Lecture 8: Gradual Drift

Alex Donov

15 March 2019

Last Week...

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

Last Week...

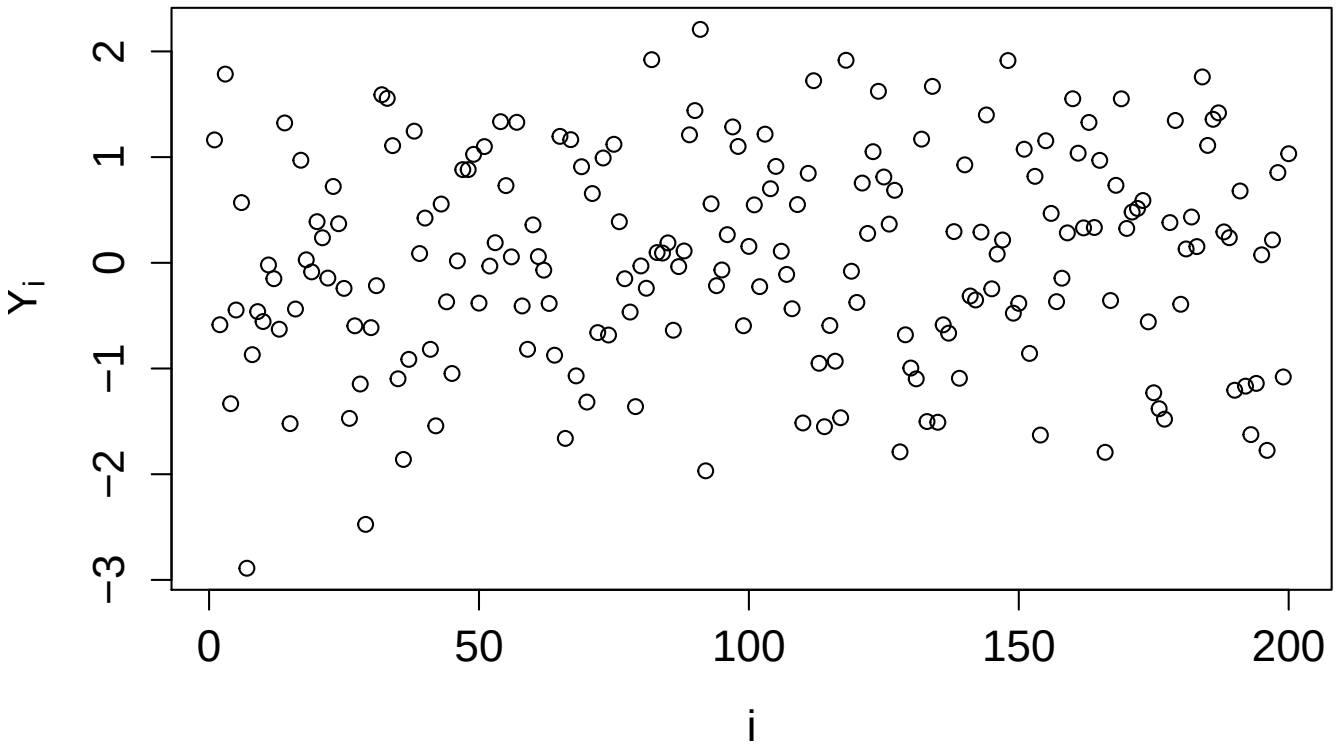
We looked at examples with **change points** where the distribution shifted. In the k change point model we have:

$$Y_t \sim \begin{cases} p(y_t|\theta_1) & \text{if } t \leq \tau_1 \\ p(y_t|\theta_2) & \text{if } \tau_1 < t \leq \tau_2 \\ p(y_t|\theta_3) & \text{if } \tau_2 < t \leq \tau_3 \\ \vdots & \\ p(y_t|\theta_{k+1}) & \text{if } \tau_k < t \leq n \end{cases}$$

