Bayesian Methods for Modelling the Star Formation History of Galaxy Zoo 2 Data

Becky Smethurst

January 22, 2014

First we define the properties of the star formation model, described through an exponentially decaying function of time, beginning when star formation is quenched, t_q at a quenching rate, τ. Prior to t_q the star formation rate is constant. We must consider these models across all possible values of $\underline{t_q}$ and $\underline{\tau}$ which will each be distributed with a mean and standard deviation, so that: $\underline{w} = (\mu_{t_q}, \sigma_{t_q}, \mu_\tau, \sigma_\tau)$. For a given combinaton of t_q and τ we define:

$$\theta = \begin{pmatrix} t_q \\ \tau \end{pmatrix},$$

so that,

$$\theta - \overline{\theta} = \begin{pmatrix} t_q - \mu_{t_q} \\ \tau - \mu_\tau \end{pmatrix}$$

where $\overline{\theta} = (\mu_{t_q}, \mu_\tau)$ and also:

$$\hat{C} = \begin{pmatrix} 1/\sigma_{t_q}^2 & 0 \\ 0 & 1/\sigma_\tau^2 \end{pmatrix}.$$

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

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

If we work in the arrays we defined earlier then:

$$P(\theta) = \frac{1}{Z_\theta} \exp\left[-\frac{1}{2}[\theta - \overline{\theta}]^T \hat{C}^{-1} [\theta - \overline{\theta}]\right]$$

where,

$$\chi_\theta^2 = [\theta - \overline{\theta}]^T \hat{C}^{-1} [\theta - \overline{\theta}]$$

and if N is the dimension NxN of $\hat{C}$ then,

$$Z_\theta = (2\pi)^{\frac{N}{2}} |\hat{C}|^{\frac{1}{2}}.$$

And so if we work in logarithmic probabilities:

$$\log\left(P(\theta)\right) = -\log\left(Z_\theta\right) - \frac{\chi_\theta^2}{2}.$$

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

$$P(\underline{d}|\theta, \underline{w}) = \prod_k P(d_k|\theta),$$

where d_k is a single data point (colour of one galaxy). We calculate $P(d_k|\theta)$ using the predicted values for the colour, $d_p(\theta)$, for a given combination of $\theta = (t_q, \tau)$.

$$P(d_k|\theta) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(d_k - d_p(\theta))^2}{\sigma_k^2}\right),$$

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

$$\chi_k^2 = \frac{(d_k - d_p(\theta))^2}{\sigma_k^2}$$

and

$$Z_k = \sqrt{2\pi\sigma_k^2}.$$

Again working in logarithmic probabilites:

$$\log\left(P(\underline{d}|\theta, \underline{w})\right) = \sum_k log(P(d_k|\theta))$$

$$\log\left(P(\underline{d}|\theta, \underline{w})\right) = \sum_k \log\left[\frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(d_k - d_p(\theta))^2}{\sigma_k^2}\right)\right]$$

$$\log\left(P(\underline{d}|\theta, \underline{w})\right) = K - \sum_k \frac{\chi_k^2}{2},$$

where K is a constant:

$$K = -\sum_k \log Z_k.$$

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

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

where,

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

and the $\int P(\underline{d}|\theta)P(\theta)d\theta$ is given by the sum across all the array elements of $P(\underline{d}|\theta)P(\theta)$.