Decision and Risk
Lecture 8: Gradual Drift

Gordon J. Ross

Last Week...

For the last two weeks we have discussed how to answer questions such as ”what is the probability of extreme events occurring?” in situations where the distribution of the data undergoes change.

We have $Y = \{Y_1, \dots, Y_n\}$ and want to know $p(\tilde{Y} > D|Y)$ where $\tilde{Y}$ represents an observation in the future

However since the distribution is not constant, we cannot assume $Y_1, \dots, Y_n \sim p(\cdot|\theta)$ are identically distributed

Last Week...

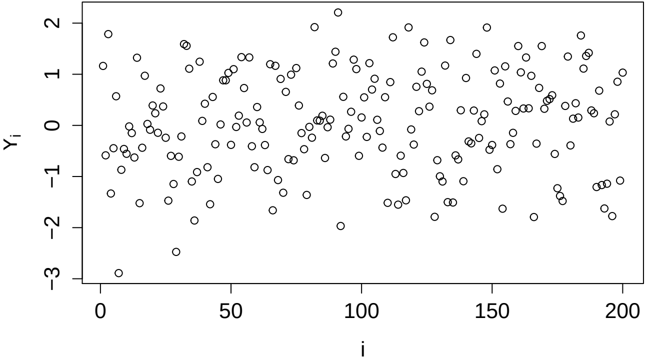
We looked at examples with **change points** where the distribution shifted. In the *k* change point model we have:

$$Y_i = \begin{cases} p(Y_i|\theta_1) & \text{if } i \leqslant \tau_1 \\ p(Y_i|\theta_2) & \text{if } \tau_1 < i \leqslant \tau_2 \\ p(Y_i|\theta_3) & \text{if } \tau_2 < i \leqslant \tau_3 \\ \dots & \\ p(Y_i|\theta_{k+1}) & \text{if } \tau_k < i \leqslant n \end{cases}$$

However in practice not all change is abrupt like this. Sometimes parameters will **drift gradually**

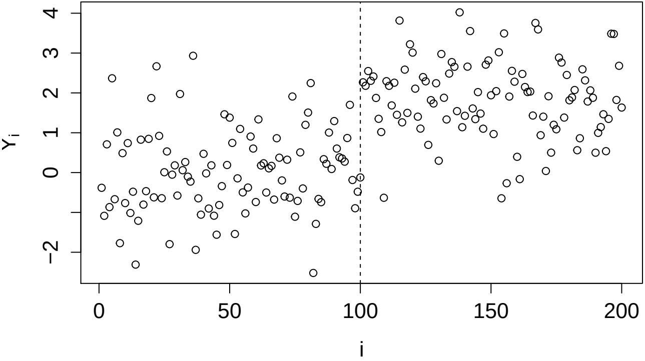
Independent Observations

$$Y_1, \ldots, Y_{200} \sim N(0, 1)$$

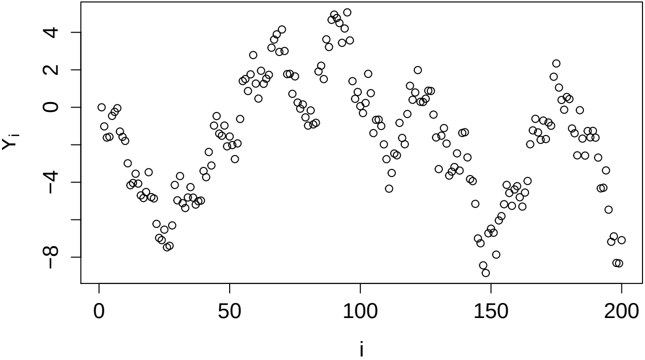


Change Point

$$Y_i = \begin{cases} N(0,1) & \text{if } i \leqslant 100 \\ N(2,1) & \text{if } i > 100 \end{cases}$$



Gradual Drift



Gradual Drift

In the gradual drift case we can see clearly that the observations are not independent – the mean of the sequence is changing over time.

