Exam

D. Jason Koskinen

koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2017

Photo by Howard Jackman

Info

- In submitting the solutions there is no need to rephrase the problem. “Solution for 1a” is sufficient.
- The submission format for explanations and plots is a PDF file. Also, include any and all software scripts used to establish your answer(s) and/or produce plots.
- Working in groups or any communication about the problems is prohibited. Using the internet as a resource is encouraged, but soliciting any help is also prohibited.
- Some questions have multiple parts. For full credit, all parts must be done.

Info

- There are no ‘optional’ problems. All problems will be considered for assessment.
- The exam will be graded out of 100 possible points
    - It will count for 40% of the final course grade
- Submit all code used!! The software you write to complete the problem is **part** of the solution.
- Must be submitted by 17:00 CET Friday March 31, 2017 for full credit. This can either to Jason via email or electronically submitted via the Digital Exam website.
- For any concerns, questions, or comments email Jason.

Starting points (5 pts.)

- On the first page of your write-up include your full name, date, name of this course, and the title of your exam submission
- Also type out (please don't copy/paste) " I (your name here) expressly vow to uphold my scientific and academic integrity by working individually on this exam and soliciting no direct external help or assistance."
- Finding help/solutions online is completely fine. But, for example, posting to a forum and receiving assistance is not okay.
- Good luck!!!

Problem 1 (25 pts.)

- There is a file posted online which has 5 columns, each representing a physical observable of interest generated from some underlying function. There are thousands of entries, i.e. rows.
  - http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2017/data/Exam_Prob1.txt
  - The variables/columns are independent distributions with no correlation to the data in the other columns
  - Be mindful about accounting for truncated ranges as well as likelihood functions that have periodic components which will create local minima/maxima
    - There is at least one column of data which is generated from a function with local minima/maxima

Lists of Distributions

$$-10 \leq a \leq 10$$
$$-10 \leq b \leq 10$$
$$4000 \leq c \leq 8000$$

- The data in each column is produced from functions ***similar to***, or potentially exactly the same as, those shown at right
- Note that these functions may be unnormalized
  - Hint: Some will require a normalization to convert them to probability distribution functions
  - The functions f(x) have bounds on their parameters a, b, and c

$$f(x) \propto \begin{cases} \frac{1}{x+5}\sin(ax) \\ \sin(ax)+1 \\ \sin(ax^2) \\ \sin(ax+1)^2 \\ x\tan(x) \\ 1+ax+bx^2 \\ a+bx \\ \sin(ax)+ce^{bx}+1 \\ e^{-\frac{(x-a)^2}{2b^2}} \end{cases}$$

$$f(k) \propto \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \quad binomial \\ \frac{\lambda^k e^{-\lambda}}{k!} & \quad poisson \\ \frac{-1}{\ln(1-p)} \frac{p^k}{k} & \quad logarithmic \end{cases}$$

Problem 1a

- Use the separate data from columns 1, 2, and 3, and identify the function on the previous slide from which each was generated. Find the best-fit values and uncertainties on those values for the distribution using a likelihood method (either bayesian or maximum likelihood is fine)
    - E.g. if f(x)=sin(ax+b)*exp(-x+c)+x/k! were one of the functions, then find the best-fit values for a, b, c, and k and their uncertainties
    - Degeneracies exist, e.g. sin(x)=cos(a+x), which can produce functionally identical data distributions
    - Any function, with associated best-fit parameters which is statistically compatible with the data in the files will be accepted as a proper solution. Only one is necessary, but needs to be justified as statistically compatible.
- Data in column 1 and 2 have artificially truncated ranges
    - Column 1 is only sampled in the independent variable from 20 to 27
    - Column 2 is only sampled in the independent variable from -1 to 1

Problem 1b

- Plot the data and the corresponding best-fit function on the same plots
  - 3 separate 1-dimensional plots
  - Plot as a function of the independent variable
  - Histogram the data, and scale the best-fit function to be ‘reasonable’ so that the features of both the data and best-fit function can be visually compared

Problem 2 (15 pts.)