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

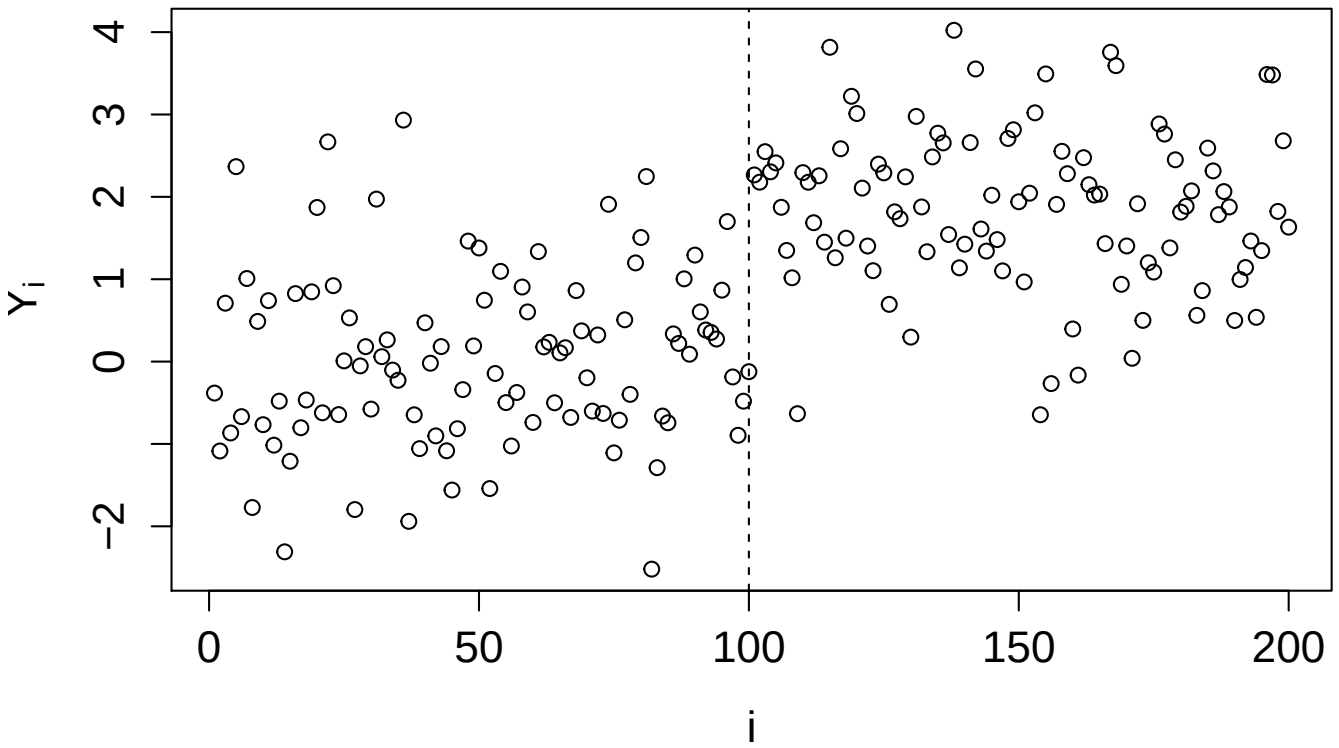
Independent Observations

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

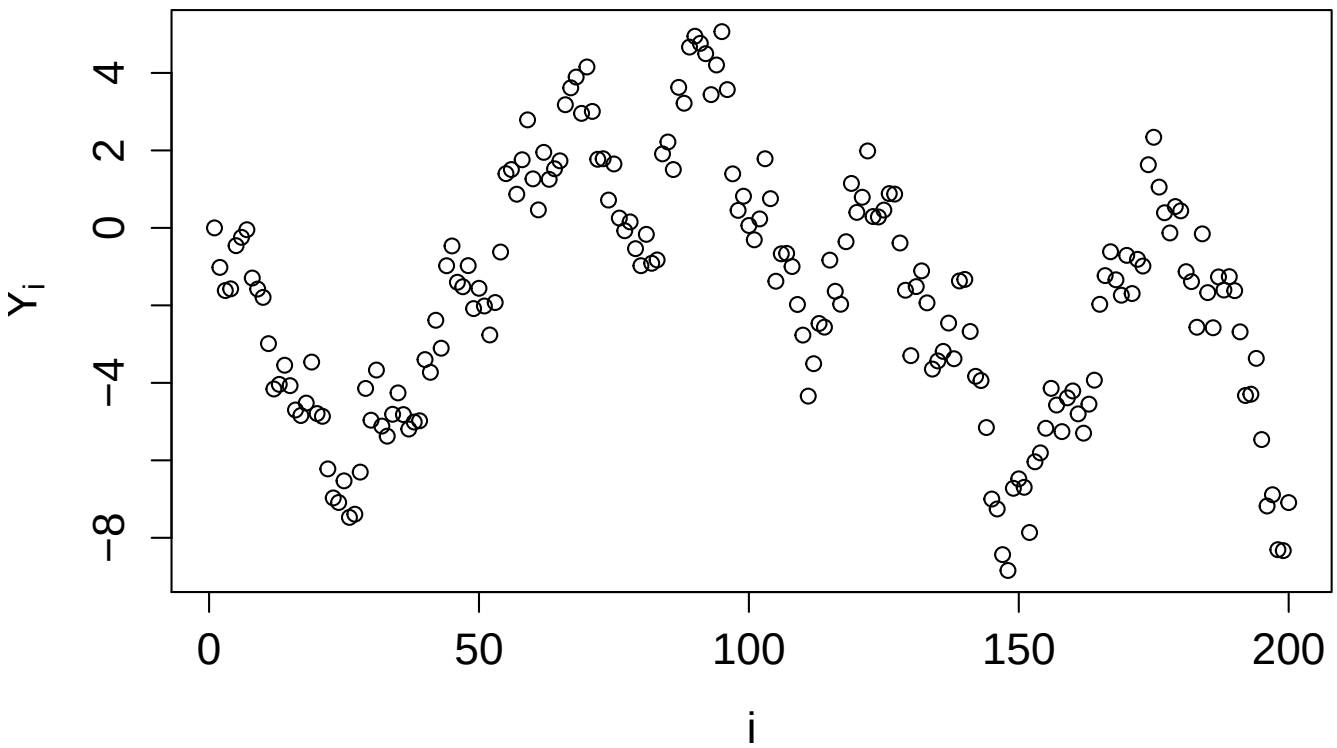


Change Point

$$Y_t = \begin{cases} N(0,1) & \text{if } t \leq 100 \\ N(2,1) & \text{if } t > 100 \end{cases}$$



Gradual Drift



Gradual Drift

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

However the mean seems to be gradually 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:

$$Y_1 = c$$

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

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

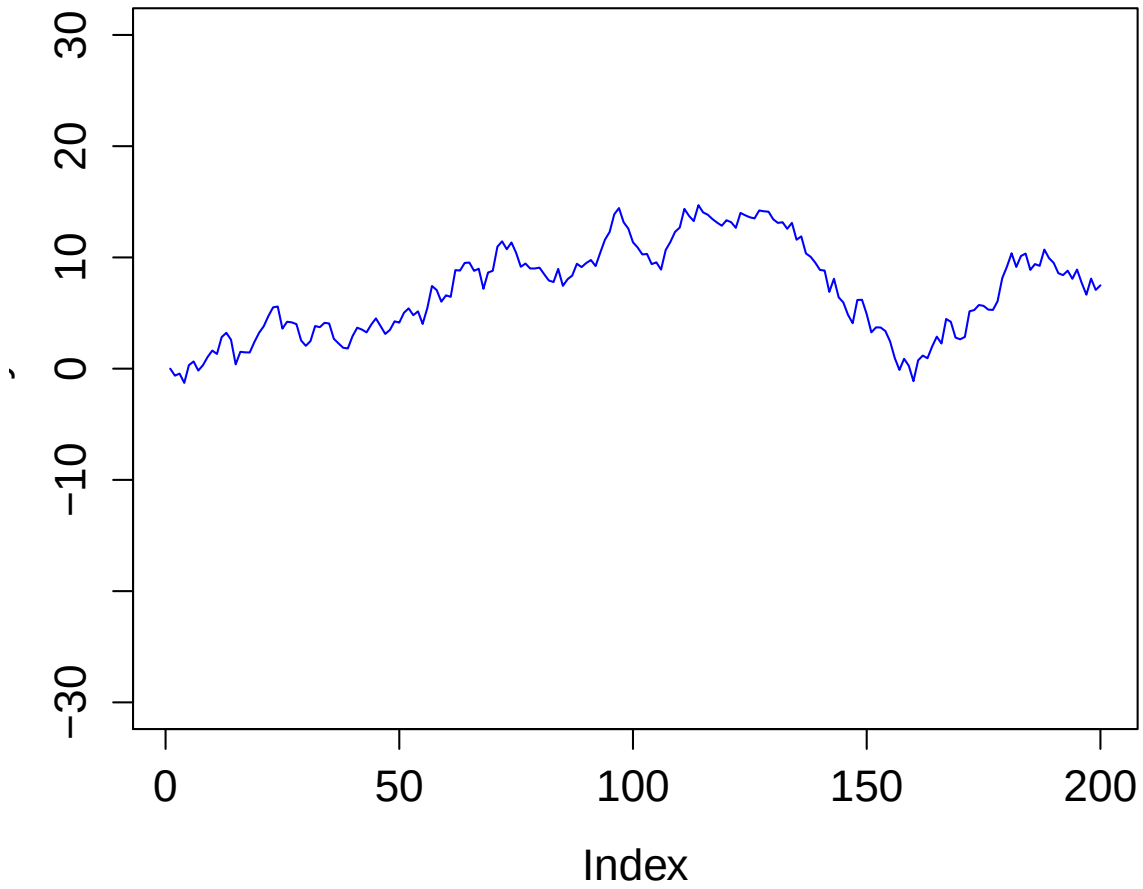
We can simulate a random walk in R very easily. Here is an example where $\sigma^2 = 1$ and $c = 0$

```
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:

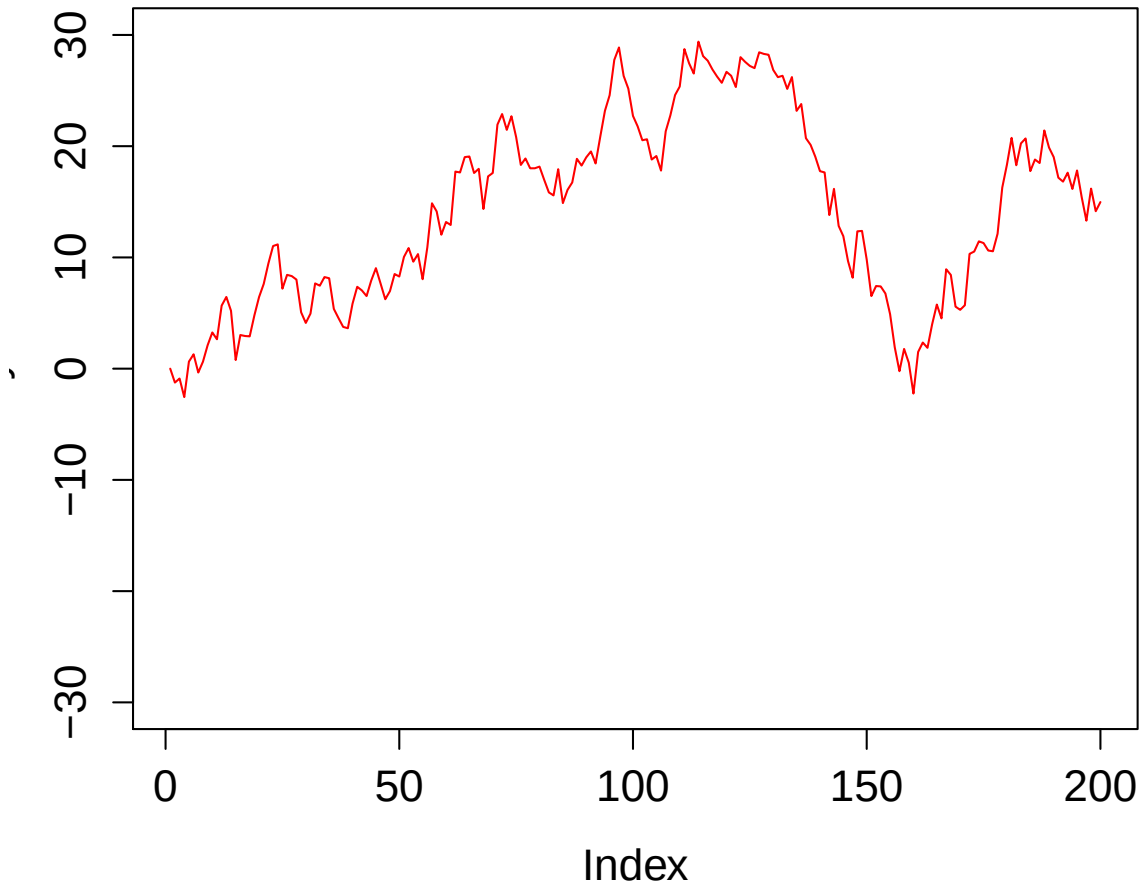
$$Y_1 = c$$

$$Y_t = Y_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2)$$

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 $\sigma = 2$

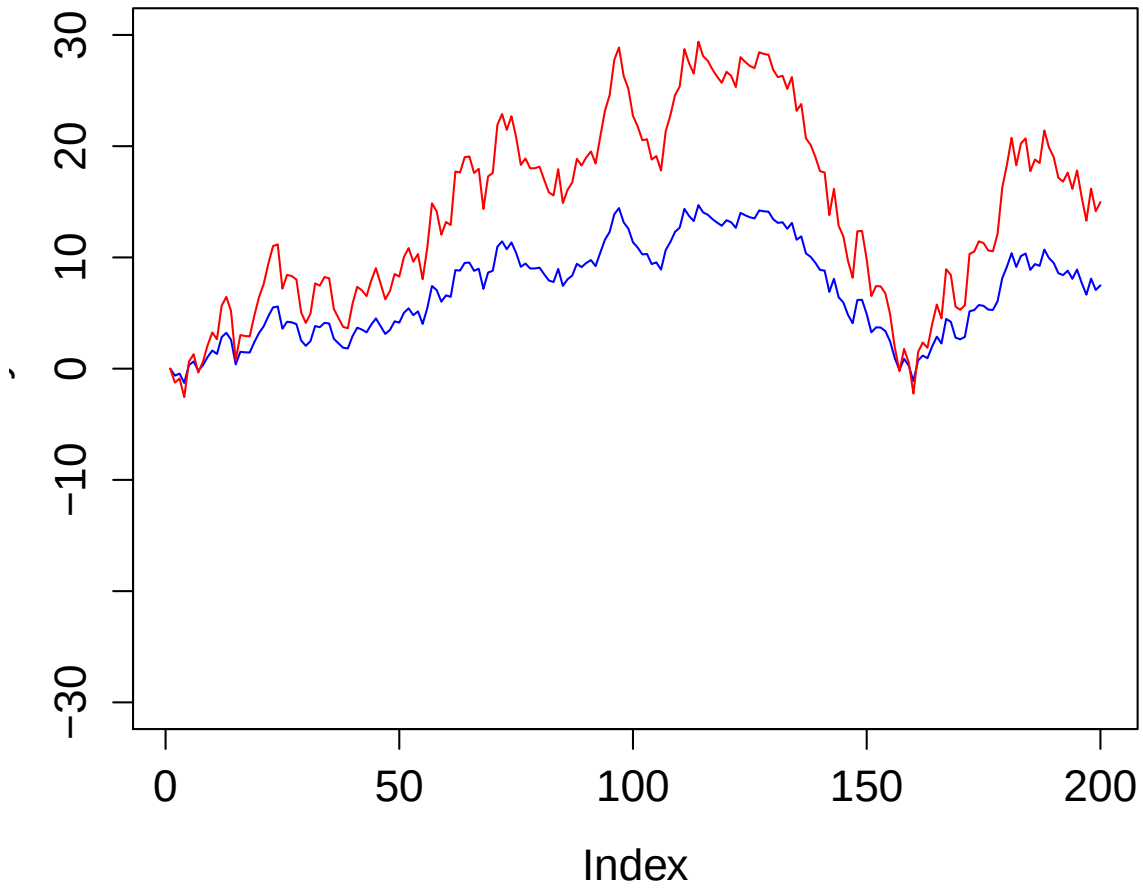
Simulating a Random Walk

Example simulated sequence, $\sigma = 2$



Simulating a Random Walk

Example simulated sequences, $\sigma = 2$, $\sigma = 1$



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_t is its distribution when we condition on the previous values $Y_1, \ldots, Y_{t-1}$, 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}

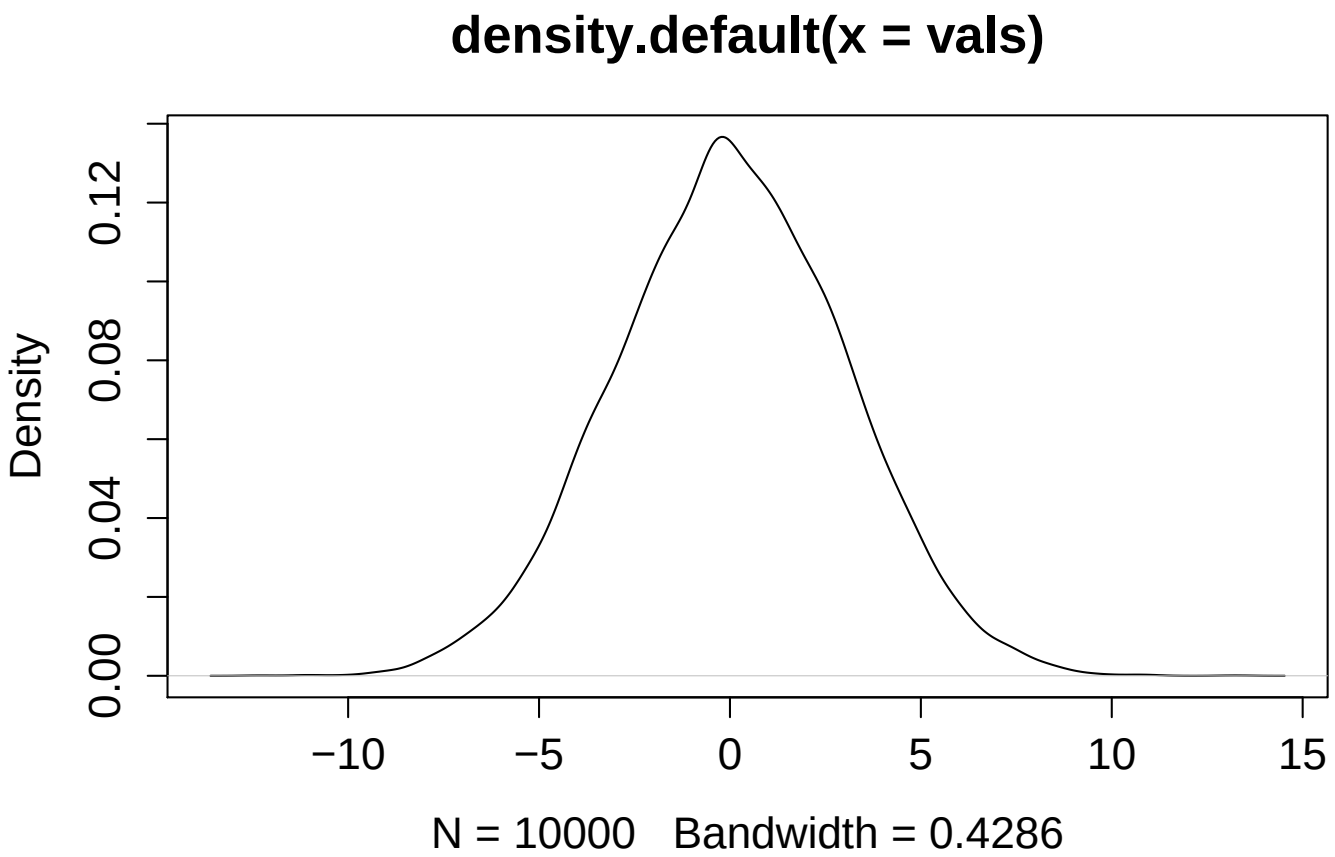
We can do this in R:

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_t:

$$Y_t = Y_{t-1} + \epsilon_t, \quad \epsilon \sim N(0, \sigma^2)$$

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_t we have:

$$Y_t \sim N(0, (t-1)\sigma^2)$$

Conditional Distribution of Y_{10}

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