However the mean seems to be slowly changing – it does not jump abruptly like in the change point model

With data like this, a different approach is required

Models for Gradual Drift

There are many ways to model gradually drifting data. One of the simplest methods for modelling a gradually changing mean is the random walk model:

$$Y_1 = c$$

$$Y_i = Y_{i-1} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

In other words, the sequence starts with some value c, and then each observation is equal to the last one, with zero mean Gaussian noise added on

Simulating a Random Walk

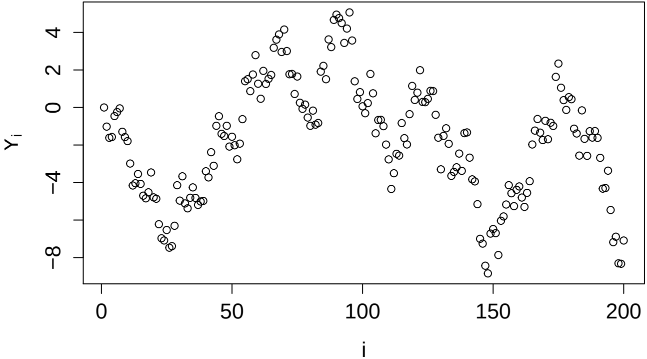
We can simulate a random walk in R very easily. Here is an example where $\sigma^2 = 1$ and $c = 0$

```
n <- 200 #length of sequence
y <- numeric(n)
sigma <- 1

y[1] <- 0
for (i in 2:length(y)) {
  y[i] <- y[i-1] + rnorm(1,0,sigma)
}
plot(y)
```

Simulating a Random Walk

Example simulated sequence, $\sigma = 1$



Simulating a Random Walk

For this random walk model:

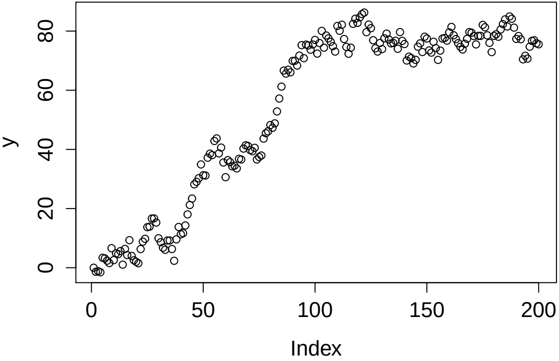
$$Y_1 = c$$

$$Y_i = Y_{i-1} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

the key parameter (usually unknown) is the variance σ^2. When this is larger, the random walk will exhibit more variance, in the sense of having periods where it can move far away from the origin. The next slide shows a random walk with $\sigma = 3$

Simulating a Random Walk

Example simulated sequence, $\sigma = 3$



Conditional vs Unconditional Distributions

When working with sequences which have a distribution which changes over time, it is useful to distinguish between the **unconditional** and **conditional** distributions of the data.

These are very different, and understanding this difference is important when it comes to predicting the future

Roughly, the conditional distribution of Y_i is its distribution when we condition on the previous values $Y_1, \ldots, Y_{i-1}$, and its unconditional distribution is its distribution when we don't.

Unconditional Distribution

The unconditional distribution is the distribution of Y_t across different many realisations of the sequence. Suppose we want to know the unconditional distribution of Y_{10} in the random walk model when $c = 0$ and $\sigma^2 = 1$

Suppose we simulated 1000 realisations of the sequence in R, each one being different from the others. For each sequence, we note the value of Y_{10}. The distribution of quantity is the unconditional distribution of Y_{10}

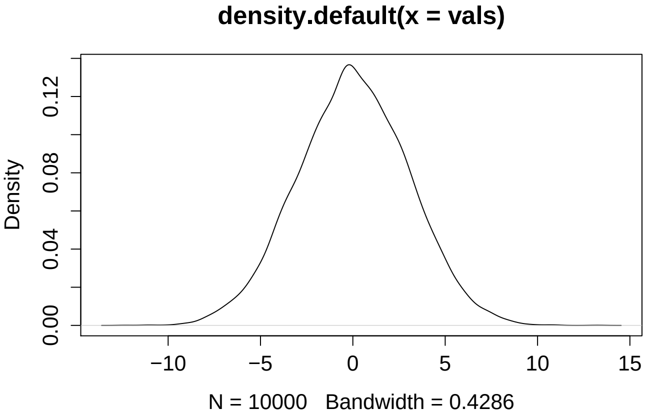
We can do this in R:

Unconditional Distribution of Y_{10}

```
sims <- 1000
vals <- numeric(sims)

for (s in 1:sims) {
  y <- numeric(10)
  y[1] <- 0
    for (i in 2:length(y)) {
      y[i] <- y[i-1] + rnorm(1,0,sigma)
  }
  vals[s] <- y[10]
}
plot(density(vals))
```

Unconditional Distribution of Y_{10}



Unconditional Distribution of Y_{10}

Note we can also find this analytically. Recall that $Y_1 = 0$ and for each Y_i:

$$Y_i = Y_{i-1} + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2)$$

So Y_{10} is just the sum of 9 independent $N(0, \sigma^2)$ random variables. By basic properties of the Normal distribution we hence have:

$$Y_{10} \sim N(0, 9\sigma^2)$$

and in general for Y_i we have:

$$Y_i \sim N(0, (i-1)\sigma^2)$$

Conditional Distribution of Y_{10}

So, the unconditional distribution of Y_i is its distribution across multiple realisations of the sequence.

