Galaxy Zoo: Bayesian Analysis of the Green Valley

R. J. Smethurst

March 26, 2014

To conduct a fully Bayesian analysis of the problem outlined above, we must consider all possible combinations of $(t_q, \tau) = \theta$ which will be distributed with a mean and standard deviation, so that:

$$w = (\mu_\theta, \sigma_\theta) = (\mu_{t_q}, \sigma_{t_q}, \mu_\tau, \sigma_\tau)$$

We can then define the Bayesian probability $P(\theta|w) = P(t_q, \tau|w) = P(t_q|w)P(\tau|w)$ assuming that $P(t_q|w)$ and $P(\tau|w)$ are independent of each other:

$$P(t_q, \tau|w) = \frac{1}{\sqrt{4\pi^2\sigma_{t_q}^2\sigma_\tau^2}} \exp\left(-\frac{(t_q - \mu_{t_q})^2}{2\sigma_{t_q}^2}\right) \exp\left(-\frac{(\tau - \mu_\tau)^2}{2\sigma_\tau^2}\right).$$

Which is equivalent to:

$$P(\theta|w) = \frac{1}{Z_\theta} \exp\left(-\frac{\chi_\theta^2}{2}\right).$$

Therefore if we work in logarithmic probabilities:

$$\log[P(\theta|w)] = -\log(Z_\theta) - \frac{\chi_\theta^2}{2}.$$

We must then find the probability of the data given these values of theta, $P(\underline{d}|\theta, t_k^{lb})$:

$$P(\underline{d}|\theta, t_k^{lb}) = \prod_k P(d_k|\theta, t_k^{lb}),$$

where d_k is a single data point (optical and NUV colours of one galaxy). We calculate $P(d_k|\theta, t_k^{lb})$ using the predicted values for the optical ($c = opt$) and NUV ($c = NUV$) colours, $d_{c,p}(\theta, t_k^{lb})$, for a given combination of $\theta = (t_q, \tau)$ and a calculated galaxy age t^{lb} (look back time, calculated from a galaxy's redshift, equivalent to the age of the galaxy assuming that all galaxies formed at $t = 0\ Gyr$):

$$P(d_k|\theta, t^{lb}) = \frac{1}{\sqrt{2\pi\sigma_{opt,k}^2}} \frac{1}{\sqrt{2\pi\sigma_{NUV,k}^2}} \exp\left(-\frac{(d_{opt,k} - d_{opt,p}(\theta, t_k^{lb}))^2}{\sigma_{opt,k}^2}\right) \exp\left(-\frac{(d_{NUV,k} - d_{NUV,p}(\theta, t_k^{lb}))^2}{\sigma_{NUV,k}^2}\right),$$

where for one combination of $\theta = (t_q, \tau)$,

$$\chi_{c,k}^2 = \frac{(d_{c,k} - d_{c,p}(\theta, t_k^{lb}))^2}{\sigma_{c,k}^2}$$

and

$$Z_k = \sqrt{2\pi\sigma_{c,k}^2}.$$

Again working in logarithmic probabilities:

$$\log\left(P(\underline{d}|\theta,\underline{t}^{lb})\right) = \sum_{c,k} log(P(d_{c,k}|\theta, t_k^{lb}))$$

$$\log\left(P(\underline{d}|\theta,\underline{t}^{lb})\right) = K - \sum_{c,k} \frac{\chi^2_{c,k}}{2},$$

where K is a constant:

$$K = -\sum_{c,k} \log Z_{c,k}.$$

What we need however is the probability of each combination of θ given the GZ2 data, $P(\theta|\underline{d})$, which we can find by:

$$P(\theta|\underline{d}) = \frac{P(\underline{d}|\theta,\underline{t}^{lb})P(\theta)}{\int P(\underline{d}|\theta,\underline{t}^{lb})P(\theta)d\theta},$$

where,

$$P(\underline{d}|\theta,\underline{t}^{lb})P(\theta) = \exp\left[\log\left[P(\underline{d}|\theta,\underline{t}^{lb})\right] + \log\left[P(\theta)\right]\right],$$

and the denominator $\int P(\underline{d}|\theta,\underline{t}^{lb})P(\theta)d\theta$ is given by the sum across all the elements of $P(\underline{d}|\theta,\underline{t}^{lb})P(\theta)$. This denominator is a mere normalisation factor, therefore when comparing the likelihoods between two different combinations of $\theta = (t_q, \tau)$ we need only compare the numerator and can also remain in logarithmic probability space. So given the data from the GZ2 sample, we can calculate $\log[P(\theta|\underline{d},\underline{t}^{lb})]$ for all possible θ values and compare these to determine the most likely values for θ given the GZ2 data. In order to this robustly, we performed a Markov Chain Monte Carlo (MCMC) sampling method to cycle through the defined parameter space using a Python implementation of an affine invariant ensemble sampler (Foreman-Mackey et al. (2013)); *emcee*.

In addition to the colours, the GZ2 data is unique in that it provides information on a galaxy's morphology. Vote fractions from GZ2 users are available for each galaxy, for example if 80 of 100 people thought a galaxy was disc shaped, whereas 20 out of 100 people thought the same galaxy was smooth in shape (i.e. elliptical), that galaxy would have $p_s = 0.2$ and $p_d = 0.8$. We can incorporate these GZ2 vote fractions into our sampling by considering them as fractions which that galaxy contributes to the likelihood $P(d_k|\theta)$. For example a galaxy which has $p_s = 0.9$ should carry more weight in the overall likelihood than a galaxy with $p_s = 0.1$. Therefore the likelihood can now be thought of as:

$$P(\underline{d}|\theta,\underline{t}^{lb}) = \prod_k p_k P(d_k|\theta, t_k^{lb}),$$

where p_k is either p_s or p_d for an individual galaxy. We can then feed the code with the colours for all of the GZ2 data along with first the p_s vote fractions to find the most likely parameters for θ for elliptical galaxies and then with the p_d vote fractions to find the most likely parameters for θ for disc galaxies. However, this is costly in computing time, therefore we perform our sampling across four parameters so that $\theta = (t_s, \tau_s, t_d, \tau_d)$ and our likelihood function is then:

$$P(\underline{d}|\theta,\underline{t}^{lb}) = \prod_k \left[p_{s,k}P(d_k|\theta_s, t_k^{lb}) + p_{d,k}P(d_k|\theta_d, t_k^{lb})\right],$$

or,

$$\log\left[P(\underline{d}|\theta,\underline{t}^{lb})\right] = \sum_k \log\left[p_{s,k}P(d_k|\theta_s, t_k^{lb}) + p_{d,k}P(d_k|\theta_d, t_k^{lb})\right].$$

The code searches through the θ paramater space to find the region that maximises $P(\theta|\underline{d})$ to return four parameter values for t_s, τ_s, t_d and τ_d.

References

Foreman-Mackey, D., Hogg, D. W., Lang, D., Goodman, J., 2013, PASP, 125, 306

Willett, K. et al., 2014, MNRAS, 435, 2835