Problem Set 3

Advanced data analysis and machine learning

> **Important**
>
> You must upload your problem set to Moodle before 23:55 on 2023-05-18.
>
> Your submission must include:
>
> - a PDF that responds to all questions; and
> - a single Python script (`.py`) that you used to answer all questions.
>
> Answers without accompanying Python code will recieve zero marks.

There is overwhelming evidence of anthropogenic climate change. This question relates to what can be inferred from only a subset of the evidence available for global warming: a ficticious set of global temperatures spanning 135 years.

In the accompanying `temperatures.txt` file you will find 1,000 fictious time series. Each series has length 135, assuming one measurement per year of temperature deviation from the mean, covering the time period from 1880-2014, inclusive. The data were first generated by drawing 1,000 random series (with some homoscedastic noise). Then, some of those series were randomly selected and had a trend added to them. The trends that were added were either +1°C / century or -1°C / century.

A bet has been offered for anyone who can correctly identify at least 900 series: those that were generated without a trend and which were generated with a trend. The prize is $100,000 US dollars.

Question 1 [10 points]

Make a figure showing all series.

Question 2 [20 points]

Specify a generative model for these data.

> **Warning**
>
> There are two ways to think of the data generating process here. One way is to think that all 135,000 data points (1,000 x 135) are independent, and some of those data points are drawn from the series with no trend added, and some data points are drawn from one of the models with a trend added.
>
> This is **incorrect**. It is not that the 135,000 data points are drawn from one of the three models. It is that each of the 1,000 series is drawn from one of those three models. If one series was drawn from a model with a trend line, then **every data point in that series** will have a trend.
>
> In other words: you have to assume the series have 'hard assignment': being drawn from one of the three models.

Question 3 [40 points]

Implement this model in a probabilistic programming language of your choice, and sample the posteriors. Plot the chains and comment on whether the MCMC has converged.

Question 4 [10 points]

For each time series, calculate the log posterior probability that it is:

- an unaffected time series,
- a time series with a +1°C / century trend added, or
- a time series with a -1°C / century added.

Make a plot showing all trend lines, coloured by their log posterior probability.

> **Tip**
>
> If you calculated the three log probability values for one series and saw the unaffected time series was the most probable scenario for that trend line, then you can plot that line in black and set the transparency to be the **relative** probability that it is unaffected (e.g., `alpha=0.6` if the relative probability is 60%). To do this you will need convert your log probabilities to relative probabilities.
>
> If you find that the most probable scenario for a trend line was -1°C / century, then colour that trend line in blue with the transparency set to the relative probability.
>
> If you find that the most probable scenario for a trend line was +1°C / century, then colour that trend line in red with the transparency set to the relative probability.

Question 5 [10 points]

For each time series, find the highest probability of membership from the available mixtures. Sum this value for all time series. This gives an expectation 1 for the number of series you may have estimated correctly:

$$E_{\text{correct}} \approx \sum_{j=1}^{N} \max(p_{j,k})$$

The uncertainty for the number of series you have estimated correctly can be calculated by

$$\sigma_{\text{correct}} \approx \sqrt{\sum_{k=1}^{N} \left[\max(p_{j,k}\,(1 - \max(p_{j,k}))\right]}$$

How many series do you expect to have calculated correctly? What is the uncertainty on that expectation value?

Question 6 [10 points]

Assuming a normal distribution, what are the chances that you would correctly identify 900 or more time series as being: an unaffected time series, a time series with +1°C / century added, or a time series with -1°C / century added?

If you had to pay $10 to submit an entry to this competition, is it a worthwhile competition to enter (a worthwhile bet)?