PS3 Data Analysis and Machine Learning

lewispicker

May 2023

1 Question 1

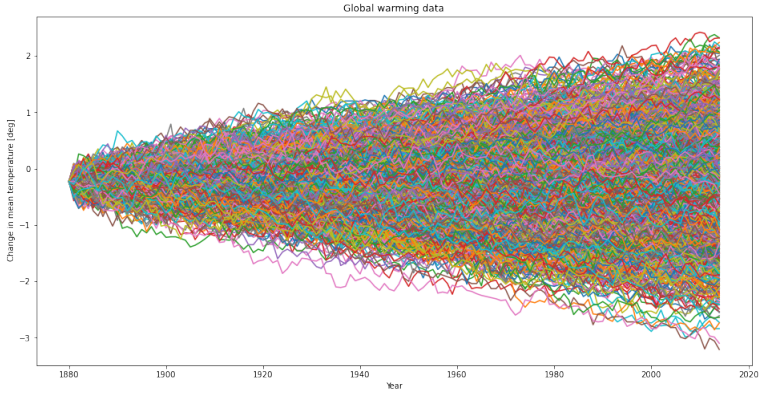

Figure 1: A plot of the change in temperature over 135 years. Each line represents a series that was drawn from either $\pm 1^\circ$ C change in temperature over 100 years or from no trend (with some homoscedastic noise).

2 Question 2

What we have in our data set is 1000 ficticious series (with some homoscedastic noise, constant variance) each containing 135 data points. Where the data in the series represents the change in the mean temperature every year from 1800 to 2014. Each series is drawn from one of three models. Some of the series were randomly selected to have a trend that would either increase/decrease the mean temperature by ± 1°C

It is important to note the the data points in the series are not independant of each other, and do depend on the previous value. I will also note that a 1°C change in a centery would correspond to a change in 0.01°C every year.

Thefore a particular series will have its data drwan from one on the three models:

$$y_i \sim \mathcal{N}(y_{i-1} - 0.01, \sigma)$$

$$y_i \sim \mathcal{N}(y_{i-1} + 0.01, \sigma)$$

$$y_i \sim \mathcal{N}(y_{i-1}, \sigma)$$

Where i is the i th data in the series and σ is the homoscedatic noise that is the same for each dataset.

To furthur tackle this problem we need to envoke a mixture of models which takes into account that a particular series could be drawn from 1 of the 3 models:

$$y_{n,i,k} \sim \{ \ N(y_{n,i-1} - 10^{-2}, \sigma)k = 1\mathcal{N}(y_{n,i-1} + 10^{-2}, \sigma)k = 2\mathcal{N}(y_{n,i-1}, \sigma)k = 3$$

$$y_{n,i,k} \sim \{ N \ (y_{n,i-1} - 10^{-2}, \sigma)k = 1N(y_{n,i-1} + 10^{-2}, \sigma)k = 2N(y_{n,i-1}, \sigma)k = 3 \tag{1}$$

n specifies which series (of which there are 1000 of). The parameter k which specifies the model that the series is drawn from. The introduction of this new parameter for ever seriers means that we have just introduces 1000 more individual parameters to fit for, luckily we can marginalise out k parameters by introducing a prior.

$$p(k) = \pi_k$$

The mixing proportions (π_k) will be bouned from 0 to 1, but there is an additional condition that the sum of the components must equal 1. To state this we say that :

$$\pi_k \sim Multinomial(K)$$

Where K is the total number of components in the model.

The overall posterior probability will be:

$$p(\pi_k, \sigma|\mathbf{y}, \mathcal{M}) \sim \sum_{n=1}^{N} \sum_{k=1}^{K} log(\pi_k y_{n,i,k})$$

