# Decision and Risk
# Lecture 7: Change Points (continued)

Gordon J. Ross

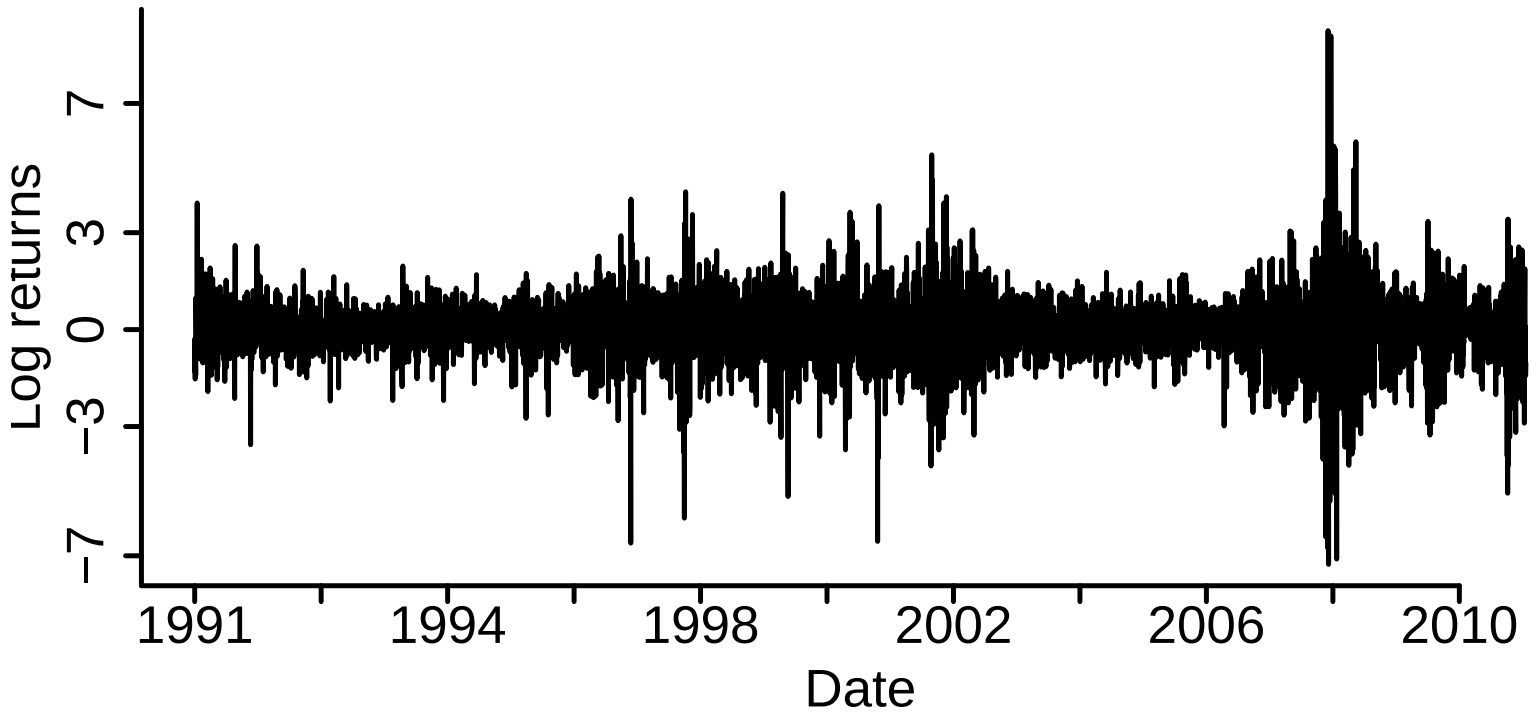
# Last Week...