- The probability of obtaining a High Threshold hit (pHT) in the ATLAS Transition Radiation Tracker depends on the logarithm of the γ-factor of the traversing particle. In the file linked below there are 60 measurements of pHT including uncertainties for various values of ln(γ).
  - http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2017/data/data_exam_prob2.txt
  - At low values of ln(γ), pHT is constant. Up to what value of ln(γ) is it consistent with being constant? Justify numerically. What (constant) value of pHT do you find?
  - Fit the distribution with suitable function(s), and possibly argue which one best describes this distribution. The function is not required to be any from slide 5.

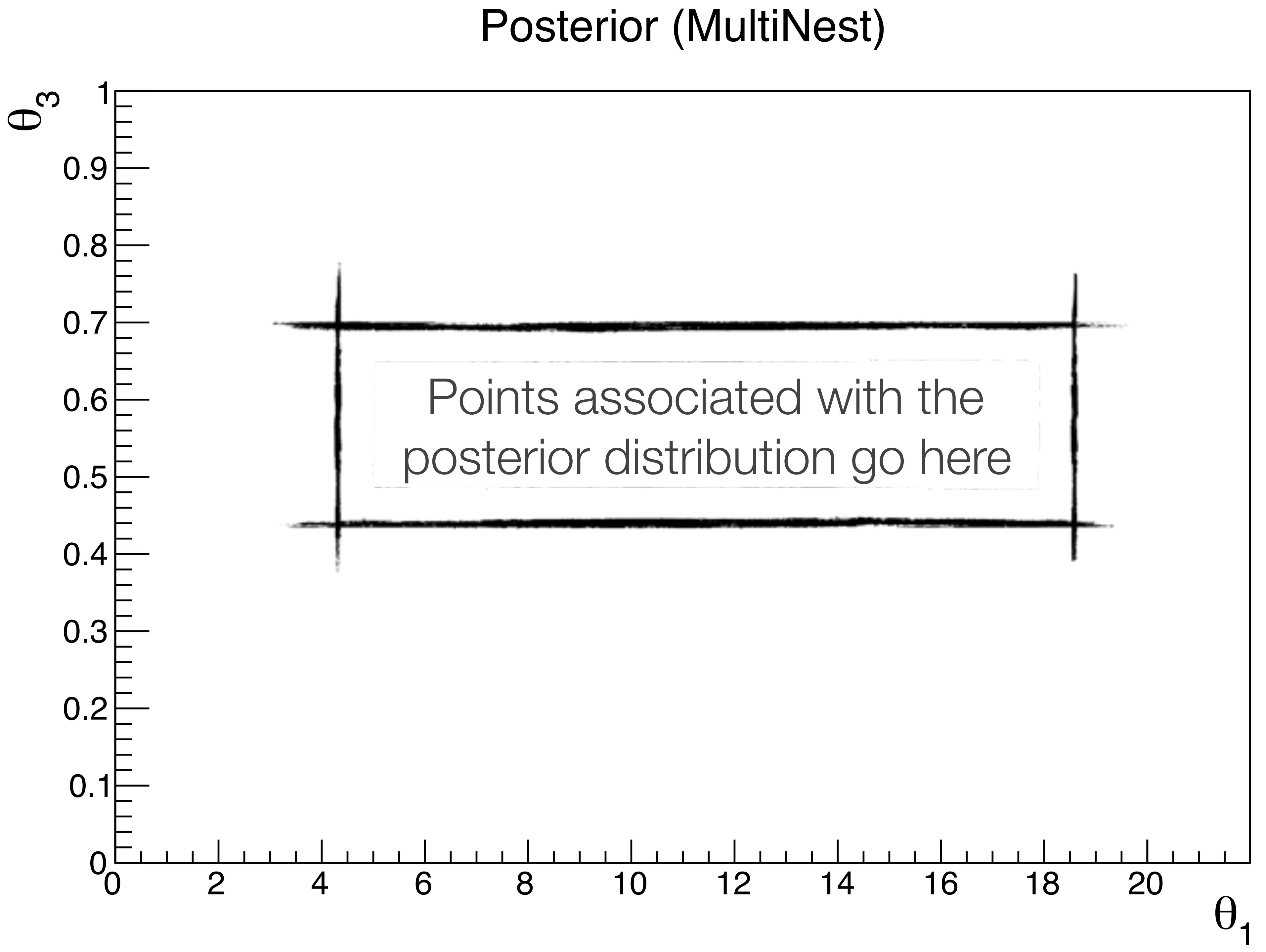
Problem 3 (20 pts.)