The **conditional** distribution of Y_t 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, $Y_t = Y_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2)$.

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 Distribution

So in general we have the unconditional distribution:

$$p(Y_t) = N(c, (t-1)\sigma^2)$$

and the conditional distribution

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

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 $Y_1, \ldots, Y_T$ and we want to know the probability that $Y_{t+1} > D$ for some D. As always, we need the predictive distribution of Y_{T+1}

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

$$Y_{T+1} \sim N(Y_T, \sigma^2)$$

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

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

$$Y_t = Y_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2)$$

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

$$Z_t = Y_{t+1} - Y_t$$

Then the Z_t 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,σ^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:

$$p(\sigma^2|y_1, \ldots, y_T) = IG\left(\alpha + \frac{(T-1)}{2}, \beta + \frac{\sum_{i=1}^{T-1} z_t^2}{2}\right)$$

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:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2)$$

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:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2)$$

It is very similar except that the distribution of Y_t depends on Y_{t-2} as well as Y_{t-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

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

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:

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

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.

When the prior is not conjugate, we can use numerical integration techniques such as Simpson's Rule or quadrature.

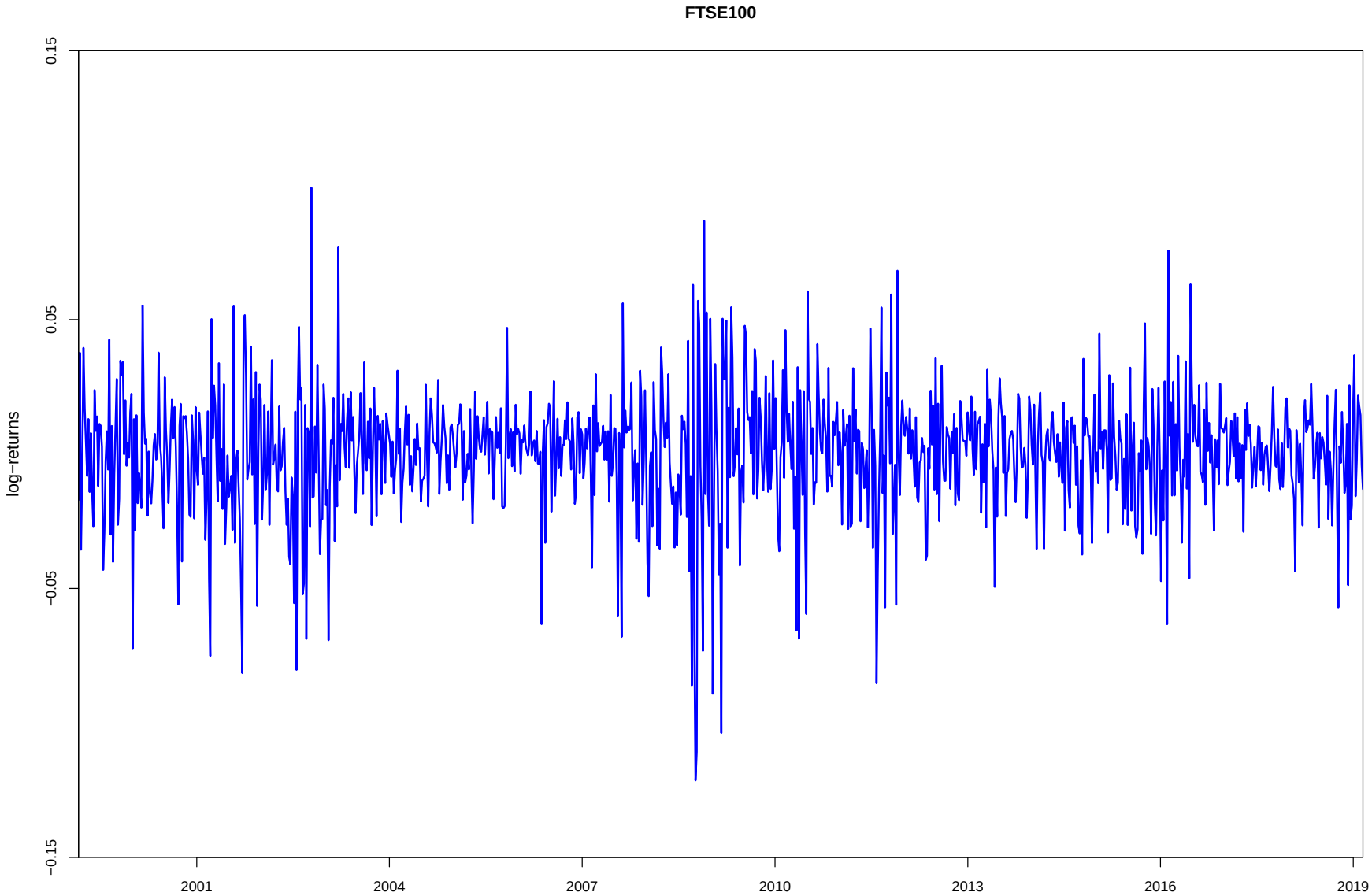
However, the complexity increases when p is large. When everything is conjugate, we can still find the posterior easily, but in the non-conjugate case we require numerical integration techniques which are beyond the scope of this module (such as Markov Chain Monte Carlo [MCMC]).

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 instead it is the **variance** of the returns which changes

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

FTSE100


FTSE100
log−returns
0.15
0.05
−0.05
−0.15
2001
2004
2007
2010
2013
2016
2019

ARCH and GARCH Models

Almost all financial return series exhibit 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.

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

ARCH Models

The $ARCH(p)$ model of Engle (1982) is defined by:

$$y_t = \mu + u_t \quad \text{where} \quad u_t = \sigma_t \epsilon_t \quad \text{and} \quad \epsilon_t \sim N(0,1)$$

$$\sigma_t^2 = \omega + \sum_{p=1}^{P} \alpha_p u_{t-p}^2$$

where σ_t^2 is the **conditional** variance given past information, and ϵ_t are $i.i.d.$ random variables, ω, $\alpha_p > 0$ assures that the conditional variance σ_t^2 is positive.

ARCH stands for "Autoregressive Conditional Heteroskedasticity".

ARCH Models

- The simplest model for time-varying conditional variance is the ARCH(1) model.
- It assumes that the returns Y_t have zero mean, and a time-varying conditional variance:

$$y_t = u_t \quad \text{where} \quad u_t = \sigma_t \epsilon_t \quad \text{and} \quad \epsilon_t \sim N(0,1)$$

$$\sigma_t^2 = \beta_0 + \beta_1 u_{t-1}^2$$

ARCH Models

$$Y_t = \sigma_t \epsilon_t, \quad \epsilon_t \sim N(0,1)$$

- The log return Y_t at time t has 0 mean, and a variance which is time dependent.
- Specifically, the variance is equal to σ_t^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 $N(0,\sigma_t^2)$, where σ_t^2 depends on the previous value Y_{t-1} (hence conditional!)

ARCH Models

$$\sigma_t^2 = \beta_0 + \beta_1 u_{t-1}^2$$

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.

Simulating from an ARCH(1) model

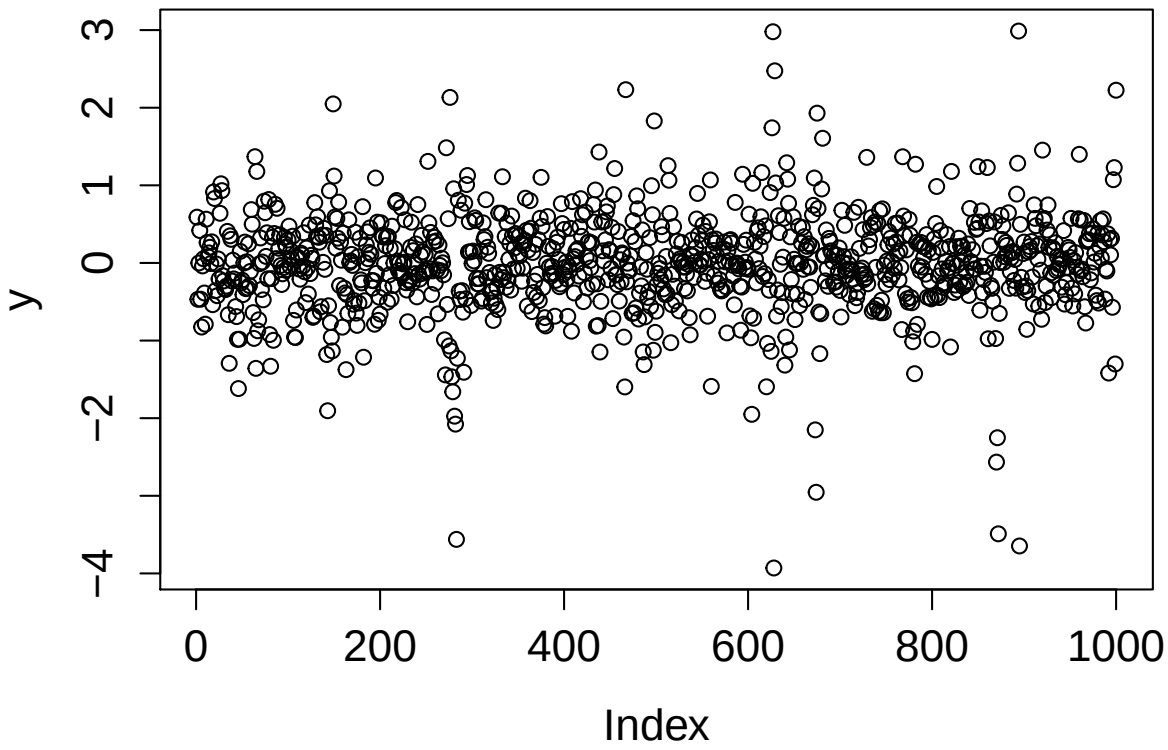
```
beta0 <- 0.1; beta1 <- 0.8
n <- 1000
sigma2s <- numeric(n);
y <- numeric(n)

