Lecture 9: Prior Choice

Alex Donov

22 March 2019

Previously...

We have seen that the Bayesian inference and prediction is based on the posterior distribution of the unknown model parameters, and that model selection is carried out in a similar manner.

We have spent a lot of time discussing models for the likelihood, and derivation of the posterior distribution. However, we have not said much about the choice of prior.

Standard Views

Throughout most of the 20th century, most Bayesian analysis fell into one of two categories: objective or subjective.

- *Objective*: use priors which do not contain any information about the unknown parameters, in order to let the data fully speak for itself
- *Subjective*: all probability statements are about a person's beliefs, so priors should encode what a person believes about the unknown parameter. Two people analysing the same data may have radically different priors

These are both extreme positions.

An Alternative

An alternative approach is to try and choose a prior to reflect information contained in previous/related studies.

This leads to **hierarchical modelling**. A more advanced course on Bayesian analysis will discuss this in more detail, however we can get the general idea across.

The essential idea of hierarchical modelling is to 'learn' the prior to use for the data we are analysing, by looking at related data sets.

Example

Suppose we are interested in predicting the occurrence of large earthquakes on a particular fault in southern California. On this particular fault, there have only ever been 5 large earthquakes in the last 100 years. This leads to 4 inter-arrival times which are (in years):

$$Y_0 = (10, 25, 37, 22)$$

We assume the time between events is exponentially distributed, $Exponential(\lambda_0)$. To predict future earthquakes, we need to estimate λ_0.

Example

We have seen before that the Gamma(α, β) distribution is the conjugate prior. The posterior is then:

$$p(\lambda_0|Y_0) = Gamma(\alpha + n, \beta + \sum Y)$$

Which in this case is Gamma$(\alpha + 4, \beta + 92)$.

But how should we choose the prior parameters α and β?

Objective Prior

The posterior is:

$$p(\lambda_0|Y_0) = Gamma(\alpha + n, \beta + \sum Y)$$

if we look more closely at this, we see that the prior is essentially contributing an α observations to the data, with a combined sum of β.

As such, as α and β get closer to 0, the prior becomes less informative. The (improper) Gamma(0,0) prior would be fully non-informative.

Subjective Prior

In the subjective approach we choose the prior to match our beliefs.

Where do these beliefs come from? Well, we could be a subject matter expert ourselves. If not, then we could ask an expert.

The field of 'prior elicitation' is a mix of Bayesian statistics and psychology, and contains methods to help people put quantitative numbers on their beliefs.

We have seen a basic example: rather than thinking about the whole distribution of your beliefs, instead focus only on the mean and variance, and choose the parameters to match these.

Subjective Prior

If λ_0 has a Gamma(α, β) distribution then:

$$E(\lambda_0) = \frac{\alpha}{\beta}, \quad Var(\lambda_0) = \frac{\alpha}{\beta^2}$$

Rearranging gives:

$$\alpha = \frac{E(\lambda_0)^2}{Var(\lambda_0)}, \quad \beta = \frac{E(\lambda_0)}{Var(\lambda_0)}$$

Subjective Prior

Suppose our beliefs were that on average there were 20 years between major earthquakes, but we were unsure about this so wanted to include a variance of 10.
We have:

$$\alpha = \frac{E(\lambda_0)^2}{Var(\lambda_0)}, \quad \beta = \frac{E(\lambda_0)}{Var(\lambda_0)}$$

So we would choose $\alpha = 20^2/10 = 40$, and $\beta = 20/10 = 2$.

Hierarchical Modelling

We are trying to predict the occurrence of large earthquakes that occur on a particular fault. However, there are also many other faults in Southern California that have earthquakes.

An alternative to objective vs subjective priors is to instead use the data from other faults to inform our choice of prior.

This is not 'subjective' since all prior information is coming from other data sets. But it is not 'objective' either since we are analysing the current data set using outside information.

Hierarchal Modelling

The inter-arrival times on our current fault are:

$$Y_0 = (10, 25, 37, 22)$$

But suppose there is another fault that has inter arrival times:

$$Y_1 = (8, 16, 39, 6)$$

And a third one that has inter-arrival times:

$$Y_2 = (9, 14, 20)$$

and many more. Suppose we have data from K other faults in total, with Y_k denoting the inter-arrival times on fault K.

Hierarchical Modelling

We want to use the data from faults Y_1 and Y_2 to help us better estimate the parameter λ_0 that generated the event times λ_0 on the original fault which we are interested in.