The **conditional** distribution of Y_i is its distribution in a particular realisation of the sequence, based on the previous values.

Again consider Y_{10}. Suppose we know the values of $Y_1, \ldots, Y_9$. Then, what is the distribution of Y_{10}?

Conditional Distribution of Y_{10}

By definition, $Y_i = Y_{i-1} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_2)$.

So, we know that given $Y_1, \ldots, Y_9$, we have that Y_{10} is equal to Y_9 plus a $N(0, \sigma^2)$ random variable. So;

$$Y_{10} | Y_1, \ldots, Y_9 \sim N(Y_9, \sigma^2)$$

Conditional vs Unconditional Distribtuion

So in general we have the unconditional distribution:

$$p(Y_i) = N(c, \sqrt{i-1}\sigma^2)$$

and the conditional distribution

$$p(Y_i|Y_1, \ldots, Y_{i-1}) = N(Y_{i-1}, \sigma^2)$$

In practice, the conditional distribution is more useful If we want to know what will happen tomorrow, it makes sense to conditional on all the available historical data.

Prediction

Suppose we have observed $Y_1, \ldots, Y_i$ and we want to know the probability that $Y_{i+1} > D$ for some D. As always, we need the predictive distribution of Y_{i+1}

Suppose that σ^2 is known exactly. In this case the predictive distribution is simply the conditional distribution:

$$Y_{i+1} \sim N(Y_i, \sigma^2)$$

So $p(Y_{i+1} > D) = 1 - pnorm(D, Y_i, \sigma)$

Parameter Estimation

In practice we do not know the value of σ^2, and it must be estimated. We do this in the standard Bayesian way.

Note that since

$$Y_i = Y_{i-1} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Then if we define a new set of variables $Z_1, \ldots, Z_{n-1}$ where:

$$Z_i = Y_{i+1} - Y_i$$

Then the Z_i variables are independent with a $N(0, \sigma^2)$ distribution

Parameter Estimation

As such, estimating σ^2 here is equivalent to estimating the unknown variance σ^2 for a sequence of independent and identically distributed variables $Z_1, \ldots, Z_n$ which have a N$(0,\sigma^2)$ distribution.

We have learned how to do this already! We use an Inverse-Gamma prior on the variance, and proceed directly using Bayes Theorem. If the prior is $IG(\alpha, \beta)$ then:

$$p(\sigma^2 | Y_1, \ldots, Y_n) = IG\left(\alpha + (n-1)/2, \beta + \sum_{i=1}^{n-1} Z_i\right)$$

and the predictive distribution comes from integrating over this as usual

AR(p) Models

The random walk model is a special case of the general $AR(p)$ model (AR here stands for 'auto-regressive'). An AR(1) model is defined as:

$$Y_i = \beta_0 + \beta_1 Y_{i-1} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Similar to the random walk model except the two additional parameters β_0 and β_1 allow the data to have an unconditional mean other than 0, and to be mean-reverting

AR(p) Models

An AR(2) model has the following form:

$$Y_i = \beta_0 + \beta_1 Y_{i-1} + \beta_2 Y_{i-2} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

It is very similar except that the distribution of Y_i depends on Y_{i-2} as well as Y_{i-1}. The unknown parameters are now $(\beta_0, \beta_1, \beta_2, \sigma^2)$

AR(p) Models

The general AR(p) model has the following form

$$Y_i = \beta_0 + \beta_1 Y_{i-1} + \beta_2 Y_{i-2} + \beta_3 Y_{i-3} + \ldots + \beta_p Y_{i-p} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

The unknown parameters are $(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2)$

AR(p) Models - Parameter Estimation

The AR(p) process has parameters $(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2)$. As always, these can either be estimated using traditional frequentist methods (e.g. maximum likelihood), or Bayesian inference.