Last week we started to discuss how we could answer questions such as "what is the probability of extreme events occurring?" in situations where **the distribution of the data is not constant over time**

We have $Y = \{Y_1, \ldots, Y_n\}$ 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 $Y_1, \ldots, Y_n \sim p(\cdot|\theta)$ are identically distributed

# Last Week...



# Last Week...

We explored the following setting: suppose we believe there is a single change point in the sequence distribution but do not know where it occurs.

Denote this unknown change point by $\tau$. Before the change point, the unknown parameter $\theta$ has value $\theta_1$, and after it changes to $\theta_2$. The distribution of the data is then:

$$Y_i \sim \begin{cases} p(Y_i|\theta_1) & \text{if } i \leqslant \tau \\ p(Y_i|\theta_2) & \text{if } i > \tau \end{cases}$$

# Example

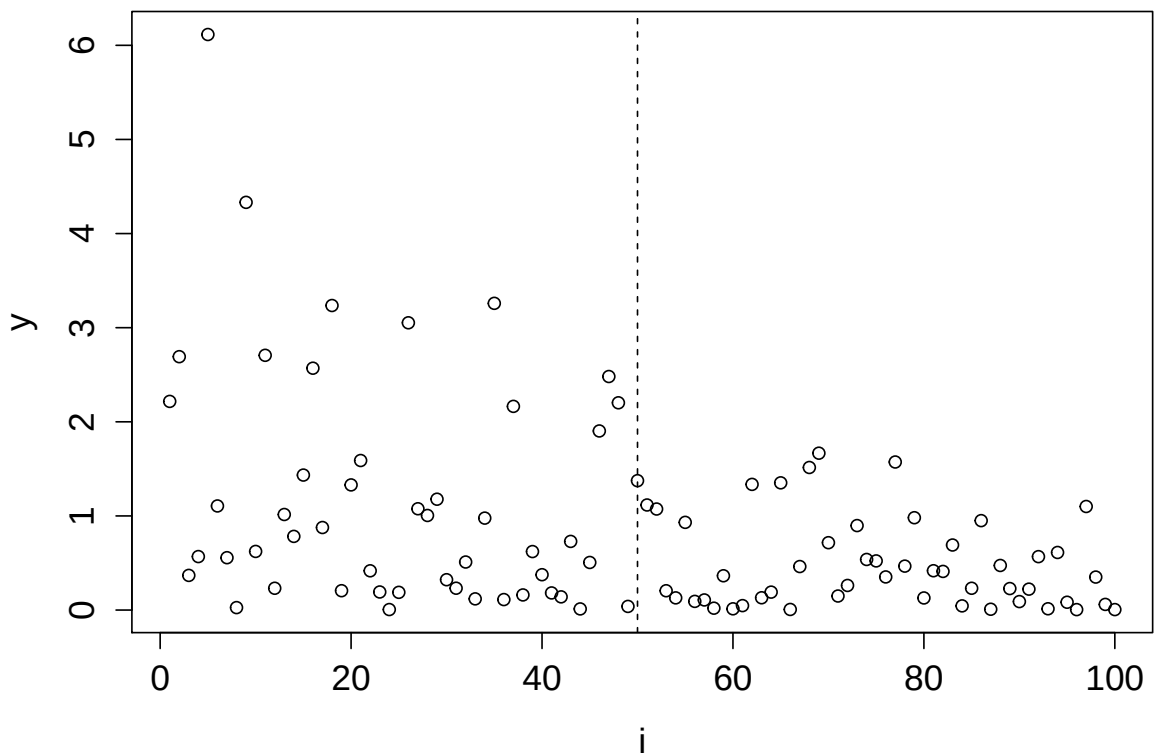
We illustrate with the example where the observations have an Exponential distribution:

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

and $\lambda_1, \lambda_2, \tau$ are all unknown. We seek to estimate $\tau$

# Example

$Y_1, \ldots, Y_{100} \sim \textit{Exponential}(\lambda)$ where the change point is at $\tau = 50$ and $\lambda_1 = 1$ before this, and $\lambda_2 = 5$ after:



# How to Detect Change Points

We saw that we could estimate $\tau$ using Bayes Theorem:

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

If we use a (non-informative) uniform prior $p(\tau) = 1/(n-1)$ then this does not depend on $\tau$ and can hence be ignored, since we will normalise the posterior at the end.