- With the function below as the defined quasi-likelihood in 3-dimensions use MultiNest, or some other nested sampling bayesian algorithm, to plot the 2-D posterior distribution for the parameters θ_1 and θ_3, i.e scatter-point plot for θ_3 vs. θ_1 (empty example on a following slide)

$$\mathcal{L}(\theta_1, \theta_2, \theta_3) = 3\Big(\cos(\theta_1)\cos(\theta_2) + \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\theta_3-\mu)^2}{2\sigma^2}} \cos(\theta_1/2) + 3 \Big)$$

  - The range should be restricted for θ_1 and θ_2 to 0-7π and for θ_3 from 0-3. Also, set μ=0.68 as the true mean of the normal distribution and σ^2=0.04
  - What are the best-fit values for θ_1, θ_2 ,and θ_3 that you find from maximizing the above function, i.e. when you generate the posterior distribution?

Problem 3


Posterior (MultiNest)
θ_3
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
2
4
6
8
10
12
14
16
18
20
θ_1
Points associated with the posterior distribution go here

Problem 3 (cont.)

- The posterior distribution is proportional to the output of the quasi-likelihood. Make two separate raster scan plots in 2-D of the output from the likelihood over the same ranges as for the previous plot. Essentially, map out the likelihood (or ln-likelihood) landscape.
  - For the scan of θ_2 vs. θ_1, fix θ_3 to the best-fit point found from MultiNest, i.e. an unchanging value. Similarly, for the scan of θ_3 vs. $\theta_{1,}$ fix θ_2 to the best-fit point.
  - For reference, look at slide 26 from the lecture notes at http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2017/Lecture13_MultiNest.pdf
- Does the posterior distribution match the raster scan plots? Discuss why it should, or why it should not.

Problem 4 (20 pts.)

- There are two files which contain sea surface water temperatures from global monthly data from HadSST3
  - May 1997 at http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2017/data/GlobalTemp_1.txt
  - May 2017 at http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2017/data/GlobalTemp_2.txt
- Using data in the 8th row (including 1 line for the header info), construct a kernel density estimator using the Epanechnikov kernel with a bandwidth of 0.4
  - The 8th row is a band of constant latitude near Denmark
  - 1.07 C is the first entry in the 8th row for 2017, and 0.74 C for 1997
  - Do **not** include entries in constructing the KDE where there are no temperature measurements

http://hadobs.metoffice.com/hadsst3/

Problem 4 - Data Format

Data are stored in ASCII
Temperatures are stored as degrees C Land squares and missing data are set to -99.99 or, in the case of numbers of observations, 0
The month and year are stored at the start of each month.
Data Array (72x36) Item (1, 1) stores the value for the 5-deg-area centred at 177.5W and 87.5N Item (72, 36) stores the value for the 5-deg-area centred at 177.5E and 87.5S

```
          ----- -----
         |     |     |
         | MON | YR  |
         |_____|_____|____________________________
    90N  |(1,1)                                   |
         |                                        |
         |                                        |
         |                                        |
         |                                        |
         |(1,18)                                  |
    Equ  |                                        |
         |(1,19)                                  |
         |                                        |
         |                                        |
         |                                        |
         |                                        |
    90S  |(1,36)____________________________(72,36)|
          180W               0               180E
```

*from the README file

Problem 4a

- Plot the P_{KDE}(temp) as a function of temperature for both 1997 and 2017 over the range of -2 C to +4 C
  - P_{KDE}(temp) is the data driven kernel density estimated probability distribution function (PDF)

- Calculate the integral of P_{KDE}(temp) for 1997 and 2017:
  - over the range -2 C to +4 C
  - over the range of -2 C to 0 C

Problem 4b

- Produce 1000 Monte Carlo draws/samples/events from the 1997 P_{KDE} over a temperature range from -1 C to +2 C
- Calculate the likelihood ratio for the 1000 Monte Carlo samples where H_0 uses the KDE from 1997 and H_1 uses the KDE from 2017