The Bayesian approach proceeds as usual: we start with a prior $p((\beta_0, \beta_1, \ldots, \beta_p, \sigma^2)$ and form the posterior:

$$p(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2 | Y_1, \ldots, Y_n) = \frac{p(\beta_0, \beta_1, \ldots, \beta_p, \sigma^2) p(Y_1, \ldots . Y_n | \beta_0, \beta_1, \ldots, \beta_p}{p(Y)}$$

AR(p) Models - Parameter Estimation

When p is low such as in the AR(1) or AR(2) models, it is possible to evaluate this posterior using the methods we have learned. In the conjugate case when everything has a Normal distribution (including the priors), we can use Bayesian inference for the Normal distribution as we have seen throughout this course - see Exercise sheet.

when the prior is not conjugate, we can use numerical integration techniques such as Simpsons Rule or quadratures.

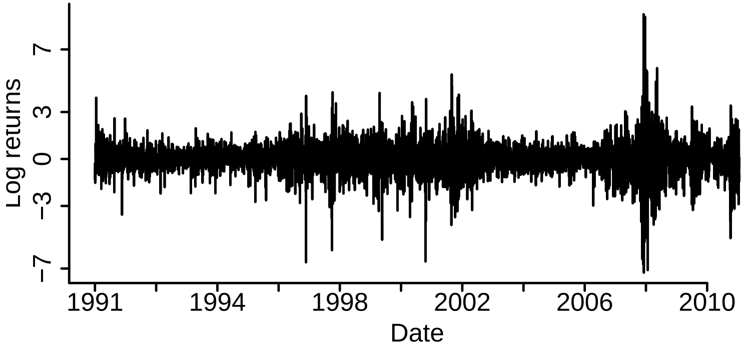
However when p is large, the number of parameters starts to get unwieldy. When everything is conjugate, we can still find the posterior easily, but in the non-conjugate case we require numerical integration techniques which are beyond the scope of this module (such as Markov Chain Monte Carlo [MCMC]).

Drifting Variance

So far we have considered models where the mean drifts gradually. However in many situations – such as financial returns – the mean is usually fairly constant over time and it is instead the **variance** of the returns which changes

Recall the Dow Jones index we have discussed before. We worked with the log returns $Y_i = \log\left(\frac{P_i}{P_{i-1}}\right)$ where P_I is the price on day i. We saw that these roughly seemed to have a Normal distribution with mean 0 and a variance that changed over time

Dow Jones



ARCH and GARCH Models

Almost all financial return series obey this pattern where the variance changes over time. In previous lectures we have seen how the change point formulation can be used to model the variance changes in this type of data.

A different approach is to treat the variance as gradually drifting, rather than time varying. Both approaches are widely used in industry

The most popular time-varying variance models used in finance are the **ARCH** and **GARCH** models

ARCH Models

The simplest way to model gradually drifting variance is the ARCH(1) model. ARCH stands for ”Autoregressive Conditional Heteroskedasticity”. It assumes that the returns Y_i have zero mean, and a gradually changing variance:

$$Y_i = \sigma_i \epsilon_i, \quad \epsilon \sim N(0,1)$$

$$\sigma_i^2 = \beta_0 + \beta_1 Y_{t-1}^2$$

Lets unpack this definition to be clear what is going on...

ARCH Models

$$Y_i = \sigma_i \epsilon_i, \quad \epsilon \sim N(0,1)$$

The log return Y_i at time i has 0 mean, and a variance which is time dependent. Specifically, the variance is equal to σ_i^2 – remember that simulating a random variable with a $N(0,1)$ distribution and multiplying it by σ is equivalent to simulating a variance from a $N(0,\sigma^2)$ distribution

So this line simply says that the **conditional** distribution of Y_t is $N(0,\sigma_i^2)$, where σ_i^2 depends on the previous value Y_{t-1} (hence conditional!)

ARCH Models

$$\sigma_i^2 = \beta_0 + \beta_1 Y_{i-1}^2$$

This line specifies how the variance evolves over time. It is similar to the random walk model from before.