We assume that :

$$Y_0 \sim Exponential(\lambda_0)$$

$$Y_1 \sim Exponential(\lambda_1)$$

$$\ldots$$

$$Y_K \sim Exponential(\lambda_K)$$

But we are only interested in λ_0. The data from the other faults is additional data which we want to use to get a better estimate.

Notation

We have our original fault we are interested in, and K others. Let n_k denote the number of inter-arrival times on fault k.

We write $Y_{k,i}$ to denote the i^{th} inter-arrival time on fault k.

No Pooling

Simplest approach: assume that there is no statistical relationship between the different faults and analyse each independently.

This is called **no pooling**: we assume that there is no information that we can share across faults. As such, each value of λ_i is estimated using only the earthquakes on that fault.

This is the standard way of doing statistical analysis. We analyse only the data that we have in front of us.

This leads us to the $p(\lambda_0|Y_0) = Gamma(\alpha + n_0, \beta + \sum_i Y_{0,i})$ posterior we have seen before.

Complete Pooling

The problem with this approach is that since we don't have many earthquakes on each fault, our estimation of the λ_k will not be accurate. The posteriors are going to be very wide, so our predictions will not be that precise.

Remember that in general the fewer observations we have available to estimate a parameter, the less accurate our estimate will be.

If we have a good subjective prior based on accurate expert knowledge, we can use this to help bias our estimate towards the most likely values. But if we don't, then it is likely that our small data set (only 5 earthquake in this case) will result in inaccurate estimation.

Complete Pooling

An alternative approach is to assume that every fault is exactly the same, so that all interevent times have the same distribution i.e. that $\lambda_0 = \lambda_1 = \ldots = \lambda_K$

We can then combine all our data together. We have only a single value of λ to estimate now, and $n_0 + n_1 + \ldots + n_K$ observations.

If our assumption is correct, then having more available data will lead to a more accurate estimation.

Complete Pooling

The posterior is now:

$$p(\lambda_0|Y_0, Y_1, \ldots, Y_K) = Gamma(\alpha + \sum_{i=0}^{K} n_i, \beta + \sum_{k=0}^{K} \sum_{i}^{n_k} Y_{k,i})$$

We now have far more data with which to estimate λ_0. As such, our estimate will be more accurate, assuming that we were correct that the inter-event times on all faults were identically distributed.

When this assumption is violated, we may do worse than in the no-pooling case, since data from faults with radically different values of λ_k will bias our estimate away from its true value.

Complete Pooling

In the earthquake case, it is probably unreasonable to assume that all faults have the same inter-arrival time distribution.

Faults vary for many reasons – different types of faulting systems, different geometries, etc

Realistically, we expect the different λ_k values to be similar, but not identical.

Summary

We hence have two approaches:

- Complete pooling: assume that all faults are identical so that all values of λ are the same. If correct, this increases the number of available observations and hence will result in a better estimate. But if our assumption is false, the estimate will be biased.
- No-pooling: assume that all values of λ_k are different and estimate each separately. This avoids bias, but we may have only a handful of observations to use to estimate each λ_k, resulting in bad estimates

Neither approach is ideal. Hierarchical modelling aims to mediate between the two extremes.

Hierarchical Modelling

This is a situation we will often encounter. We are interested in analysing a particular data set, but we also have access to similar, related data sets.

We cannot reasonably assume that these other data sets have been generated by the exact same parameter value as ours. But they should still tell us something about the parameter value for our current data set.

Basic idea: we use these other data sets to choose an appropriate prior for our analysis.

Hierarchical Modelling

Basic idea: we use these other data sets to choose an appropriate prior for our analysis.

Why is this reasonable?

Remember that the posterior combines likelihood and prior. If the other datasets influence only the prior (not the likelihood), they give us a good 'initial guess' for the parameter value. But as we collect more observations on our fault, the likelihood will start to outweigh the prior so it matters less.

Unlike no-pooling case, we are taking the other faults into account. But unlike complete pooling, we are not assuming our fault is identical (i.e. same parameter value) to any other faults.