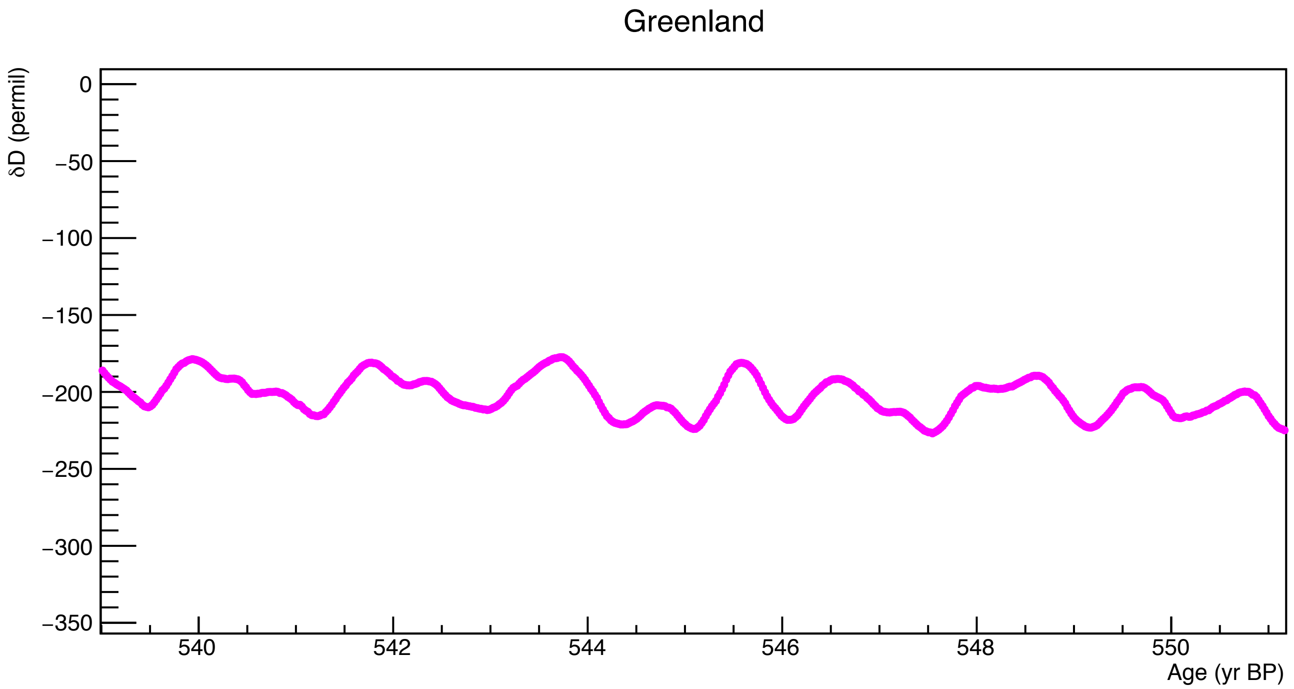
$$\frac{\mathcal{L}(H_0|x)}{\mathcal{L}(H_1|x)}$$