# The Likelihood

The likelihood $p(Y|\tau)$ depends on the unknown parameters $\lambda_1, \lambda_2$. If these were known, it would be:

$$p(Y|\tau, \lambda_1, \lambda_2) = \prod_{i=1}^{\tau} p(Y_i|\lambda_1) \prod_{i=\tau+1}^{n} p(Y_i|\lambda_2)$$

Since they are unknown, we integrate over them:

$$p(Y|\tau) = \int \left( \prod_{i=1}^{\tau} p(Y_i|\lambda_1) \right) p(\lambda_1) d\lambda_1 \int \left( \prod_{i=\tau+1}^{n} p(Y_i|\lambda_2) \right) p(\lambda_2) d\lambda_2$$

# How to Detect Change Points

In the Exponential-Gamma case, we can do this integral using the same trick we always use. The likelihood of the observations $Y_1, \ldots, Y_\tau$ to the left of the change point is:

$$p(Y_1, \ldots, Y_\tau | \tau) = \int \prod_{i=1}^{\tau} p(Y_i | \lambda_1) p(\lambda_1) d\lambda_1$$

$$= \int \prod_{i=1}^{\tau} \left(\lambda_1 e^{-\lambda_1 Y_i}\right) \left(\frac{\beta^\alpha}{\Gamma(\alpha)} \lambda_1^{\alpha-1} e^{-\beta\lambda_1}\right) d\lambda_1$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \int \lambda_1^{\alpha+\tau-1} e^{-\lambda_1(\beta+S_1)} \lambda_1, \quad S_1 = \sum_{i=1}^{t} Y_i$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+\tau)}{(\beta+S_1)^{\alpha+\tau}}$$

# How to Detect Change Points

Combining with the prior $p(\tau)$ gives the posterior distribution for the change point $\tau$:

$$p(\tau|Y) = \frac{p(\tau)p(Y|\tau)}{p(Y)} = \frac{\frac{1}{n-1}\left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right)^2 \frac{\Gamma(\alpha+\tau)}{(\beta+S_1)^{\alpha+\tau}} \frac{\Gamma(\alpha+n-\tau)}{(\beta+S_2)^{\alpha+n-\tau}}}{p(Y)}$$

Since there are only a finite number of values $\tau$ can have (1,2,. . . ,n-1) we can avoid calculating $p(Y)$ and instead normalise the posterior "by hand" so that it sums to 1. Note we can also ignore the $1/(n-1)$ term in the prior for the same reason – it does not depend on $\tau$

After normalising, we have the posterior distribution for the change point $p(\tau|Y)$ which represents our beliefs about its location after seeing the data.

# This Week

This week we will focus on the following three questions:

1. How do we use the information we learned about the change point to reason about the probability of extreme values occurring?
2. How do we decide whether there actually is a change point at all?
3. What do we do if there is more than one change point?

# Predicting the Occurrence of Large Values

Suppose we have observed one observation a day for $n$ days $Y_1, \ldots, Y_n$ (e.g. the number of terrorist attacks per day, or the amount of financial loss, etc)

We want to reason about the the probability of seeing large values in the future, i.e. we want the predictive distribution $p(\tilde{Y}|Y_1, \ldots, Y_n)$.