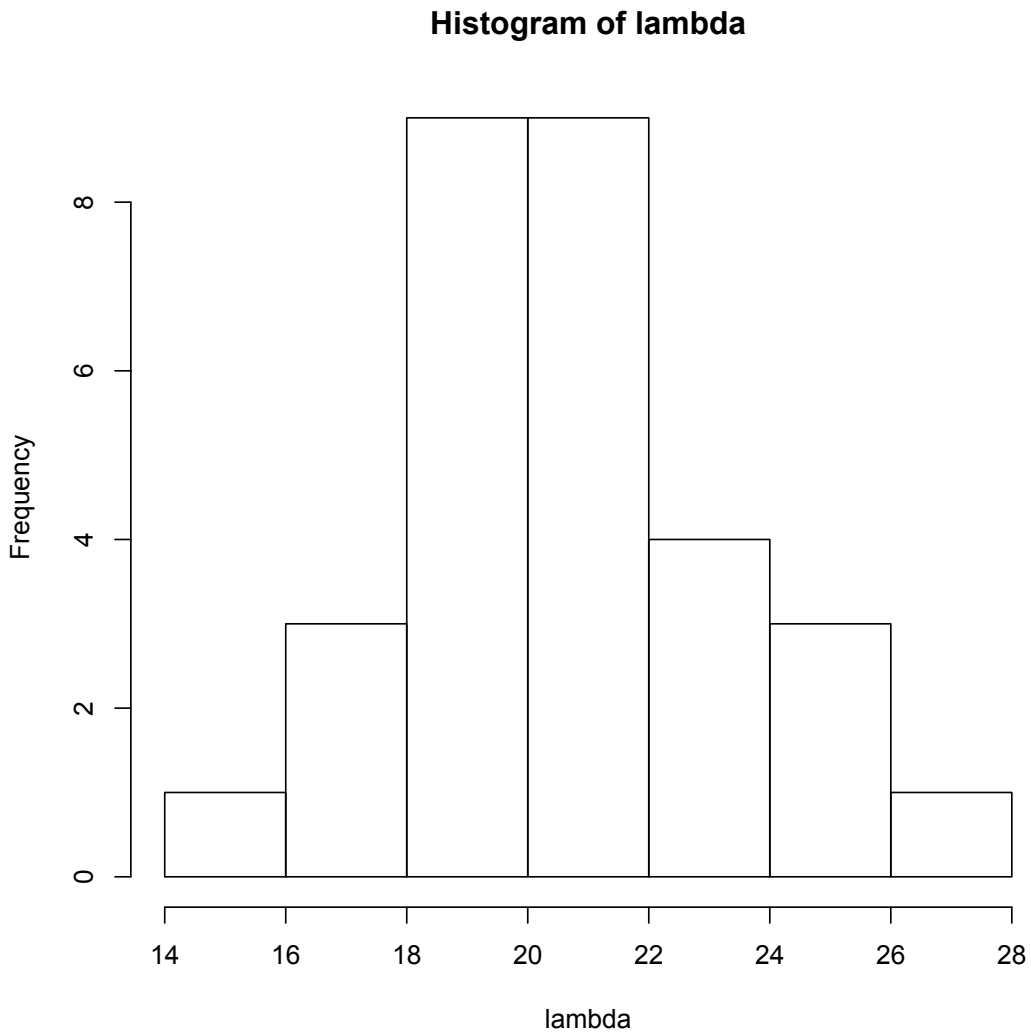
Example

Suppose we want to estimate λ_0 using data Y_0 on some fault. We also have data for 30 other faults in South California, and let λ_k denote the value of λ which generated the inter-arrival times on fault k.

Assume for now that **we know the true values of each** λ_k for $k > 0$ (i.e. for all the other 30 faults, but not the one we are trying to estimate/predict). In other words, each fault has sufficient earthquakes that these can be estimated very accurately.

We can visualise these values of λ_k using a histogram.

Example


Histogram of lambda
Frequency
0
2
4
6
8
14
16
18
20
22
24
26
28
lambda

Example

Idea: if we have no obvious reason to think our fault is different from the other faults, we can combine the no-pooling and complete-pooling approaches by using the other faults to set the parameters of the prior distribution for our fault.

In other words, we will conduct our analysis by choosing a prior distribution which matches the empirically observed histogram of λ values. We pick α and β to match the shape on the previous slide.

Example

In the simplest case, we could do this by matching the empirical mean and variance.
Considering other 30 λ_k values, suppose that the mean is 11 and the variance is 12. We know that if $\lambda_0 \sim Gamma(\alpha, \beta)$ then:

$$E(\lambda_0) = \frac{\alpha}{\beta}, \quad Var(\lambda_0) = \frac{\alpha}{\beta^2}$$

$$\alpha = \frac{E(\lambda_0)^2}{Var(\lambda_0)}, \quad \beta = \frac{E(\lambda_0)}{Var(\lambda_0)}$$

So we would choose $\alpha = 11^2/12 = 10.08$, and $\beta = 11/12 = 0.92$.

Example

This is similar to the subjective approach but we are no longer choosing the prior based on our beliefs – we are choosing it to match other data.