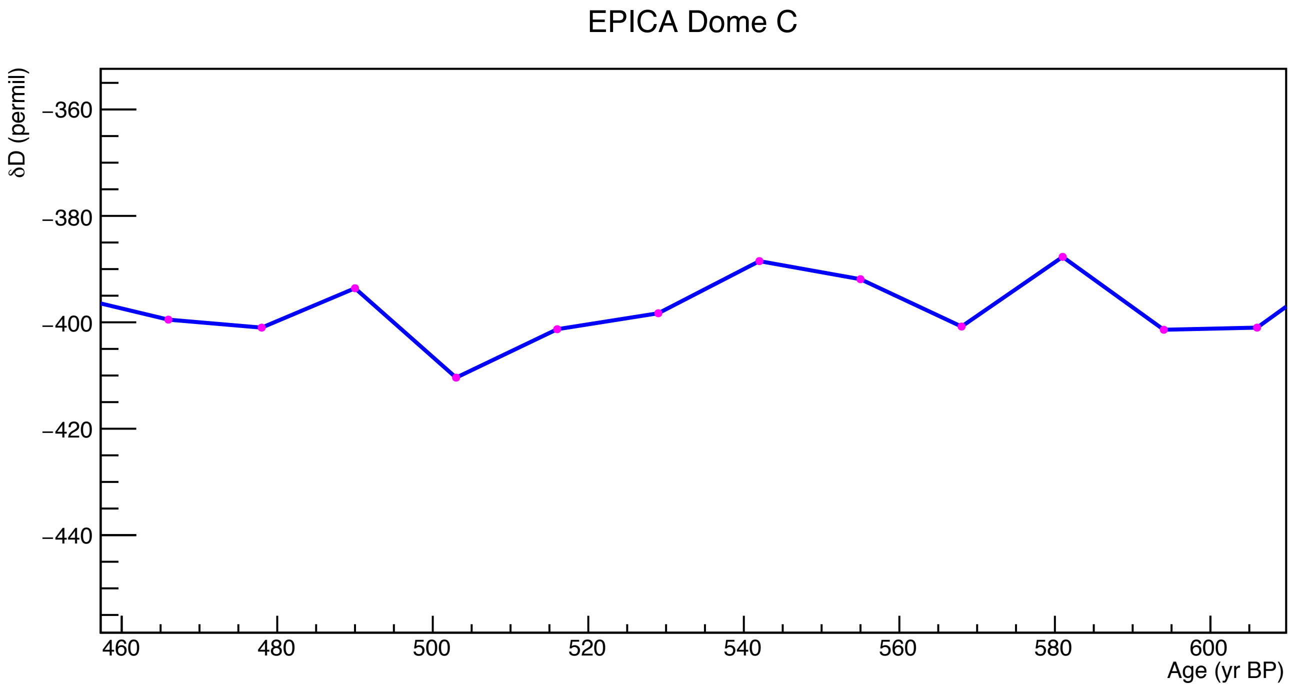
Hint: Do the calculation as a function of natural log(s) and then convert back to 'normal' likelihood values

- Submit your 1000 samples as an ASCII txt file:
  - each entry on a separate line for 1000 total lines in the file
  - File name should be your last name and "_KDE_1000_samples.txt", e.g. "koskinen_KDE_1000_samples.txt"

Problem 5 (15 pts.)

- Wavelets using ice core data from Niccoló Maffezzoli





This is just a nice image of the data

Notes from Niccoló

OVERVIEW

Two ice core records are available. REN = Renland ice core (Greenland) EDC = Epica Dome C (Antarctica). Both cores are measured for stable water isotopes on a depth-scale. Particularly, the quantities d18O and dD are proxies for temperature. Therefore, by looking at their time series, one can have a look at the variability of glacial and interglacial cycles. The depth scale is converted to an age scale by methods which are not important here.

THE SIGNAL

The REN core spans the Holocene (last 15000 years approx), the last glacial cycle and reaches the previous interglacial (Eemian, 125000 years ago) By looking at the REN core, you can resolve annual signals (!!!) in the top(most recent part). The glacial part contains abrupt warming events, called Dansgaard-Oeschger events. The frequency and the reasons of these climate oscillations are unknown. https://en.wikipedia.org/wiki/Dansgaard-Oeschger_event
The EDC core (the oldest core available) is an Antarctic core well suited for reconstructing climate very far back in time (800000 years). What you see in this record is climate variability induced by insolation variability due to orbital parameter changes (https://en.wikipedia.org/wiki/Milankovitch_cycles)

DATA

The time series are imported using pandas in python. You have only dD (delta Deuterium) variable for comparison between the two cores. You can access the variable x serie of each core by CORE[‘X']. Missing values in the data are consider NaN. This could be a problem when plotting*.

