Decision and Risk
Lecture 6: Change Points

Gordon J. Ross

Where Are We...

So far, we have focused on asking questions such as ”what is the probability of extreme events happening?” in a number of application areas such as financial loss, and terrorist attacks.

Our typical situation is that we have observed the data $Y = Y_1, \ldots, Y_n$ and want to make predictions about future values $\tilde{Y}$ which are assumed to have the same distribution as Y

We have a probability model $p(Y|\theta)$ where the parameter θ is unknown, and a prior $p(\theta)$ which represents our prior beliefs about θ

The Two Key Equations

Our two key equations are:

1. The posterior distribution for θ:

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

2. The predictive distribution for $\tilde{Y}$ given the historical data Y:

$$p(\tilde{Y}|Y) = \int p(\tilde{Y}|\theta)p(\theta|Y)d\theta$$

The form for the predictive distribution comes from the Theorem of Total Probability. A special case we have used often is:

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

Real life is often more complicated...

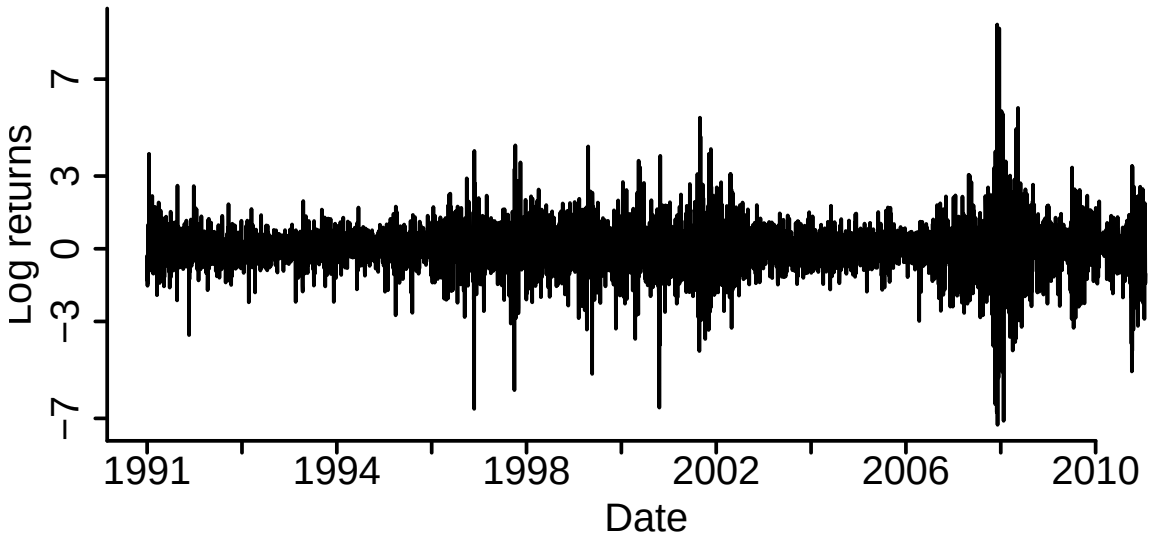
So far we have discussed situations where the data Y_i are independent and identically distributed so that $Y_i \sim p(Y|\theta)$

However is often not the case in practice

Consider the Value-at-Risk problem where we wish to find the probability of losing more than \$D in a single day. We specify a probability model for the returns $Y_1, \ldots, Y_n$, for example that they are $N(0, \theta)$, and use historical data to estimate θ

The problem is that in practice, real financial returns tends to look like this (next slide)

Some Real Financial Data



This is the daily log-returns ($Y_i = \log(P_i/P_{i-1})$, where P_i is the price on day i) of the Dow Jones stock index, which is one of the most commonly traded financial assets. The pattern we see here with the constantly changing variance occurs in pretty much all real financial data. In the case the returns $Y_1, \ldots, Y_n$ are clearly not identically distributed, due to the changing variance.

The Problem

With data such as this, we cannot just naively estimate the unknown parameters θ as we have been doing. Suppose we have 10 years of observed data. If a change in the distribution $p(Y|\theta)$ occurred last year (e.g. if the variance increased) then the observations from before that point should not be used to estimate the current value of θ or they will give us a misleading estimate.