As we have seen before, we get this by using the theorem of total probability and integrating over the posterior distribution of any parameters we don't know. I.e. if $\theta$ is a vector of unknown parameters then:

$$p(\tilde{Y}|Y_1, \ldots, Y_n) = \int p(\tilde{Y}|\theta)p(\theta|Y_1, \ldots, Y_n)d\theta$$

# Predicting the Occurrence of Large Values

Lets refresh our memory about how to do this in the Exponential case where there is **no change point**. We have:

$$Y_i \sim Exponential(\lambda), \quad \text{for all } i$$

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

We learn $\lambda$ based on $Y_1, \ldots, Y_n$ and our prior knowledge (represented as a Gamma prior) Recall that the posterior distribution $p(\lambda | Y_1, \ldots, Y_n)$ is:

$$p(\lambda | Y_1, \ldots, Y_n) = Gamma(\alpha + n, \beta + \sum_{i=1}^{n} Y_i) = Gamma(\tilde{\alpha}, \tilde{\beta})$$

# Predicting the Occurrence of Large Values

$$p(\tilde{Y}|Y_1,\ldots,Y_n) = \int p(\tilde{Y}|\theta)p(\theta|Y_1,\ldots,Y_n)d\theta =$$

$$= \int \lambda e^{-\lambda\tilde{Y}} \frac{\tilde{\beta}^{\tilde{\alpha}}}{\Gamma(\tilde{\alpha})}\lambda^{\tilde{\alpha}-1}e^{-\tilde{\beta}\lambda}d\lambda =$$

$$= \frac{\tilde{\beta}^{\tilde{\alpha}}}{\Gamma(\tilde{\alpha})}\frac{\Gamma(\tilde{\alpha}+1)}{(\tilde{\beta}+\tilde{Y})^{\tilde{\alpha}+1}} = \quad \text{(using } \Gamma(z+1)/\Gamma(z) = z\text{)}$$

$$= \alpha\frac{\tilde{\beta}^{\tilde{\alpha}}}{(\tilde{\beta}+\tilde{Y})^{\tilde{\alpha}+1}} = \frac{\tilde{\alpha}}{\tilde{\beta}}\left(\frac{\tilde{\beta}+\tilde{Y}}{\tilde{\beta}}\right)^{-\alpha-1} =$$

$$= \frac{\tilde{\alpha}}{\tilde{\beta}}\left(1+\frac{\tilde{Y}}{\tilde{\beta}}\right)^{-\tilde{\alpha}-1}$$

# Predicting the Occurrence of Large Valuesg

if we want to find $p(\tilde{Y} > D|Y_1, \ldots, Y_n)$ then we do this by:

$$p(\tilde{Y} > D|Y_1, \ldots, Y_n) = \int_D^{\infty} p(\tilde{Y}|Y_1, \ldots, Y_n) d\tilde{Y}$$

which in the Exponential case is:

$$p(\tilde{Y} > D|Y_1, \ldots, Y_n) = \int_D^{\infty} \frac{\tilde{\alpha}}{\tilde{\beta}} \left(1 + \frac{\tilde{Y}}{\tilde{\beta}}\right)^{-\tilde{\alpha}-1} d\tilde{Y}$$

In cases where we can't do this integral by hand, numerical integration (e.g. Simpson's Rule or quadratures) can be used.

# Predicting the Occurrence of Large Values

Now lets go back to the case where there is a change point.. Suppose we make the following assumptions:

- We know the true value of $\tau$
- No more change points will occur in future

Then:

$$p(\tilde{Y}|Y_1, \ldots, Y_n, \tau) = p(\tilde{Y}|Y_{\tau+1}, \ldots, Y_n)$$

i.e we simply ignore the observations before the change point since they are no longer relevant. – things have changed, so only the observations in the new segment are relevant for predicting the future

# Predicting the Occurrence of Large Values

In the Exponential case we hence have:

$$p(\tilde{Y}|Y_1,\ldots,Y_n,\tau) = \frac{\tilde{\alpha}}{\tilde{\beta}}\left(1+\frac{\tilde{Y}}{\tilde{\beta}}\right)^{-\tilde{\alpha}-1}$$

where:

$$\tilde{\alpha} = \alpha + n - \tau$$

$$\tilde{\beta} = \beta + \sum_{i=\tau+1}^{n} Y_i$$

# Predicting the Occurrence of Large Values

Of course, in practice we do not know where $\tau$ is. But we saw earlier how to compute its posterior distribution $p(\tau|Y_1, \ldots, Y_n)$.