When the variance evolves according to the above equation, this is called an ARCH(1) model. Essentially in the ARCH model, it is the **variance** that is following a random walk, not the mean

Simulating from an ARCH(1) model

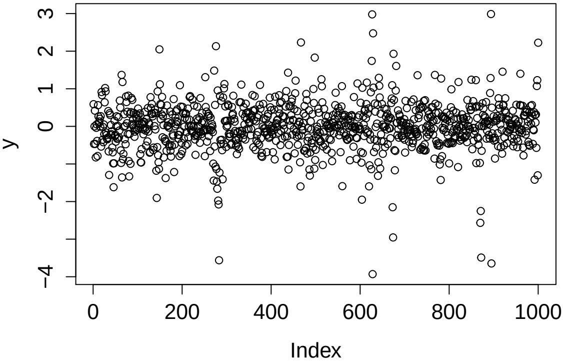
```
beta0 <- 0.1; beta1 <- 0.8
n <- 1000
sigma2s <- numeric(n);
y <- numeric(n)

sigma2s[1] <- 1;
y[1] <- rnorm(1,0,sqrt(sigma2s[1]))

for (i in 2:n) {
  sigma2s[i] <- beta0 + beta1 * y[i-1]^2
  y[i] <- rnorm(1,0,sqrt(sigma2s[i]))
}
```

ARCH(1)

Using the above parameter values, this is an example simulated sequence $Y_1, \dots, Y_n$:



GARCH(1,1) Model

The GARCH(1,1) model is very similar to the ARCH(1) model, but has a single extra term:

$$Y_i = \sigma_i \epsilon_i, \quad \epsilon \sim N(0,1)$$

$$\sigma_i^2 = \beta_0 + \beta_1 Y_{t-1}^2 + \beta_2 \sigma_{i-1}^2$$

The only difference is the $\beta_2 \sigma_{i-1}^2$ term. For those who have taken a time-series class before, this is essentially an ARMA(1,1) model applied to the **variance**

The extra term essentially allows high values of the variance to persist for longer over time. This typically results in a better fit to real financial returns data.

The GARCH(1,1) model has 3 unknown parameters: $(\beta_0, \beta_1, \beta_2)$.

GARCH(1,1) Model

Similar to the AR(p) model, are extensions of the GARCH model which include more terms.

However these are mainly of academic interest. In practice, when people (both in academia and industry – banks, hedge funds, etc) use a GARCH model for time-varying variance, it is overwhelmingly the GARCH(1,1) model which is chosen. Despite being a simple model with only 3 parameters, it tends to give a very good fit to real financial data.

The following slides show how we can simulate example GARCH(1,1) sequences in R

Simulating from a GARCH(1,1) model

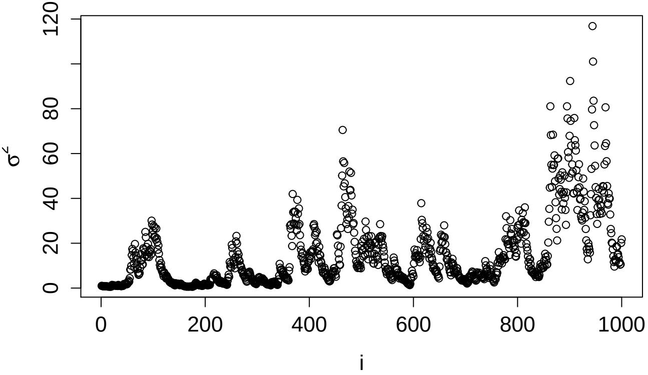
```
beta0 <- 0.1; beta1 <- 0.2; beta2 <- 0.8
n <- 1000
sigma2s <- numeric(n);
y <- numeric(n)

sigma2s[1] <- 1;
y[1] <- rnorm(1,0,sqrt(sigma2s[1]))

for (i in 2:n) {
  sigma2s[i] <- beta0 + beta1 * y[i-1]^2 + beta2 * sigma2s[i-1]
  y[i] <- rnorm(1,0,sqrt(sigma2s[i]))
}
```

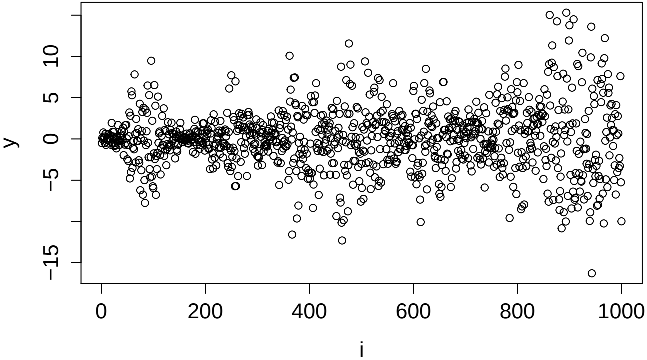
GARCH(1,1) Model

Using the above parameter values, this is an example simulated realisation of σ^2:



GARCH(1,1) Model

And this is the corresponding simulated Y_i values, where each has a variance of σ_i^2, from the previous slide. We can see this has similar features to what we have observed in financial data



GARCH(1,1) Parameter Estimation

The parameters of the GARCH model are $\theta = \{\beta_0, \beta_1, \beta_2\}$. Given a sequence of real data such as the Dow Jones index returns, we need to estimate these to fit the model

Estimating these can be tricky - there are typically no conjugate priors, and the integrals must be done numerically. The estimation is also difficult in a frequentist context - the maximum likelihood estimates do not have a standard form, and numerical maximisation of the likelihood function must be performed instead.

Since there are only 3 parameters, we can use numerical integration/quadratures (see Exercise sheet). More advanced methods such as Markov-Chain Monte Carlo are also useful, but beyond the scope of this module.

Model Selection

A Problem

Last week, we discussed multiple change point models. Given that we know there are k change points in a sequence of data:

$$Y_i = \begin{cases} p(Y_i|\theta_1) & \text{if } i \leqslant \tau_1 \\ p(Y_i|\theta_2) & \text{if } \tau_1 < i \leqslant \tau_2 \\ p(Y_i|\theta_3) & \text{if } \tau_2 < i \leqslant \tau_3 \\ \dots & \\ p(Y_i|\theta_{k+1}) & \text{if } \tau_k < i \leqslant n \end{cases}$$

We showed how to estimate their locations $\tau_1, \dots, \tau_k$

A Problem

Today we have discussed models such as the AR(p) and various GARCH models.

But: **how do we know which model to use in practice**?

How do we decide whether there should be 0, 1, 2, or more change points? How do we decide whether a change point model is more appropriate model than a GARCH?

This is the task of **model selection**

Model Selection

We previously discussed model selection in Lecture 3, in the context of choosing between the Exponential and Lognormal models for modeling the time between terrorist attacks.

It is worth reviewing this material again, now that we have reached a more advanced stage of the course

Model Selection

Model selection is the task of choosing which of two or more different probability models (usually likelihood functions) are better suited to modelling a particular data set.

For example, we here want to choose between a model with 0 change points, and a model with 1 change point. Or we want to choose between an ARCH(1) and a GARCH(1,1) model.

Model selection is one of the most controversial areas of statistics – if you ask a group of statisticians the best way to do it, you will get several different answers.

Bayesian Model Selection

In theory, Bayesian model selection is simple and logical. Suppose we have K different models $M_1, \ldots, M_K$. For example, in a change point context with data $Y = (Y_1, \ldots, Y_n)$ the models could be:

- $M_0 : Y_1, \ldots, Y_n$ does not contain any change points, and the observations are independent and identically distributed
- $M_1 : Y_1, \ldots, Y_n)$ contains a single change point $(k = 1)$
- $M_2 : Y_1, \ldots, Y_n)$ contains exactly two change points $(k = 2)$
- $\ldots$

The Bayesian approach is simply to compute the posterior distribution of all models $p(M_0|Y), p(M_1|Y), \ldots$. These respectively correspond to the belief we have about Models $0, 1, \ldots$ being correct after seeing the data. We then go with the most probable model, i.e. the one for which the posterior is highest.

Bayesian Model Selection

We can compute these posterior distributions using Bayes theorem. For Model 0:

$$p(M_0|Y) = \frac{p(Y|M_0)p(M_0)}{p(Y)}$$

Here $p(M_0)$ is the prior belief we have the Model 0 is correct before seeing the data, and $p(Y|M_0)$ is the **marginal likelihood** of the data Y under Model 0. Similarly for Model 1:

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

and so on

Bayesian Model Selection

Note that $p(Y)$ occurs in the denominator of every one of these posteriors, and does not depend on the model. Since it is common to all posteriors, we can simply ignore it.

We can sometimes (but not always!) 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.