In other words we need to estimate where the most recent change point has occurred and only look at the data after that point. We will then use only this data to compute the quantities we are interested in (e.g. the predictive distribution, and probably of extreme events occurring)

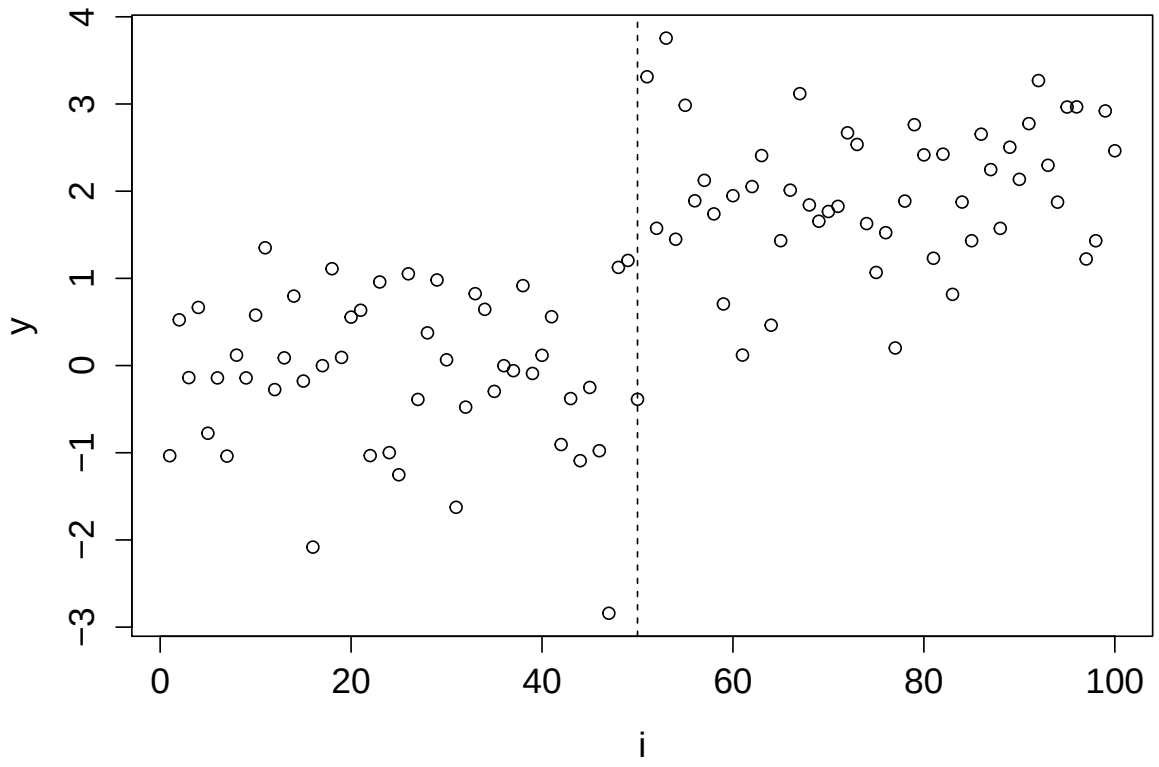
Change Point Detection

The setting is hence as follows: we have a sequence of time ordered data $Y_1, \ldots, Y_n$. Suppose we believe there is a single change point but do know when it occurs. Denote this unknown change point by τ. Before the change point, the unknown parameter θ has value θ_1, while after it changes to θ_2. The distribution of the data is then:

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

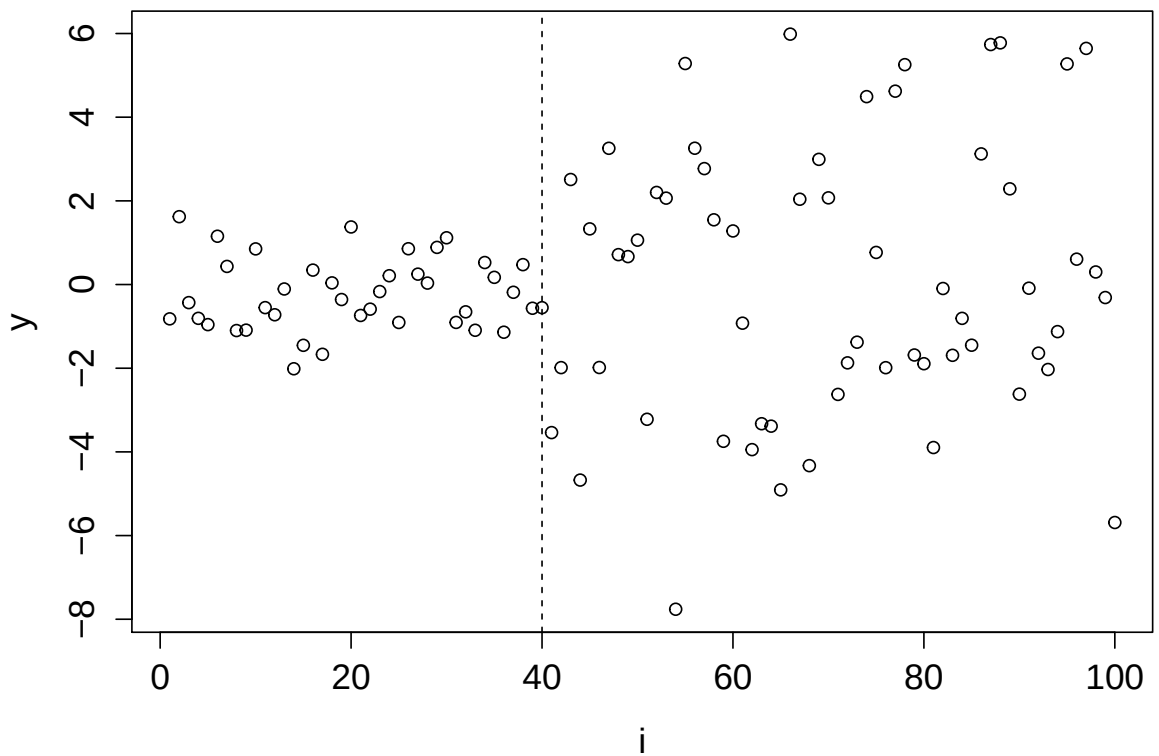
Example 1

$Y_1, \ldots, Y_{100} \sim N(\theta, 1)$ where the change point is at $\tau = 50$ and $\theta = 0$ before this, and $\theta = 1$ after:



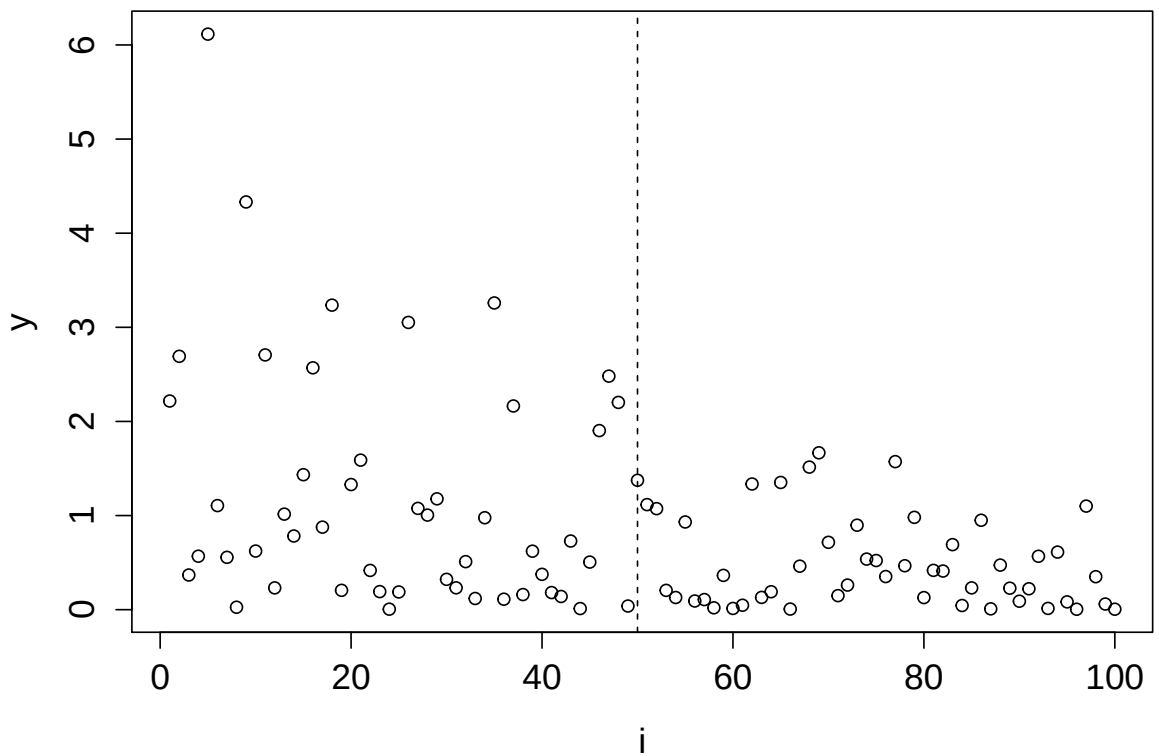
Example 2

$Y_1, \ldots, Y_{100} \sim N(0, \theta)$ where the change point is at $\tau = 40$ and $\theta = 1$ before this, and $\theta = 3^2$ after:



Example 3

$Y_1, \ldots, Y_{100} \sim \textit{Exponential}(\lambda)$ where the change point is at $\tau = 50$ and $\theta = 1$ before this, and $\theta = 3$ after:



How to Detect Change Points

In these examples the location of the change point is fairly obvious to the eye since I have deliberately chosen large changes. However this will not be the case when the magnitude of change is smaller.

For the rest of this lecture we assume that there is a **single** change point τ and need to estimate its location (we will discuss multiple change points in future).

We will work out the mathematics for the third example above (Exponential sequence) in detail, but identical techniques apply to essentially any other type of distribution (Normal, Binomial, etc)

Our Setting

So our setting is that we have observed $Y_1, \ldots, Y_n$ where:

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

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

How to Detect Change Points

The nice thing about Bayesian statistics is that all problems are solved in essentially the same way.

To estimate τ, we do what we always do: start with a prior $p(\tau)$ then combine it with the likelihood $p(\tau|Y)$ to get the posterior distribution $p(\tau|Y)$ which represents all our knowledge about the change point after seeing the data. As always, this is done using Bayes Theorem:

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

The Prior

Often we do not have strong beliefs about where the change point τ occurs, so we use a non-informative prior which treats every possible location as equally likely. There are n observations and the last one cannot be a change point since we require at least one observation on both sides of τ. As such, there are $n-1$ possible locations, so the non-informative prior is just a discrete uniform distribution on $1, \ldots, n-1$:

$$p(\tau = k) = \frac{1}{n-1}, \quad k = 1, \ldots, n-1$$

The Likelihood

Next we consider the likelihood $p(Y|\tau)$. This depends on the unknown parameters λ_1, λ_2. Suppose these were known. Then, the likelihood would simply be:

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

How to Detect Change Points

However, λ_1, λ_2 are not known. As always in Bayesian statistics, we require a prior distribution for them. Suppose we use the usual conjugate Gamma(α, β) prior for some choice of α and β.

Consider the first segment to the left of the change point, i.e. observations $Y_1, Y_2, \ldots, Y_\tau$ only. From the last slide, if λ_1 was known the likelihood would be

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

But since λ_1 is unknown we instead use the theorem of total probability and integrate over the prior:

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

How to Detect Change Points

In the Exponential-Gamma case, we can do this integral using the same trick we always use:

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

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

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

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

How to Detect Change Points

It is easy to show (do this yourself) using the same argument that the likelihood for the observations $Y_{\tau+1}, Y_{\tau_2}, \ldots, \tau_n$ to the right of the change point is:

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

So combining this together with the expression in the last slide gives the full likelihood:

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

How to Detect Change Points

Finally, combining with the prior $p(\tau)$ gives the posterior distribution for the change point τ:

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

This expression obviously looks horrible but it is easy to evaluate in a language such as R (see exercise sheet). The important thing is to understand where the various parts come from, particularly how to find the $p(Y|\tau)$ expression

How to Detect Change Points

What do we do about the $p(Y)$ part on the denominator? It turns out we actually do not need to compute it. Why is this?

Well, think about what the expression $p(\tau|Y)$ really means. It means that if we plug in $\tau = k$ into the formula on the previous page $p(\tau = k|Y)$, we get the posterior probability that $\tau = k$

There are only $n-1$ possible locations for the change point τ, i.e. $p(\tau|Y)$ is only non-zero at the values $\tau = 1, 2, \ldots, n-1$

The part on the bottom $p(Y)$ does not depend on τ - it will be the same regardless of which value of τ we plug in. It is hence just a constant which scales the posterior distribution so that it sums to 1 (as all probability distributions must)

How to Detect Change Points

Since in this case there are only a finite number of values for τ (1,2,. . . ,n-1), we can instead just evaluate the **numerator** of $p(\tau|Y)$ at these values and ignore the $p(Y)$ term on the denominator since it does not depend on τ.

