Decision and Risk
Lecture 9: Model Selection

Gordon J. Ross

Last Week...

In the last few weeks we have looked at different types of data we might find in the real world

- Data with a single change point
- Data with multiple change points
- Data with gradual drift (ARCH, GARCH, etc)

These all pose challenges when it comes to making predictions about the future.

Problem

However in practice we will not know which modelling approach best fits the data

- Is there a change point or not?
- If we think there are changes, how many change points should there be?
- Does a change point model fit better than a gradual drift model?

Model Selection

This is essentially a task of **model selection**

Given a group of models (no change, a single change point, multiple change points, etc), which one is most appropriate for the data?

We briefly considered this in a previous lecture on modelling terrorism data. We will now look in more detail,

Example

Suppose that in a particular country, the number of terrorist attacks in the last 10 years have been:

$$4, 10, 12, 8, 6$$

We want to predict the number of attacks next year

Example

Suppose we assume that the number of attacks each year follow a Poisson distribution with parameter λ. The likelihood is:

$$p(Y|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{Y_i} e^{-\lambda}}{Y_i!}$$

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

$$p(\lambda|Y) = Gamma(\alpha + \sum Y_i, \beta + n)$$

The predictive distribution is hence $p(\tilde{Y}|Y) = \int p(\tilde{Y}|\lambda)p(\lambda|Y)d'\lambda$

Example

But how can do we know the Poisson model is correct? Perhaps the number of attacks is actually generated by a Geometric distribution with parameter θ. In this case the likelihood is:

$$p(Y|\lambda) = \prod_{i=1}^{n} (1-\theta)^{Y_i - 1}\theta$$

We have seen (recent ICA) that the Beta(α, β) prior is conjugate, and the posterior distribution is:

$$p(\lambda|Y) = Beta(\alpha + n, \beta + \sum Y_i)$$

This of course leads to a different predictive distribution. How we decide which is correct? We saw in the terrorism lecture that using the wrong model can lead to wildly different answers.

Bayesian Model Selection (Lecture 4 recap)

In theory, Bayesian model selection is simple and logical. Suppose we have K different models $M_1, \ldots, M_K$. In our case $K = 2$, and the models are:

- $M_1 : p(Y|\theta)$ should be modelled using an Exponential distribution
- $M_2 : p(Y|\theta)$ should be modelled using a Geometric distribution

The Bayesian approach is simply to compute the posterior distribution of both models $p(M_1|Y)$ and $p(M_2|Y)$. These respectively correspond to the belief we have about Models 1 and 2 being correct after seeing the data. We then go with the most probable model – e.g. use an Exponential distribution if $p(M_1|Y) > p(M_2|Y)$

Bayesian Model Selection

We can compute these posterior distributions using Bayes theorem. For Model 1:

$$p(M_1|Y) = \frac{p(Y|M_1)p(M_1)}{p(Y)}$$

Here $p(M_1)$ is the prior belief we have the Model 1 is correct before seeing the data, and $p(Y_1|M_1)$ is the **marginal likelihood** of the data Y under Model i. Similarly for Model 2:

$$p(M_2|Y) = \frac{p(Y|M_2)p(M_2)}{p(Y)}$$

Bayesian Model Selection

Note that $p(Y)$ occurs in both formula and does not depend on the model. Since it is common to both, we can simply ignore it.

We will also usually assume that the prior $p(M_i)$ on each model is equal – i.e. we do not assume any model is more likely than the others.

As such, the only terms that matter are the $p(Y|M_i)$ terms. We will choose the model for which $p(Y|M_i)$ is largest.

Bayesian Model Selection - Marginal Likelihood

The $p(Y|M_i)$ terms denote marginal likelihoods. Let θ_i denote the vector of unknown parameters that occurs in Model i. In our case, $\theta_1 = \{\lambda\}$ and $\theta_2 = \{\theta\}$ corresponding to the parameters of the Exponential and Geometric distributions respectively. Then:

$$p(Y|M_i) = \int p(Y|\theta_i)p(\theta_i|M_i)d\theta_i$$

where $p(\theta_i|M_i)$ is the prior in model M_i

We previously discussed using the BIC to approximate these quantities. But we can also compute them directly when everything is conjugate.