The underlying theoretical justification for this is a concept called **exchangeability**.

Roughly, the parameter values are exchangeable if we have no grounds for believing that any of the faults differ from each other in a structured way.

Hierarchical Model

We are essentially assuming that each of the values of λ are independent samples from some distribution $p(\lambda)$. Our model is hence:

$$\lambda_k \sim Gamma(\alpha, \beta)$$

$$Y_{k,i} \sim Exponential(\lambda_k)$$

Since we have access to many values of λ_k, we can essentially use these to estimate the parameters α and β.

We are essentially learning the prior from the data.

A Second Example

Suppose we want to survey a community in a region at risk of tsunamis, to learn about the proportion of a population that believes they are adequately prepared in the event of disaster. We give a poll to n people, where we ask the yes/no question "Are you adequately prepared for the risk of a major tsunami occurring?"

The likelihood is Binomial (similar to coin flipping) since the outcome in binary. Let Y be the number of people we survey who answer yes, and let θ be the population-level proportion who answer yes, which we are trying to estimate.

A Second Example

The model is:

$$Y \sim Binomial(n, \theta)$$

We know the Beta(α, β) distribution is the conjugate prior, and the posterior is:

$$p(\theta|Y) = Beta(\alpha + Y, \beta + n - Y)$$

We could use a non-informative prior like Beta(0,0), but if we have only surveyed a small number of people then we may not have enough observations to accurately estimate θ.

A Second Example

Again suppose we have access to the results of K previous polls about tsunami preparedness that have been carried out in communities that are similar to the one we are surveying.

Let θ_i denote the proportion of people who answered 'yes' in community i. We assume these previous polls had large enough samples that each θ_i can be treated as known exactly.

A Second Example

If θ has a Beta(α, β) distribution then the relationship between the parameters α, β and the mean/variance of θ are:

$$E[\theta] = \frac{\alpha}{\alpha + \beta}, \quad Var[\theta] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

rearranging gives:

$$\alpha = \left(\frac{1 - E[\theta]}{Var[\theta]} - \frac{1}{E[\theta]}\right)[E[\theta]]^2, \quad \beta = \alpha\left(\frac{1}{E[\theta]} - 1\right)$$

A Second Example

Suppose that from the K empirically observed θ_i values we have:

$$E[\theta_i] = 0.6, \quad Var[\theta_i] = 0.2$$

Plugging them in would give values of α and β:

$$\alpha = \left(\frac{1 - E[\theta]}{Var[\theta]} - \frac{1}{E[\theta]} \right) [E[\theta]]^2 = 0.12, \quad \beta = \alpha \left(\frac{1}{E[\theta]} - 1 \right) = 0.08$$

to use as a prior for analysing the original community of interest.

Hierarchical Model

So far we have assumed that all the other parameter values λ_i or θ_i are known exactly.

Of course in practice, we typically won't know the exact values of the other λ_i's. These must also estimated from the data on the corresponding faults.

Suppose for example we have earthquake inter-event times on 3 faults:

$$Y_0 = (10, 25, 37, 22)$$

$$Y_1 = (8, 16, 39, 6)$$

$$Y_2 = (9, 14, 20)$$

Hierarchical Model

The general hierarchical model is then:

$$\theta_i \sim p(\theta_i|\gamma)$$

$$Y_i \sim p(Y_i|\theta_i)$$

where γ is the vector of prior parameters we want to 'learn' (e.g. $\gamma = (\alpha, \beta)$ in the case of the Exponential likelihood and Gamma prior).

Hierarchical Model

In this case we want to treat γ as an extra parameter to be learned. We hence assign it a further prior distribution,to give:

$$\gamma \sim p(\delta)$$

$$\theta_i \sim p(\theta_i|\gamma)$$

$$Y_i \sim p(Y_i|\theta_i)$$

Hierarchical Model

For example in the earthquake example, we might put a Uniform prior on both α and β in the Gamma prior:

$$\alpha \sim Uniform[0, \infty]$$

$$\beta \sim Uniform[0, \infty]$$

$$\theta_i \sim Gamma(\alpha, \beta)$$

$$Y_i \sim Exponential(Y_i|\lambda_i)$$

Hierarchical Model

Fitting models like this to data goes slightly beyond the scope of this course. Essentially the conjugacy breaks down, due to the multiple layers of hierarchy.

Instead, computational methods like Gibbs sampling and Markov Chain Monte Carlo are typically used. You will cover these in a course specifically dedicated to Bayesian analysis.