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 $Y_1, \ldots, Y_T$:



Limitations of ARCH(p) models

ARCH(p) models are rarely used in practice since they have limitations associated with them:

- There is no clear guidance on the choice of the number of lags p of the squared residual.
- The number of lags of the squared error required to capture all of the dependence could be very large, and hence the number of parameters to be estimated. In this case, the resulting model would not be parsimonious.
- Non-negativity constraints might be violated due to the large number of parameters in the model. This is because the more parameters there are in the model, the more likely it is that there will be at least one negative estimated value.

GARCH(p, q) model

To address limitations associated with ARCH model, Bollerslev (1986) and Taylor (1986) independently developed the GARCH model. The $GARCH(p,q)$ model of Bollerslev (1986) is defined by:

$$y_t = \mu + u_t \quad \text{where} \quad u_t = \sigma_t \epsilon_t \quad \text{and} \quad \epsilon_t \sim N(0,1)$$

$$\sigma_t^2 = \omega + \sum_{p=1}^{P} \alpha_p u_{t-p}^2 + \sum_{q=1}^{Q} \beta_q \sigma_{t-q}^2$$

where σ_t is the conditional variance given past information, and ϵ_t are $i.i.d.$ random variables, ω, β_q, $\alpha_p > 0$ assures that the condition variance σ_t^2 is positive and $\sum_{p=1}^{P} \alpha_p + \sum_{q=1}^{Q} \beta_q < 1$ ensures stationarity.

GARCH(1,1) Model

- The simplest model for time-varying conditional variance is the GARCH(1,1) model.

$$y_t = u_t \quad \text{where} \quad u_t = \sigma_t \epsilon_t \quad \text{and} \quad \epsilon_t \sim N(0,1)$$

$$\sigma_t^2 = \omega + \alpha_1 u_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

- The only difference is the $\beta_1 \sigma_{t-1}^2$ term.
- The β_1 coefficient essentially allows the conditional variance to persist over time. This typically results in a better fit to real financial returns data.

  The GARCH(1,1) model has 3 unknown parameters: $(\omega, \alpha_1, \beta_1)$.

GARCH(1, 1) model

- The GARCH model can be shown to be an ARMA model for the conditional variance.
- For example, consider again the GARCH(1, 1) model:

$$\sigma_t^2 = \omega + \alpha_1 u_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

- Then it can be shown to be an ARMA(1,1) process for the squared errors.

$$u_t^2 = \omega + (\alpha_1 + \beta_1) u_{t-1}^2 + v_t - \beta_1 v_{t-1}$$

GARCH(1,1) Model

- In practice, when people (both in academia and industry – banks, hedge funds, etc) use a GARCH model for time-varying conditional 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 GARCH(1,1) process in R.

Simulating from a GARCH(1,1) model

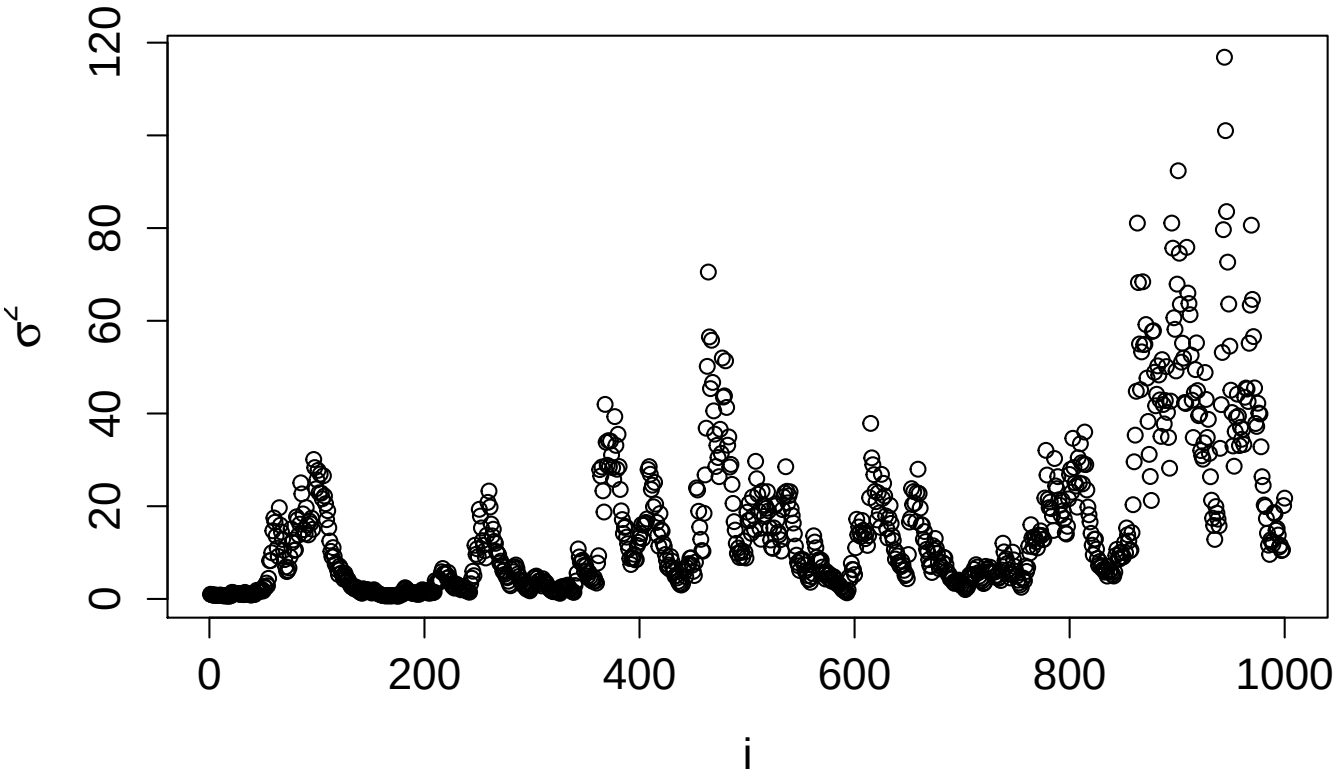
```
beta0 <- 0.1; beta1 <- 0.2; beta2 <- 0.8
n <- 1000
sigma2s <- numeric(n);
y <- numeric(n)

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]))
}
```