Bayesian Model Selection - Marginal Likelihood

Lets begin with the Poisson distribution. We require:

$$p(Y|M_1) = \int p(Y|\theta_1)p(\theta_1|M_1)d\theta_1$$

Using a Gamma($\alpha,' \beta$) prior we have:

$$p(Y|\theta_1) = \prod_{i=1}^{n} \frac{\lambda^{Y_i} e^{-\lambda}}{Y_i!}$$

$$p(\theta|M_1) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

Bayesian Model Selection - Marginal Likelihood

So:

$$p(Y|M_1) = \int p(Y|\theta_1)p(\theta_1|M_i)d\theta_i$$

$$= \int \left(\prod_{i=1}^{n} \frac{\lambda^{Y_i} e^{-\lambda}}{Y_i!} \right) \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \lambda$$

$$= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \frac{1}{\prod_{i=1}^{n} Y_1} \int \lambda^{S_1+\alpha-1} e^{-\lambda(\beta+n)} d\lambda, \quad S_1 = \sum Y_i$$

$$= \frac{1}{\prod_{i=1}^{n} Y_1} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \frac{\Gamma(\alpha+S_1)}{(\beta+n)^{\alpha+S_1}}$$

Bayesian Model Selection - Marginal Likelihood

Now for the Geometric distribution. We again require:

$$p(Y|M_2) = \int p(Y|\theta_2)p(\theta_2|M_2)d\theta_2$$

Using a conjugate Beta$(\alpha,' \beta)$ prior we have:

$$p(Y|\theta_2) = \prod_{i=1}^{n}(1-\theta)^{Y_i-1}\theta$$

$$p(\theta|M_2) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Bayesian Model Selection - Marginal Likelihood

So:

$$p(Y|M_2) = \int p(Y|\theta_2)p(\theta_2|M_2)d\theta_2$$

$$= \int \left(\prod_{i=1}^{n}(1-\theta)^{Y_i-1}\theta\right)\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta$$

$$= \frac{1}{B(\alpha,\beta)}\int \theta^{\alpha}(1-\theta)^{S_1-n+\beta-1}d\theta$$

$$= \frac{B(\alpha+1,\beta+S_1-n)}{B(\alpha,\beta)}$$

Bayesian Model Selection - Marginal Likelihood

So we have:

$$p(M_1|Y) = \frac{1}{\prod_{i=1}^{n} Y_1} \frac{\beta^{\alpha}}{\Gamma(\alpha)} \frac{\Gamma(\alpha + S_1)}{(\beta + n)^{\alpha + S_1}}$$

$$p(M_2|Y) = \frac{B(\alpha + 1, \beta + S_1 - n)}{B(\alpha, \beta)}$$

Both these quantities can be evaluated exactly in R. Whichever is largest is the model we should choose

Bayesian Model Selection - Nonconjugate Case

Suppose we could not choose a conjugate prior for the parameters in a particular model M_i. In this case, we could not evaluate the integral in $p(Y|M_i) = \int p(Y|\theta_i)p(\theta_i M_i)d\theta_i$.

One option would be to use numerical integration (e.g. Gaussian quadratures) as we saw in Lecture 5

Another might be to approximate the marginal likelihood using the BIC as we saw in the terrorism lecture. Lets review this

BIC

The **Bayesian Information Criterion** (BIC) approximation works as long as we have 'enough' observations Y. In this case:

$$\log p(Y|M_i) \approx \log p(Y|\hat{\theta}_i) - 0.5 k_i \log(n)$$

where:

- n is the number of observations
- $\hat{\theta}_i$ is the maximum likelihood estimate of θ_i in Model i
- k_i is the number of parameters in Model i. Both the Geometric and Poisson distributions have a single parameter

BIC - Example

For the Poisson distribution, the likelihood is again

$$p(Y|\theta_1) = \prod_{i=1}^{n} \frac{\lambda^{Y_i} e^{-\lambda}}{Y_i!}$$

The MLE can be shown to be:

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

The BIC is hence:

$$\log p(Y|M_i) \approx \log \left(\prod_{i=1}^{n} \frac{\hat{\lambda}^{Y_i} e^{-\hat{\lambda}}}{Y_i!} \right) - 0.5 \log(n)$$

$$= \sum_{i=1}^{n} \log \left(\frac{\hat{\lambda}^{Y_i} e^{-\hat{\lambda}}}{Y_i!} \right) - 0.5 \log(n)$$

BIC - Example

Similarly for the Geometric distribution the likelihood is:

$$p(Y|\theta_2) = \prod_{i=1}^{n}(1-\theta)^{Y_i-1}\theta$$

The MLE can be shown to be:

$$\hat{\theta} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} Y_i}$$

The BIC is hence:

$$\log p(Y|M_i) \approx \log\left(\prod_{i=1}^{n}(1-\hat{\theta})^{Y_i-1}\hat{\theta}\right) - 0.5\log(n)$$

$$= \sum_{i=1}^{n}\log\left((1-\hat{\theta})^{Y_i-1}\hat{\theta}\right) - 0.5\log(n)$$

Example - Change Points

Now lets return to the question of determining whether a change point exists in some data.