If we assume equal priors, 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. Then:

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

Previously in Lecture 3, we said these integrals were hard to compute, and approximated them using the Bayesian Information Criterion. This is still a feasible way to proceed in cases where (e.g.) the prior is not conjugate. But we now know how to do these integrals in many cases

Example - Exponential Distribution

Lets consider the case where the observations have an Exponential distribution, as we done last week:

$$Y_i = \begin{cases} Exponential(\lambda_1) & \text{if } i \leqslant \tau_1 \\ Exponential(\lambda_2) & \text{if } \tau_1 < i \leqslant \tau_2 \\ Exponential(\lambda_3) & \text{if } \tau_2 < i \leqslant \tau_3 \\ \quad \ldots & \\ Exponential(\lambda_{k+1}) & \text{if } \tau_k < i \leqslant n \end{cases}$$

But now we do not know the value of k – we do not know how many change points there are

We assume that all the λs have the same Gamma(α, β) prior which does not depend on the model (i.e. it does not depend on how many change points there are)

Bayesian Model Selection - Marginal Likelihood

We begin with Model M_0, i.e. no change points. In this case, the observations are all i.i.d from an Exponential(λ) distribution.The marginal likelihood is:

$$p(Y|M_0) = \int p(Y|\lambda)p(\lambda)d\lambda$$

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

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+n)}{(\beta + \sum_{i=1}^{n} Y_i)^{(\alpha+n)}}$$

Note that unlike the predictive distribution for future observations, this integral is with respect to the prior, **not** the posterior.

Bayesian Model Selection - Marginal Likelihood

For Model M_1 with a single change point:

$$Y_i = \begin{cases} \mathit{Exponential}(\lambda_1) & \text{if } i \leqslant \tau \\ \mathit{Exponential}(\lambda_2) & \text{if } \tau < i \leqslant n \end{cases}$$

We saw last week that given the change point location τ the marginal likelihood is:

$$p(Y|\tau) = \int p(Y|\lambda)p(\lambda)d\lambda$$

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

$$= \left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right)^2 \frac{\Gamma(\alpha+\tau)}{(\beta + \sum_{i=1}^{\tau} Y_i)^{(\alpha+\tau)}} \frac{\Gamma(\alpha+n-\tau)}{(\beta + \sum_{i=\tau+1}^{n} Y_i)^{(\alpha+n-\tau)}}$$

Bayesian Model Selection - Marginal Likelihood

The marginal likelihood for Model M_1 with 1 change point is found by averaging over the unknown change point τ, i.e.:

$$p(Y|M_1) = \sum_{\tau=1}^{n-1} p(Y|\tau)p(\tau)$$

if we have a uniform prior on the change point location, this becomes:

$$\frac{1}{n-1}\sum_{\tau=1}^{n-1}\left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right)^2 \frac{\Gamma(\alpha+\tau)}{(\beta+\sum_{i=1}^{\tau} Y_i)^{(\alpha+\tau)}} \frac{\Gamma(\alpha+n-\tau)}{(\beta+\sum_{i=\tau+1}^{n} Y_i)^{(\alpha+n-\tau)}}$$

Note: I would never ask you to evaluate an expression this complicated in an exam. But I would expect you to understand where it comes from. We are essentially applying the Theorem of Total Probability at every stage (very similar to last week)

Bayesian Model Selection - Marginal Likelihood

So we have seen how to evaluate the expressions $p(Y|M_0)$ and $p(Y|M_1)$. The marginal likelihoods for models with 2 or more change points are obtained in identical ways.

, to choose the number of change points, we simply compute these expressions given the data Y and our choice of prior parameters α and β. We then go with the model which has the highest marginal likelihood

So for example if we have that $p(Y|M_1) > p(Y|M_i)$ for all $i \neq 1$, we would choose the model with a singe change point

Choosing between (e.g.) ARCH(1) and GARCH(1,1) models is handled in the same way – we compute marginal likelihoods of the data under both models and go with the best.

Two Potential Problems

There are two potential problems with Bayesian model selection:

- In cases where we have non-conjugate priors, it can be very hard to compute marginal likelihoods. In some cases, we may want to use approximations instead (e.g. the Bayesian Information Criterion from Lecture 3). But these may not be accurate with limited data
- You may have noticed in the Exponential change point example that Bayesian model selection depends crucially on the choice of Gamma(α, β) prior parameters. This is almost always the case – Bayesian model selection can be very sensitive to the choice of prior. Some people see this as a benefit, while others see ti as a drawback. The benefit is that since marginal likelihoods average over the prior, complex models are penalized so the Bayesian approach is less likely to overfit. But the drawback is that it is impossible to test the 'model itself' (e.g. GARCH vs ARCH, Exponential vs Lognormal, etc), only the combination of model+prior