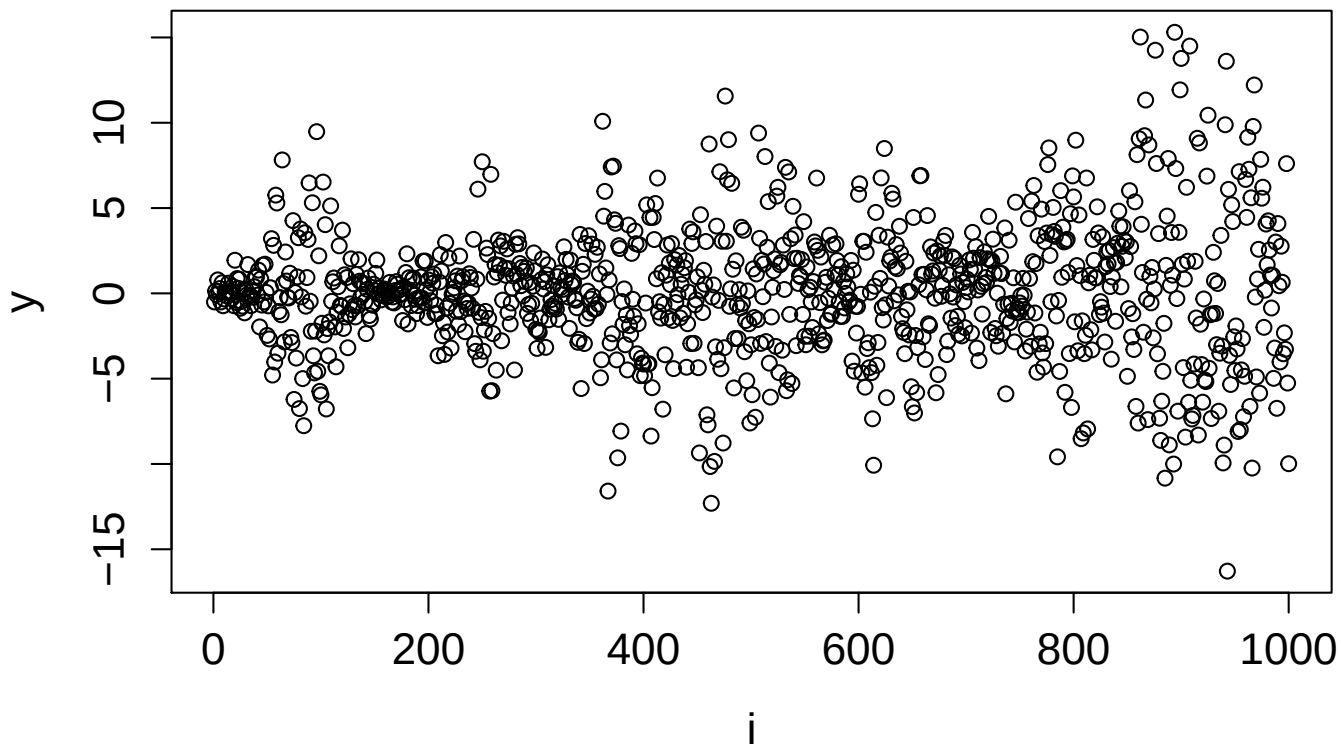
GARCH(1,1) Model

Using the above parameter values, this is a simulated realisation of σ_t^2:



GARCH(1,1) Model

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



GARCH model as an ARCH(∞) process

- Why GARCH models are more popular than ARCH models?
- GARCH model is a more parsimonious model of the conditional variance than a high-order ARCH model.
- For instance, GARCH(1,1) model can be written as an infinite order ARCH model:

$$\begin{aligned}\sigma_t^2 &= \omega + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \alpha_3 u_{t-3}^2 + \dots \\ &= \omega + \sum_{p=1}^{\infty} \alpha_p u_{t-p}^2\end{aligned}$$

The unconditional variance under a GARCH specification

- It is important to note that although the conditional variance of u_t^2 is changing, the unconditional variance is constant.
- Again, consider the GARCH(1, 1) model.
- Then the unconditional variance of u_t is constant and defined as follows:

$$\text{var}\left(u_t\right) = \frac{\omega}{1 - \left(\alpha_1 + \beta_1\right)}$$

- From this definition it is evident why the restriction $\alpha_1 + \beta_1 < 1$ is required.
- For $\alpha_1 + \beta_1 \geq 1$, the unconditional variance of u_t is not defined.
- When $\alpha_1 + \beta_1 = 1$, the model is termed IGARCH, i.e. integrated GARCH, which means a "unit root in variance".

TGARCH model

- An extension of GARCH model has been proposed in order to account for the *leverage effects*.
- *Leverage effects* refers to the tendency of volatility to react differently to a positive change in price than to a negative change.

The threshold $GARCH(p,q)$ model is defined by:

$$y_t = \mu + u_t \quad \text{where} \quad u_t = \sigma_t \epsilon_t$$

$$\sigma_t^2 = \omega + \sum_{p=1}^{P} \alpha_p u_{t-p}^2 + \sum_{o=1}^{O} \gamma_o u_{t-o}^2 I_{[u_{t-o}<0]} + \sum_{q=1}^{Q} \beta_q \sigma_{t-q}^2$$

where σ_t^2 is the conditional variance given past information, and ϵ_t are $i.i.d.$ random variables, ω, β_q, γ_o, $\alpha_p > 0$ assures that the conditional variance σ_t^2 is positive.

- This volatility model is also known as the GJR model, named after the authors Glosten, Jagannathan and Runkle (1993).

GARCH(1,1) Parameter Estimation

The parameters of the GARCH model are $\theta = \{\omega, \alpha_1, \beta_1\}$. 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. More advanced methods such as Markov-Chain Monte Carlo are also useful, but beyond the scope of this module.

Model Selection

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). 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. 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.