We observe the values $Y_1, \ldots, Y_n$. Rather than just assuming there is a change, we want to determine whether the model with a single change point fits better than the model which assumes there are no change points at all

Example - Change Points

Let $Y_1, \ldots, Y_n$ be a sequence of random variables with an Exponential distribution. Define:

- M_0: The observations are identically distributed $Y_{[} \sim \mathit{Exponential}(\lambda)$ (no change point). The model parameters are hence $\theta_0 = (\lambda)$
- M_1: A single change point exists at τ so that $Y_1, \ldots, Y_\tau \sim \mathit{Exponential}(\lambda_0)$ and $Y_{\tau+1}, \ldots, Y_n \sim \mathit{Exponential}(\lambda_1)$. The model parameters are hence $\theta_1 = (\lambda_0, \lambda_1, \tau)$

Example - Change Points

Suppose that each of $\lambda, \lambda_0, \lambda_1$ is given an independent conjugate Gamma(α, β) prior. The change point τ is given a discrete uniform prior on $(1, 2, \ldots, n-1)$.

To ease notation, lets define:

$$S_{r,s} = \sum_{i=r}^{s} Y_i,$$

Example - Change Points

The marginal likelihood for Model M_0 without a change point is then

$$p(M_0|Y) = \int p(Y|\lambda)p(\lambda|M_0)d\lambda = \int \prod_{i=1}^{n} \left(\lambda e^{-\lambda Y_i}\right) \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} d\lambda$$

$$= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \int \lambda^{n} e^{-\lambda S_{1,n}} \lambda^{\alpha-1} e^{-\beta\lambda} d\lambda =$$

$$= \frac{\beta^{\alpha}}{\Gamma(\alpha)} \int e^{-\lambda(\beta+S_{1,n})} + \lambda^{(\alpha+n)-1} d\lambda = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \frac{\Gamma(\alpha+n)}{(\beta+S_{1,n})^{\alpha+n}}$$

Example - Change Points

and similarly for Model M_1 with a single change point at an unknown location τ.

$$p(M_1|Y) = \int p(Y|\lambda_0, \lambda_1, \tau)p(\lambda_0, \lambda_1, \tau|M_2)\, d\lambda_0\lambda_1\tau$$

$$= \frac{1}{n-1}\sum_{\tau=1}^{n-1}\left(\int \left[\prod_{i=1}^{\tau} p(Y_i|\lambda_0) \prod_{i=\tau+1}^{n} p(Y_i|\lambda_1)\right] p(\lambda_0, \lambda_1|M_1)\, d\lambda\right)$$

Note that the $\frac{1}{n-1}$ term is the prior for τ. The summation comes from the fact that τ is discrete, so replace its integral with a sum to marginalise it out.

Example - Change Points

$$p(M_1|Y) = \int p(Y|\lambda_0, \lambda_1, \tau) p(\lambda_0, \lambda_1, \tau|M_2) d\lambda_0 \lambda_1 \tau$$

$$= \frac{1}{n-1} \sum_{\tau=1}^{n-1} \left(\int \left[\prod_{i=1}^{\tau} p(Y_i|\lambda_0) \prod_{i=\tau+1}^{n} p(Y_i|\lambda_1) \right] p(\lambda_0, \lambda_1|M_1) d\lambda \right)$$

$$= \frac{1}{n-1} \sum_{\tau=1}^{n-1} \left(\int \int \left[\frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda_0^{(\alpha+\tau)-1} e^{-\beta(\lambda_0 + S_{0,\tau})} \right] d\lambda_0 \left[\frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda_1^{(\alpha+n-\tau)-1} e^{-\beta(\lambda_1 + S_{\tau+1,n})} \right] d\lambda_0 \right)$$

$$= \frac{1}{n-1} \sum_{\tau=1}^{n-1} \left[\left(\frac{\beta^{\alpha}}{\Gamma(\alpha)} \right)^2 \frac{\Gamma(\alpha+\tau)}{(\beta + S_{0,\tau})^{\alpha+\tau}} \frac{\Gamma(\alpha+n-\tau)}{(\beta + S_{\tau+1,n})^{\alpha+n-\tau}} \right]$$