If we do this, the resulting posterior will not sum to 1. But we can rescale it to sum to 1 by dividing through by the sum of all the numerator values.

Remember in general that if we have any sequence of numbers $Z_1, Z_2, \ldots, Z_m$ then we can rescale these to sum to 1 by dividing each one by their sum. I.e. if we define:

$$\tilde{Z}_1 = \frac{Z_1}{\sum_{i=1}^{m} Z_i}, \quad \tilde{Z}_2 = \frac{Z_2}{\sum_{i=1}^{m} Z_i}, \ldots, \tilde{Z}_m = \frac{Z_m}{\sum_{i=1}^{m} Z_i},$$

Then $\sum_{i=1}^{m} \tilde{Z}_i = 1$.

Example

That might be hard to understand. Lets work through an example to make it clearer.

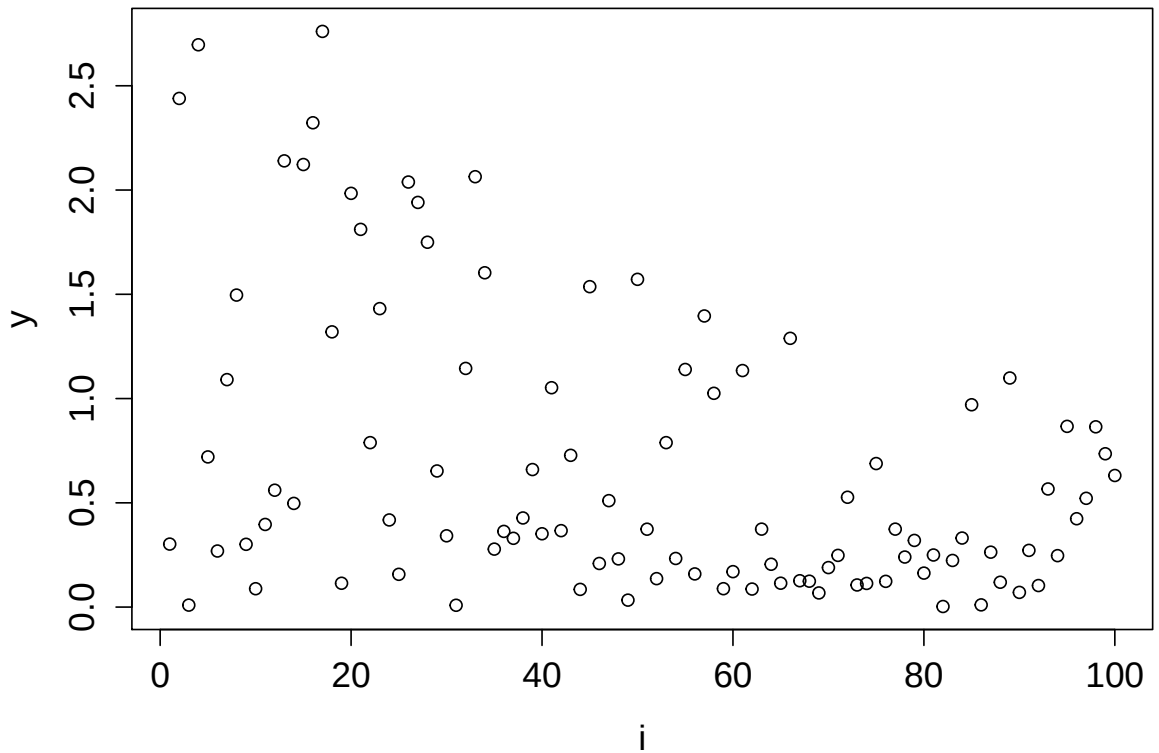
Assume we have 100 observations $Y_1, \ldots, Y_{100}$. The true (and unknown) parameters are $\tau = 50, \lambda_1 = 1, \lambda_2 = 3$.

We use the previous uniform prior on τ, and give λ_1 and λ_2 a Gamma(1,1) prior

(Note: in this week's exercises i am going to ask you to reproduce the below analysis yourself in R)

Example

The sequence looks like this:



Example

There are 99 possible locations for the unknown τ, corresponding to $\tau \in \{1, 2, \ldots, 99\}$.

