Chapter 11

Model selection

11.1 Our modelled world

Let us have a set of k models that we will consider in our analyses:

$$\underline{\mathcal{M}} = \{\mathcal{M}_1, \ldots, \mathcal{M}_k\}.$$

To be clear, a model in this context is a joint probability distribution over data and parameters of interest. If we were to specify a different prior for a parameter and keep the same likelihood structure, we would be dealing with a different model. The purpose of our modelling may be to predict what might occur if we were to have more data. For simplicity, let us assume that these future data are denoted $\mathbf{x}$.

Given $\underline{\mathcal{M}}$, we will be operating in one of the following regimes:

(1) M-closed

We believe the true data generating process is in an element of $\underline{\mathcal{M}}$ with the prior distributions set up so as not to rule out the true parameterisation. In this case, we have a simple equation for our predictions of $\mathbf{x}$:

$$\pi(\mathbf{x}) = \sum_{i=1}^{k} \pi\left(\mathbf{x}|\mathcal{M}_i\right) \pi\left(\mathcal{M}_i\right)$$

by the law of total probability. As we collect more data, we will get $\pi\left(\mathcal{M}_j\right)$ getting closer to one if $\mathcal{M}_j$ is the true model.

(2) M-complete

We believe that a true data generating process exists, and, despite us being able to conceptualise it, we cannot put it in a from that makes the model accessible to us for computational purposes. However, in this case, we assume $\underline{\mathcal{M}}$ include the best models that we could currently utilise.

Here, we cannot really use the same formula as in the M-closed scenario because assigning prior probabilities for the models does not really make sense because we know that none of them is

the true model. Therefore, we are restricted to reporting $\pi(\mathbf{x}|\mathcal{M}_i)$ for each of our models and focus shifts to evaluating individual model performance.

(3) M-open

The reality is that we are almost always in a situation where we know the true model is not in $\underline{\mathcal{M}}$. Again, assigning prior probabilities for the models does not make sense, and we are left trying to evaluate predictive performance.

The solutions to handling (2) and (3) are not entirely satisfactory. In the remainder of this chapter, we will be focussing on the M-closed situation.

11.2 Nested models

The ideal situation for us is to have our alternative models contained within an overarching model. Nested models are usually used to describe regression models where simpler models (with fewer explanatory variables and interactions) are within an overarching model containing all variables and possible interactions. In our context, we want all the models in $\underline{\mathcal{M}}$ to be a special case of a all-encompassing model.

Example 11.2.1

Imagine that we have two models:

$\mathcal{M}_1$:

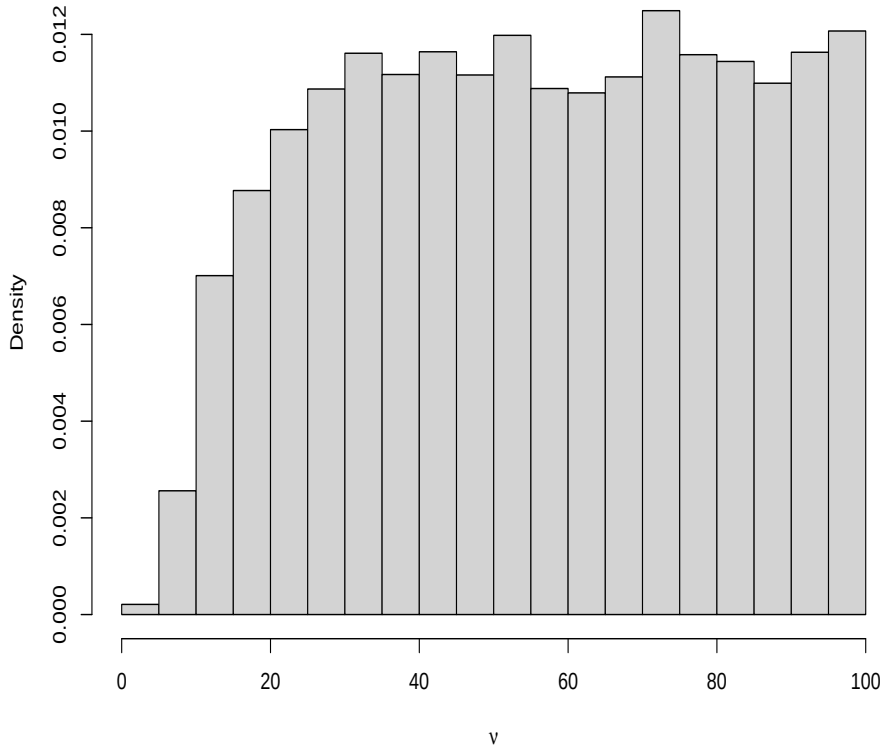
$$\begin{aligned} x_i|\mu,\sigma &\sim \mathsf{N}(\mu,\sigma^2), \quad i=1,\ldots,n, \\ \pi(\mu,\sigma); & \end{aligned}$$

$\mathcal{M}_2$:

$$\begin{aligned} x_i|m,\nu,s &\sim \mathsf{t}_\nu(m,s), \quad i=1,\ldots,n, \\ \pi(m,\nu,s). & \end{aligned}$$

As $\nu \to \infty$, $\mathcal{M}_2 \to \mathcal{M}_1$ (provided that the priors on the parameters are consistent). In fact, by the time $\nu = 30$, we are getting close to parity.

Now, let's imagine that we have collected 100 observations of x that are taken from a normal distribution and we sample from our posterior for ν under $\mathcal{M}_2$. We get the following histogram:



This would seem to give support to $\mathcal{M}_1$ because the ν are mainly favoured over values that would mean that the t-distribution and the normal would be almost indistinguishable.