Example - Change Points

So we have:

$$p(M_0|Y) = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha + n)}{(\beta + S_{1,n})^{\alpha+n}}$$

$$p(M_1|Y) = \frac{1}{n-1} \sum_{\tau=1}^{n-1} \left[\left(\frac{\beta^\alpha}{\Gamma(\alpha)} \right)^2 \frac{\Gamma(\alpha + \tau)}{(\beta + S_{0,\tau})^{\alpha+\tau}} \frac{\Gamma(\alpha + n - \tau)}{(\beta + S_{\tau+1,n})^{\alpha+n-\tau}} \right]$$

These quantities can again be computed in a language like R. We assume no change point exists if $p(M_0|Y) > p(M_1|Y)$.

The case for multiple change points is identical. Define model M_i to be the model containing i change points. We can compute each $p(M_i|Y)$ in a similar manner to how we computed $p(M_1|Y)$. We choose the model with the number of change points which maximises this quantity

Example - Change Points

Similarly, suppose we wanted to determine whether a GARCH(1,1) model was more appropriate than a change point model for a particular set of financial data.

This would proceed in the same manner. We would compute the marginal likelihoods under the change point model as above, except using (e.g.) a Normal distribution for the log-returns rather than an Exponential

We would then compute a similar quantity for the GARCH(1,1) model. Since no conjugate prior exists, this would require numerical integration or the BIC approximation

Limitations of Marginal Likelihood Based Model Selection

In all of Bayesian statistics, we require a prior distribution $p(\theta)$ for model parameters which encodes our beliefs about them before taking the data into account.

For most of this course, our interest has been in the posterior distribution $p(\theta|Y) \propto p(Y|\theta)p(\theta)$, i.e estimating the parameters within a particular model.

In this case, the prior often doesn't matter too much since it gets swamped by the data if we have enough observations

Limitations of Marginal Likelihood Based Model Selection

Note the posterior is:

$$p(\theta|Y) \propto \left(\prod_{i=1}^{n} p(Y_i|\theta)\right) p(\theta)$$

As *n* gets larger, the product on the left contributes more and more terms, so dominates the prior. The prior becomes less relevant as we collect more data

Limitations of Marginal Likelihood Based Model Selection

However when computing marginal likelihoods, this is **not** generally the case. The prior always makes a significant difference, even when the amount of data is large.

This is because when we compute $p(Y|M_i)$, we are integrating with respect to the prior $p(\theta_i|M_i)$

$$p(Y|M_i) = \int p(Y|\theta_i)p(\theta_i|M_i)d\theta_i$$

The marginal likelihood is essentially a measure of how well the prior (not posterior) and the likelihood explain the data

Limitations of Marginal Likelihood Based Model Selection

When we use a model with no change point, the marginal likelihood is computing how well the prior for λ explains the data:

$$p(M_0|Y) = \int p(Y|\lambda)p(\lambda|M_0)d\lambda = \int \prod_{i=1}^{n} \left(\lambda e^{-\lambda Y_i}\right) \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}d\lambda$$

When the prior is reasonable, it should explain the data well. But if we choose a highly non-informative prior (e.g. Gamma(0,0)) it will not explain the data well. This is because it is essentially saying that all values of λ are equally likely. But most values will not give a good fit to any particular data set

Limitations of Marginal Likelihood Based Model Selection

Similarly for the model with one change point, the marginal likelihood is computing how well the prior for λ_1, λ_2 and τ explains the data. We see in the final marginal likelihood formula that we have to sum over every value of τ, treating all equally since they are all equal in the prior:

$$p(M_1|Y) = \frac{1}{n-1}\sum_{\tau=1}^{n-1}\left[\left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right)^2 \frac{\Gamma(\alpha+\tau)}{(\beta+S_{0,\tau})^{\alpha+\tau}}\frac{\Gamma(\alpha+n-\tau)}{(\beta+S_{\tau+1,n})^{\alpha+n-\tau}}\right]$$

Limitations of Marginal Likelihood Based Model Selection

The practical implication of this is that all marginal likelihoods depend on the choice of prior.

