numerical methods (*e.g.*, acceptance fractions, convergence statistics).

   Plot the resultant two-parameter posterior density $\mathbb{P}(E_{\rm H}, \Gamma_{\rm H}|N_{1:B}, \hat{\Gamma}_{\rm bkg}, \hat{\alpha}, \mathsf{H}, \mathsf{I})$ and the resultant one-parameter posterior density $\mathbb{P}(E_{\rm H}|N_{1:B}, \hat{\Gamma}_{\rm bkg}, \hat{\alpha}, \mathsf{H}, \mathsf{I})$ obtained from the samples. Summarise the constraints on the Higgs mass and whether this constitutes any evidence for a detection.

5. Extend the above Metropolis code to simultaneously infer both the background model and the Higgs parameters, by sampling from the joint background+source posterior distribution $\mathbb{P}(E_{\rm H}, \Gamma_{\rm H}, \sigma, \Gamma_{\rm bkg}, \hat{\alpha}|N_{1:B}, \mathsf{H}, \mathsf{I})$.

   Plot the joint constraints in i) the two background parameters and ii) the Higgs energy and rate, and compare these joint results to those obtained previously.

6. Explain how you could use Bayesian model comparison methods to use assess the degree to which this data provides evidence of the Higgs boson. In particular, state how you would deal with the fact that both models (*i.e.*, $\neg\mathsf{H}$ and H) have unspecified internal parameters.