So, we evaluate the numerator of $p(\tau|Y)$ using R at each of these 99 values of τ (remember we do not need to evaluate $p(Y)$ on the denominator). I.e we are evaluating:

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

for $k = 1, 2, \ldots, 99$ where

$$S_1 = \sum_{i=1}^{k} Y_i, \quad S_2 = \sum_{i=k+1}^{n} Y_i$$

Note: I have used the tilde symbol here to denote that this is **not** the true posterior since it will not sum to 1

Example

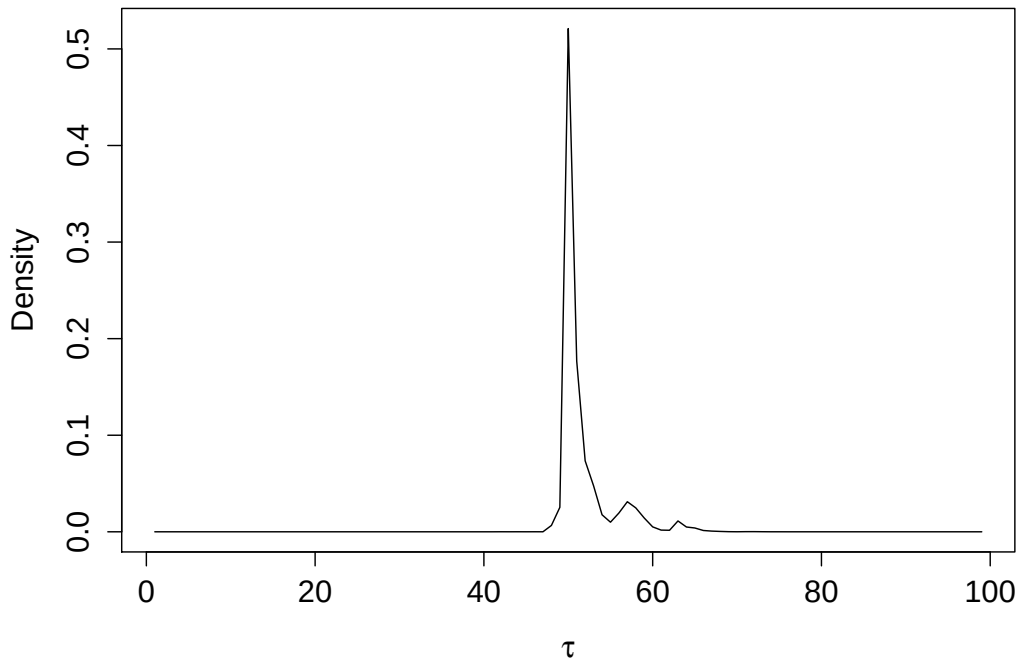
Next, we normalise these values to sum to 1. So, for each k we have:

$$p(\tau = k|Y) = \frac{\tilde{p}(\tau = k|Y)}{\sum_{i=1}^{99} \tilde{p}(\tau = i|Y)}$$

This gives us the true posterior, which is now a properly normalised probability distribution that sums to 1

Example

We can now plot the posterior distribution $p(\tau|Y)$



(note I have smoothed this in the plot to make it clearer but it should really be a histogram: remember that the posterior is only defined on integer values $1, 2, 3, \ldots, n-1$)

Example

So our posterior beliefs are strongly peaked at the change point occurring at around $\tau = 50$. However there is some uncertainty. The actual values are (you will compute these on the exercise sheet)

$$p(\tau = 49|Y) = 0.02$$

$$p(\tau = 50|Y) = 0.52$$

$$p(\tau = 51|Y) = 0.17$$

$$p(\tau = 52|Y) = 0.07$$

And $p(\tau = k|Y) = 0$ for all other values of k.

So we can quantify our posterior belief that the change point occurs at any particular location. We believe there is a 52% chance it occurs at $\tau = 50$, a 17% probability it occurs at $\tau = 51$, and so on.

Point Estimate

As always, we can get a point estimate for τ by taking an appropriate summary of the posterior. Recall that if we seek to minimise the mean square error, we choose the posterior mean. In this case, our point estimate is:

$$\hat{\tau} = \sum_{k=1}^{99} \tau p(\tau|Y)$$

which is $\hat{\tau} = 51.62$ in this case.