3 Question 3

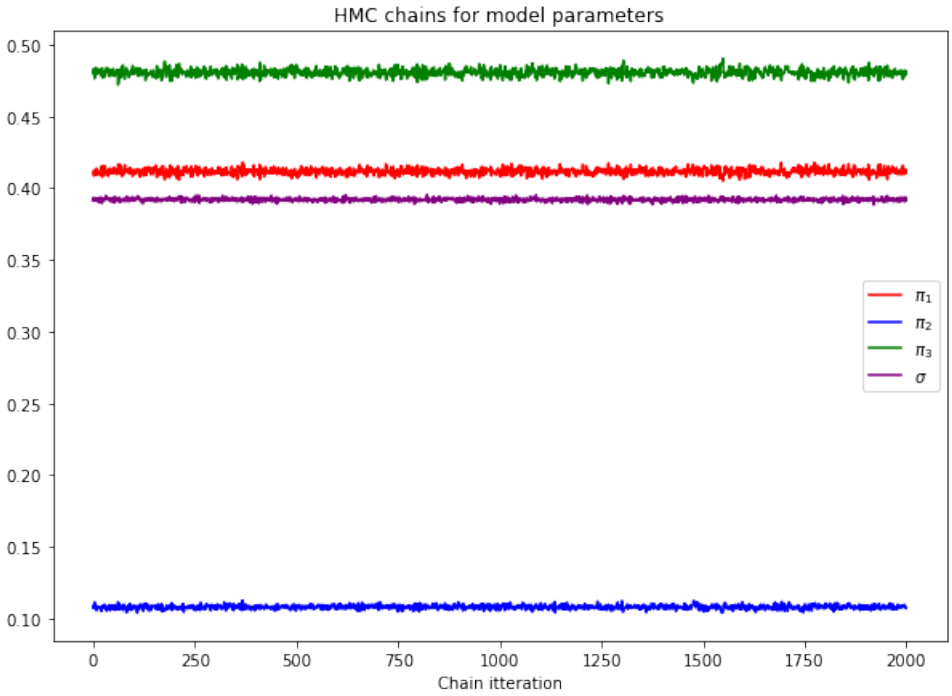

Figure 2: A plot depicting the four HMC chains of the parameters of my specified model. It appears that the have all converged and that the probability sums up to 1 which is a good sign.

4 Question 4

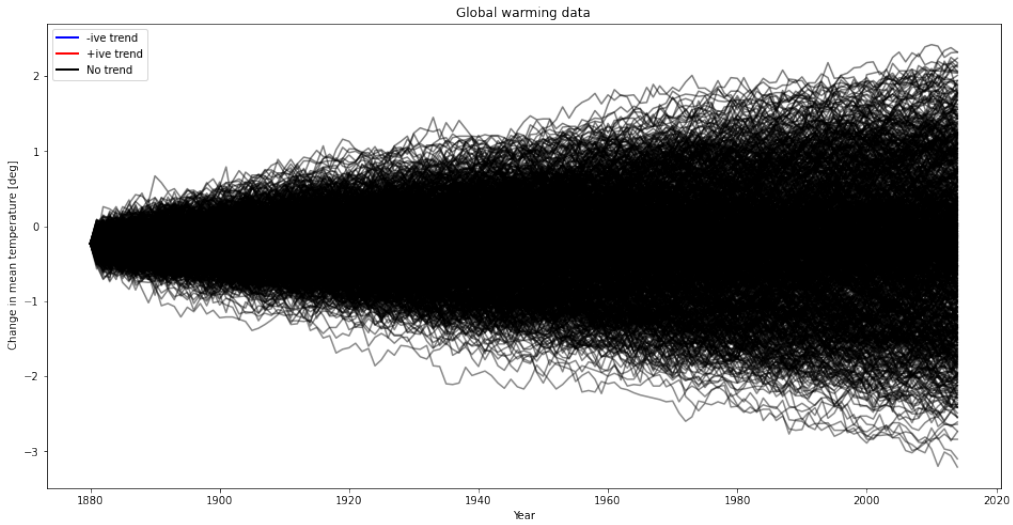

Figure 3: A plot depicting data series coloured by the likelihood that it was drawn from a particular model and opacity w.rt to that probability. This does not look reliable and I think is a consequence of how I measured the component probabilities.

5 Question 5

$$E_{correct} = 490.8 \pm 15.8$$

6 Question 6

As one might of expected from the mean and standard deviation of the expected correct values, that the probability of correctly identifying 900 or more series is pretty much 0.

Therefore I would not take the bet as I would see better returns in a money shredder.