So, we just do what we always do: average over the posterior:

$$p(\tilde{Y}|Y_1, \ldots, Y_n) = \int p(\tilde{Y}|Y_1, \ldots, Y_n, , \tau)p(\tau|Y_1, \ldots, Y_n)d\tau$$

Since $\tau$ can only take finitely many values, this can be written as:

$$p(\tilde{Y}|Y_1, \ldots, Y_n) = \sum_{\tau=1}^{n-1} p(\tilde{Y}|Y_1, \ldots, Y_n, \tau)p(\tau|Y_1, \ldots, Y_n) =$$

$$= \sum_{\tau=1}^{n-1} p(\tilde{Y}|Y_{\tau+1}, \ldots, Y_n)p(\tau|Y_1, \ldots, Y_n)$$

i.e. we are taking the weighted average (by the posterior) of the $p(\tilde{Y}|Y_1, \ldots, Y_n, \tau)$ term we computed on the previous slides

# Example

Suppose we have a sequence of 10 observations $Y_1, \ldots, Y_{10}$ from the Exponential distribution, with a single change point

Suppose that when we compute $p(\tau | Y_1, \ldots, Y_{10})$ above we find:

$$p(\tau = 6 | Y_1, \ldots, Y_{10}) = 0.1$$

$$p(\tau = 7 | Y_1, \ldots, Y_{10}) = 0.7$$

$$p(\tau = 8 | Y_1, \ldots, Y_{10}) = 0.2$$

and the probability that $p(\tau = k) | Y_1, \ldots, Y_{10})$ for any other value of $k$ is 0.

If we then want to find the predictive distribution for the next unseen observation ($Y_{11}$) we have

# Example

$$p(\tilde{Y}|Y_1,\ldots,Y_n) = \sum_{\tau=1}^{n-1} p(\tilde{Y}|Y_{\tau+1},\ldots,Y_n,\tau)p(\tau|Y_1,\ldots,Y_n) =$$

$$= 0.1 \times p(\tilde{Y}|Y_7, Y_8, Y_9, Y_{10}) + 0.7 \times p(\tilde{Y}|Y_8, Y_9, Y_{10}) + 0.2 \times p(\tilde{Y}|Y_9, Y_{10})$$

where

$$p(\tilde{Y}|Y_j,\ldots,Y_n) = \frac{\tilde{\alpha}}{\tilde{\beta}}\left(1+\frac{\tilde{Y}}{\tilde{\beta}}\right)^{-\tilde{\alpha}-1}$$