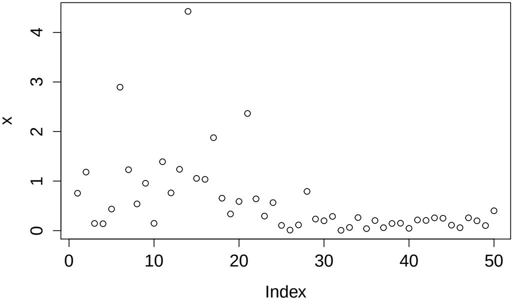
In the change point case, choosing a prior for λ that is ‘too non informative’ will result in us not detecting any change points, since very few values of λ explain the data well

Recall that the conjugate Gamma(α, β) for the Exponential becomes essentially uniform (giving all values equal weight) as α and β get close to 0.

The Gamma(0,0) prior is improper but we can approximate this by taking the limit $p(\lambda) = \mathrm{Gamma}(\epsilon, \epsilon)$ as $\epsilon \rightarrow 0$.

Limitations of Marginal Likelihood Based Model Selection

To illustrate the effect of the prior, I simulated 50 Exponential random variables from a true model with a change point at $\tau = 25$ with $\lambda_0 = 1, \lambda_1 = 5$.



Limitations of Marginal Likelihood Based Model Selection

The table on the next slide shows the corresponding marginal likelihoods for both models. The third column gives the posterior probability $p(M_0|y)$, found by normalising these to sum to 1

Limitations of Marginal Likelihood Based Model Selection

ϵ	$p(Y\|M_0)$	$p(Y\|M_1)$	$p(M_0\|y)$
1	1.3×10^{-12}	1.2×10^{-07}	1.1×10^{-5}
0.1	3.0×10^{-13}	3.6×10^{-08}	8.3×10^{-6}
0.001	3.8×10^{-14}	7.2×10^{-10}	5.3×10^{-5}
0.0001	4.0×10^{-16}	8.1×10^{-14}	4.9×10^{-3}
0.0000001	4.0×10^{-19}	8.2×10^{-20}	0.83

Table: Marginal likelihoods for models under a Gamma(ϵ, ϵ) prior

We see that as we approximate the improper prior (essentially uniform on $[-\infty, \infty]$) we stop detecting the change point.

Lindley’s Paradox

In general marginal likelihood model selection works as long as we pick priors that are ‘sensible’. We don’t really believe that a value of (e.g.) 1,000,000,000,000,000,000 is likely for financial returns or terrorism causalites, so a prior which gives it non-trivial weight isn’t realistic

This is all an example of something more general called Lindley’s paradox, which highlights the stark difference between marginal likelihood based model selection, and typical frequentist approaches. Lindley’s paradox is beyond the scope of this course (and hence not examinable) but very interesting and worth reading about.

The wikipedia page is quite nice and has a good example (again, not examinable): `https://en.wikipedia.org/wiki/Lindley's_paradox`

Alternatives to Marginal Likelihoods

As such, rather than using marginal likelihoods/Gibbs sampling, some statisticians recommend alternative procedures. A detailed discussion of this is beyond the scope of this course. however we can briefly discuss some ideas

Note this is not examinable since we are not going to go into much detail.

Prediction

Prediction: Pick the model which predicts the data best. In the time series context we have considered, we could do one step ahead prediction. Given a model M_i with parameter θ, we estimate the parameter θ using only the data up to time t:

$$p(\theta_l | Y_1, \ldots, Y_t)$$

We then predict observation Y_{t+1} using the standard prediction equation we have seen many times:

$$p(Y_{t+1} | Y_1, \ldots, Y_t) = \int p(Y_{t+1} | \theta) p(\theta | Y_1, \ldots, Y_t) d\theta$$

Note this integral is respect to the posterior, not prior!

Prediction

We repeat this for every choice of t, i.e. for each observation Y_t we estimate θ using only the observations up to Y_t, and then predict Y_{t+1}.
We are hence doing repeated one-observation-ahead prediction.
The overall score is then the product of these:

$$\prod_{t=1}^{n-1} p(Y_{t+1}|Y_1, \ldots, Y_t)$$

We then choose the model which gives the best predictive score. This approach avoids marginal likelihoods, and all integrals use only posteriors

Cross-validation

Another popular choice is cross-validation, which is very similar to the above except in a non-time series context

Cross-validation is similar to bootstrapping. We randomly split the data up into 2 subsets, known as the training and test set. Denote these by Y_{train} and Y_{test}.

We estimate θ using only the training set to get the posterior $p(\theta_l|Y_{train})$, and then predict the test set:

$$p(Y_{test}|Y_{train}) = \int p(Y_{test}|\theta)p(\theta_l|Y_{train})d\theta$$

This is then repeated many times, with the random split into training and test sets being different each time. We agin choose the model with the best predictive accuracy