A direct way to include seemingly non-compatible data-generating models within a single overarching model is to use a mixture distribution:

$$\begin{aligned} z_i|\boldsymbol{\theta} &\sim \text{Categorical}(\boldsymbol{\theta}), \quad i = 1, .., n, \\ x_i|z_1, \boldsymbol{\alpha}_1 &\sim G_1(\boldsymbol{\alpha}_1), \quad i = 1, .., n, \\ &\vdots \\ x_i|z_k, \boldsymbol{\alpha}_k &\sim G_k(\boldsymbol{\alpha}_k), \quad i = 1, .., n, \\ \pi(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_k, \boldsymbol{\theta}). & \end{aligned}$$

In this formulation, the z_n are latent variables that indicate which model the data have been drawn from. In practice, we will be uncertain about which model gave rise to the data and we will need to integrate out those variables, and we need to consider the interplay between the different α_j.

Example 11.2.2

We have two competing models for data that are known to be positive:

$$\begin{aligned} z_i|\theta &\sim \text{Bernoulli}(\theta), \quad i = 1, .., n, \\ x_i|z_1, \lambda &\sim \text{Exp}(\lambda), \quad i = 1, .., n, \\ x_i|z_2, \sigma &\sim \text{HalfNormal}(\sigma), \quad i = 1, .., n. \end{aligned}$$

If we let the pdf associated with mixture element j be $\pi_j(x|\text{parameters})$, the likelihood here (dropping the index on x) is

$$\begin{aligned} l(\lambda, \sigma, \theta; x) &\propto \pi(x|\lambda, \sigma, \theta) \\ &= \sum_{j=1}^{2} \pi(x|\lambda, \sigma, z_j)\pi(z_j|\theta) \\ &= \theta\pi_1(x|\sigma, z_j) + (1-\theta)\pi_2(x|\sigma, z_j). \end{aligned}$$

11.3 Posterior odds

If we are in the M-closed situation and we have two competing models for some data $\mathbf{x}$ say: $\mathcal{M}_1$ and $\mathcal{M}_2$. Each model will have its own set of parameters that we will need to deal with: θ_1 and θ_2 say. *A priori*, my odds for $\mathcal{M}_1$ against $\mathcal{M}_2$ are

$$\frac{\Pr(\mathcal{M}_1)}{\Pr(\mathcal{M}_2)} = \frac{\Pr(\mathcal{M}_1)}{1 - \Pr(\mathcal{M}_1)} = O_f(\mathcal{M}_1).$$

After observing data, my posterior odds for $\mathcal{M}_1$ against $\mathcal{M}_2$ are given by

$$\frac{\Pr(\mathcal{M}_1|x)}{\Pr(\mathcal{M}_2|\mathbf{x})} = \frac{\Pr(\mathcal{M}_1)}{\Pr(\mathcal{M}_2)} \frac{\pi(\mathbf{x}|\mathcal{M}_1)}{\pi(\mathbf{x}|\mathcal{M}_2)}$$

or

$$\text{POSTERIOR ODDS} = \text{PRIOR ODDS} \times \text{BAYES FACTOR}.$$

A question remains over how we calculate $\pi(\mathbf{x}|\mathcal{M}_i)$. This is an instance of the preposterior distribution under $\mathcal{M}_i$.

Throughout the module, we have been assuming some data generating model $\mathcal{M}$, but we have not been explicitly conditioning on it:

$$\pi(\theta \mid \mathbf{x}, \mathcal{M}) \propto \pi(\mathbf{x} \mid \theta, \mathcal{M})\pi(\theta|\mathcal{M}).$$

Recall that, finding the proportionality constant, we get

$$\pi(\theta \mid \mathbf{x}, \mathcal{M}) = \frac{\pi(\mathbf{x} \mid \theta, \mathcal{M})\pi(\theta|\mathcal{M})}{\pi(\mathbf{x}|\mathcal{M})},$$

where $\pi(\mathbf{x}|\mathcal{M})$ is the *evidence* for model $\mathcal{M}$.

Example 11.3.1

Consider two models:

$$\begin{aligned} \mathcal{M}_1 &: X|\theta \sim \text{Bin}(10, \theta), \quad \theta \sim \text{Be}(2, 2); \\ \mathcal{M}_2 &: X|\theta \sim \text{Bin}(20, \theta), \quad \theta \sim \text{Be}(2, 2). \end{aligned}$$

I strongly believe that the first model is correct: $\Pr(\mathcal{M}_1) = 0.9$. So my prior odds for $\mathcal{M}_1$ are 9. I observe x=10.

$$\begin{aligned} \pi(x = 10|\mathcal{M}_1) &= \int_0^1 \pi(x = 10|\theta, \mathcal{M}_1)\pi(\theta|\mathcal{M}_1)d\theta \\ &= \binom{10}{10} \frac{1}{B(2,2)} \int_0^1 \theta^{11}(1-\theta)d\theta \\ &= \frac{B(12,2)}{B(2,2)} \text{ (The previous integral was of a Be(12,2) density.)} \\ &= \frac{11! \times 1!}{13!} \frac{3!}{1! \times 1!} = \frac{1}{26}. \end{aligned}$$

Similarly, $\pi(x = 10|\mathcal{M}_2) = 11/161$. So my posterior odds for $\mathcal{M}_1$ are

$$\frac{\Pr(\mathcal{M}_1|x)}{\Pr(\mathcal{M}_2|x)} = 9 \times \frac{161}{11 \times 26} = 5.07 \text{ to 2 d.p.}$$

or my posterior probability for the first model is

$$\Pr(\mathcal{M}_1|x) = \frac{5.07}{1 + 5.07} = 0.84.$$

If I had believed that both models were equally likely *a priori*, my posterior odds for the first model would equal the Bayes factor, 0.56, and my posterior probability for the first model would be 0.36.