with

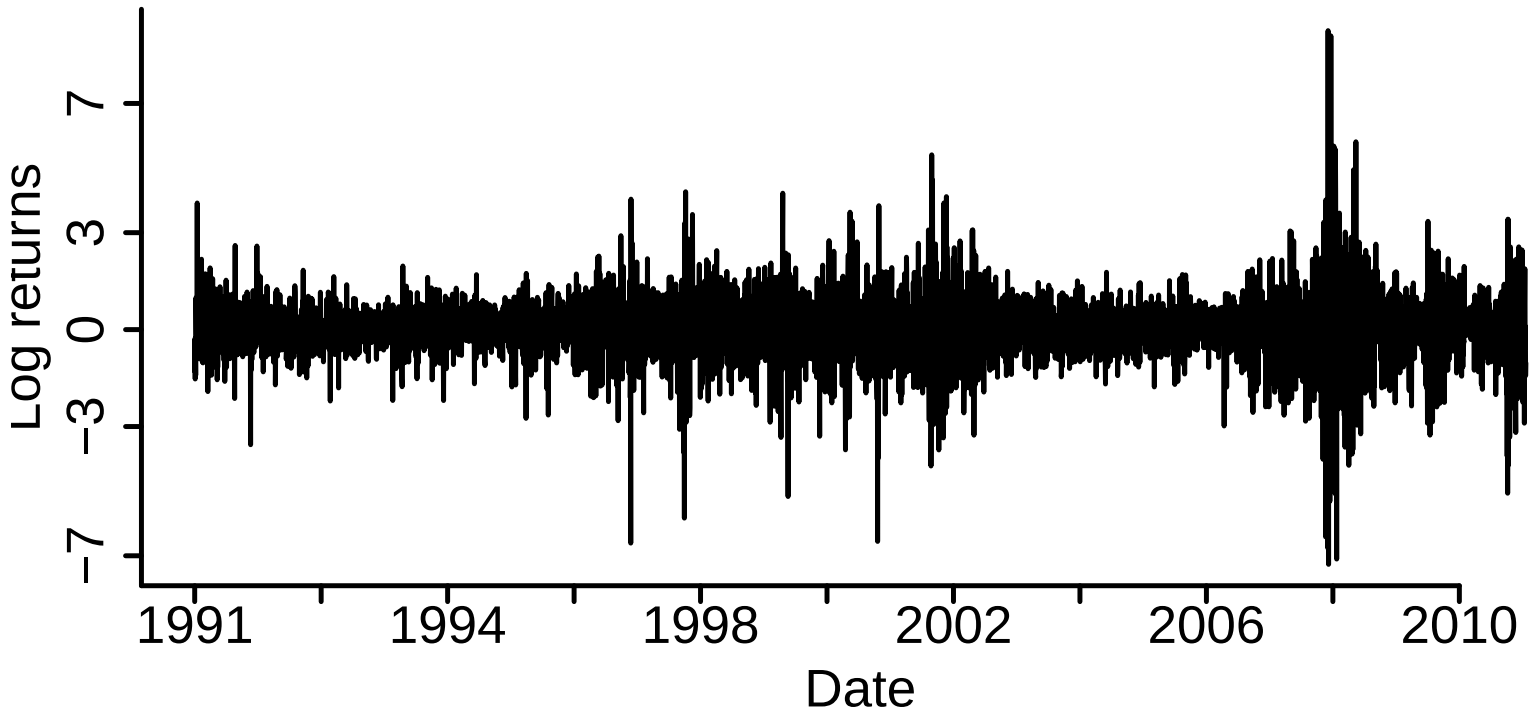
$$\tilde{\alpha} = \alpha + 10 - j + 1, \quad \tilde{\beta} = \beta + \sum_{i=j}^{10} Y_i$$

# Multiple Change Points

So far, we have considered only the case where the data contains at most a single change point.

This is the simplest version of the change point problem. However in practice the series may contain multiple change points Remember the Dow Jones data:

# Dow Jones



# Multiple Change Points

Suppose there are $k$ change points dividing the series into $k+1$ segments. Let $\theta_j$ denote the (unknown) value of $\theta$ in the $j^{th}$ segment. Se the change point model is:

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

We need to estimate the change point locations $\tau_1, \dots, \tau_k$.

Conceptually, everything we here in this case is **exactly the same** as in the single change point case

# Multiple Change Points

We estimate the change point locations using Bayes Theorem as always:

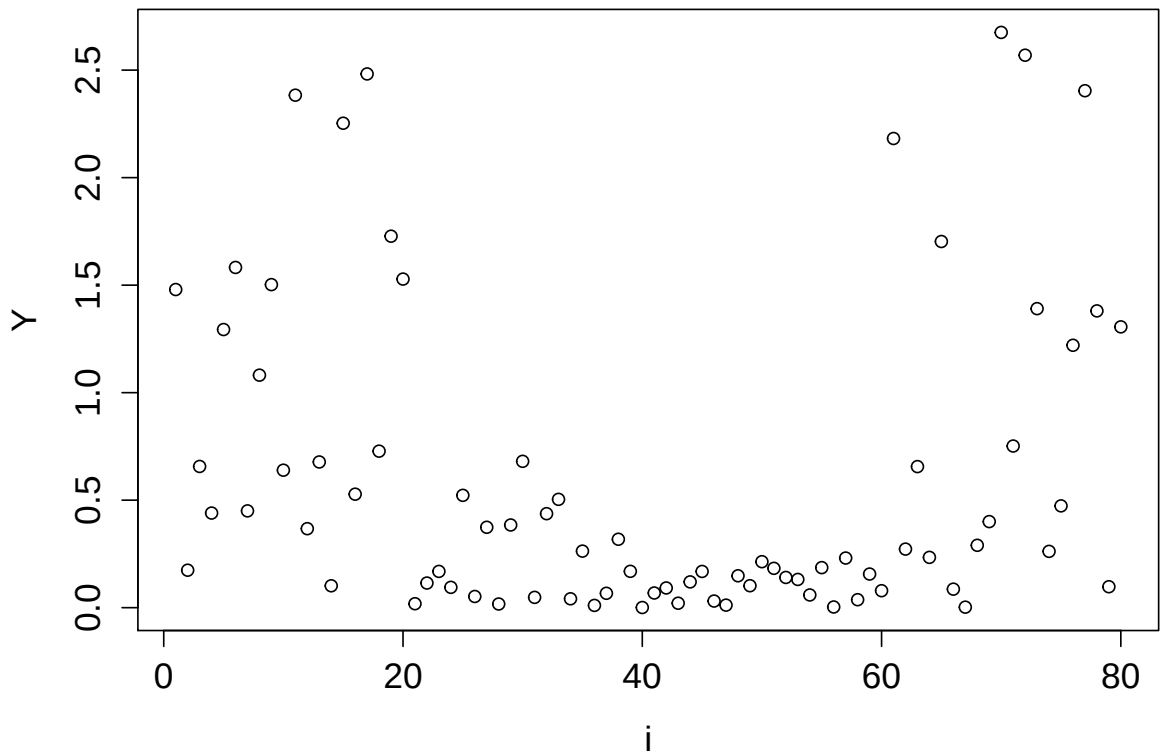
$$p(\tau_1, \ldots, \tau_k | Y_1, \ldots, Y_n) = \frac{p(Y_1, \ldots, Y_n | \tau_1, \ldots, \tau_k) p(\tau_1, \ldots, \tau_k)}{p(Y_1, \ldots, Y_n)}$$

# Multiple Change Points - Exponential Distribution

So for example in the case where the observations have an Exponential distribution:

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

# Multiple Change Points - Exponential Distribtuion



# Multiple Change Points - Prior

As in the single change point case, we usually assume the change points are uniformly distributed over $1, 2, \ldots, n$.

Since the prior distribution is uniform, this means that it does not depend on the $\tau_1, \ldots, \tau_k$ values (as before

As such, it can be ignored when it comes to computing the posterior since it will be rolled into the normalising constant when we normalise the posterior

# Multiple Change Points - Likelihood

Recall that in the single change point case if the $\lambda$ parameters were known, the likelihood was:

$$p(Y|\tau) = \prod_{i=1}^{\tau} p(Y_i|\theta_1) \prod_{i=\tau+1}^{n} p(Y_i|\theta_2)$$

In the multiple change point case this becomes:

$$p(Y|\tau_1, \ldots, \tau_k) = \prod_{i=1}^{\tau_1} p(Y_i|\theta_1) \prod_{i=\tau_1+1}^{\tau_2} p(Y_i|\theta_2) \prod_{i=\tau_2+1}^{\tau_3} p(Y_i|\theta_3) \ldots \prod_{i=\tau_k+1}^{n} p(Y_i|\theta_{k+1})$$

i.e. we are still breaking the likelihood up into segments as before, we just have $k+1$ segments now rather than two

# Multiple Change Points - Likelihood

Similarly in the single change point case when the parameters $\theta$ were unknown we have to integrate over them:

$$p(Y|\tau) = \int \prod_{i=1}^{\tau} p(Y_i|\theta_1)p(\theta_1)d\theta_1 \int \prod_{i=\tau+1}^{n} p(Y_i|\theta_2)p(\theta_2)d\theta_2$$

In the $k$ change point case this becomes:

$$p(Y|\tau_1,\ldots,\tau_k) = \left(\int \prod_{i=1}^{\tau_1} p(Y_i|\theta_1)p(\theta_1)d\theta_1\right) \times \left(\int \prod_{i=\tau_1+1}^{\tau_2} p(Y_i|\theta_2)p(\theta_2)\right) \times$$

$$\times \left(\int \prod_{i=\tau_2+1}^{\tau_3} p(Y_i|\theta_3)p(\theta_3)d\theta_3\right) \times \ldots \times \left(\int \prod_{i=\tau_k+1}^{n} p(Y_i|\theta_{k+1})p(\theta_{k+1})d\theta_{k+1}\right)$$

# Multiple Change Points - Example

Recall that with the Exponential sequence in the single change point case when we did these integrals, the likelihood of the observations to the left of the change point was

$$p(Y_1, \ldots, Y_\tau|\tau) = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+\tau)}{(\beta+S_1)^{\alpha+\tau}}, \quad S_1 = \sum_{i=1}^{\tau} Y_i$$

In the multiple change point case, an identical argument can be used to do the integrals above. If we focus on (e.g.) the second segment (between $\tau_1$ and $\tau_2$) we get:

$$p(Y_{\tau_1+1}, \ldots, Y_{\tau_2}|\tau) = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+\tau_2-\tau_1)}{(\beta+S_2)^{\alpha+\tau_2-\tau_1}}, \quad S_2 = \sum_{i=\tau_1+1}^{\tau_2} Y_i$$

# Multiple Change Points - Example

So we have:

$$p(Y_1, \ldots, Y_n|\tau_1, \ldots, \tau_k) = \left[\frac{\beta^\alpha}{\Gamma(\alpha)}\right]^{k+1} \times \frac{\Gamma(\alpha+\tau_1)}{(\beta+S_1)^{\alpha+\tau_1}} \times \frac{\Gamma(\alpha+\tau_2-\tau_1)}{(\beta+S_2)^{\alpha+\tau_2-\tau_1}}, \times$$

$$\times \frac{\Gamma(\alpha+\tau_3-\tau_2)}{(\beta+S_3)^{\alpha+\tau_3-\tau_2}} \times \ldots \times \frac{\Gamma(\alpha+n-\tau_k)}{(\beta+S_{k+1})^{\alpha+n-\tau_k}},$$

# Multiple Change Points - Example

Remember the posterior is:

$$p(\tau_1, \ldots, \tau_k | Y_1, \ldots, Y_n) = \frac{p(Y_1, \ldots, Y_n | \tau_1, \ldots, \tau_k) p(\tau_1, \ldots, \tau_k)}{p(Y_1, \ldots, Y_n)}$$

We can ignore the $p(Y)$ and $p(\tau_1, \ldots, \tau_k)$ terms since they don't depend on the $\tau_i's$ (recall the latter is constant since the prior was uniform)

So, the unnormalised posterior is equal to the likelihood on the previous slide

# Multiple Change Points

As before, we normalise the posterior by evaluating it at each possibly combination of the change points, and dividing through by the sum

Since this is too fiddly to do by hand, it really requires a computer. The exercise sheets ask you to implement this in R, and the final workshop In two weeks time) explore using this on real data