*we’re not doing any plotting where this will be a problem

Data location

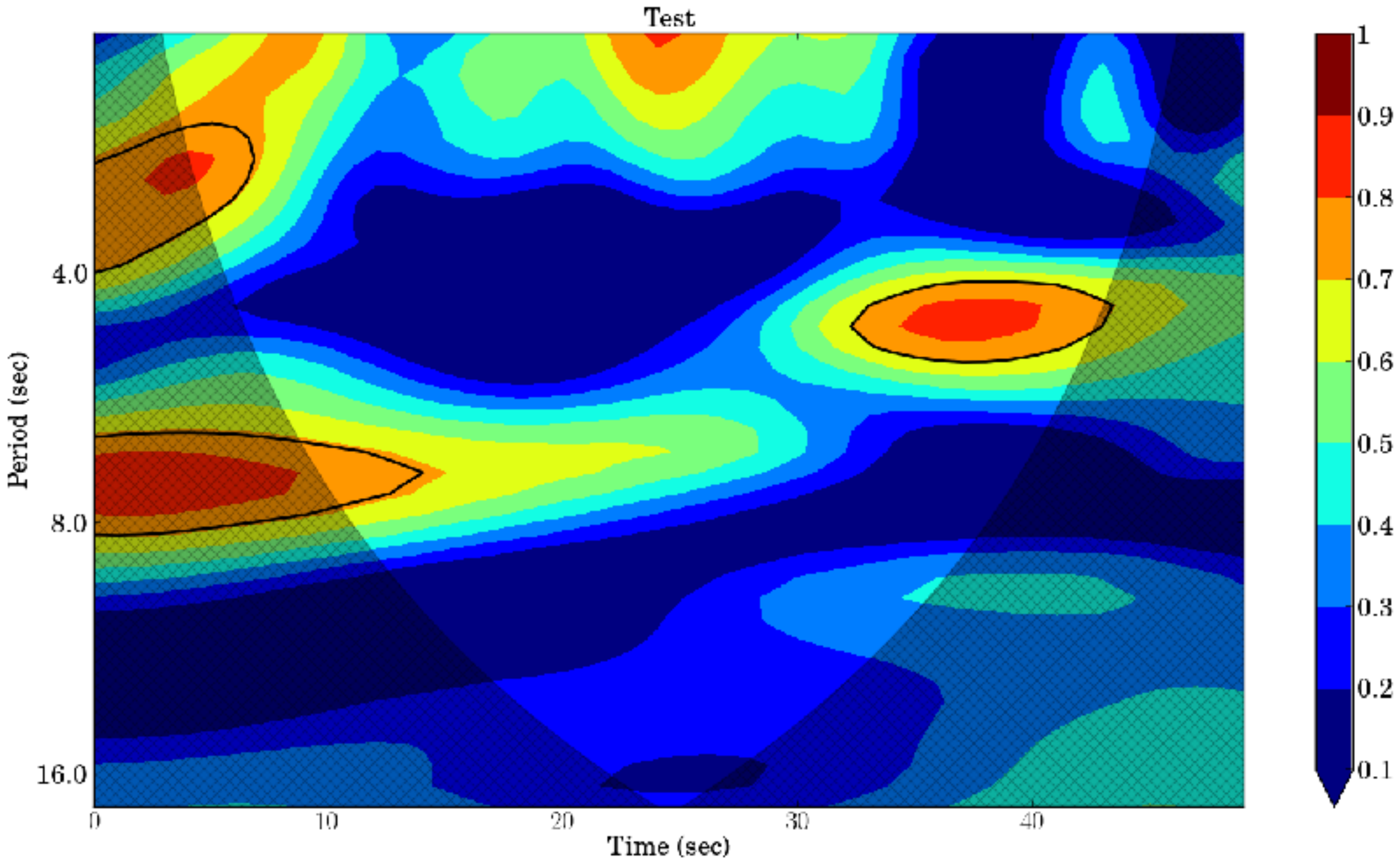
- For the data the focus will be on dD (delta-Deuterium) for only one of the ice core data samples (Ren), which can be found at either:
  - http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2017/data/Exam_RECAP_d18O_dD_d_5mm_dated_withTimescaleV4.csv
  - Already parsed version: http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2017/data//Exam_IceCore_Ren.csv
- Python pseudo-code using PANDAS from Niccolo to load the original .csv file can be found at http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2017/Niccolo_IceCore_loading.py

Problem 5a

- The task is to use wavelets to search for features in the data.
- Use the Ricker wavelet and produce the 2D continuous wavelet coefficient map using
  - The year range of 540 to 550 before present (BP)
  - Scales/widths defined as 2^{N-2} for N in integers from 1 to 10, i.e. [0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256] for the wavelet
  - For inspiration look at slides 7-9 from the wavelet lecture to get a reminder about wavelet coefficient map(s)
  - Include a color scale bar which includes the numerical range of the coefficients (see an example on the next slide)

Example

For illustrative purposes only



Eduardo S. Pereira and Regla D. Somoza
PIWavelet & duducosmos

Problem 5b

- The data may express more features if we look for deviations from the mean instead of using the absolute value of dD as an inherent amplitude
- Use the Ricker wavelet again and produce a 2D continuous wavelet coefficient map using
  - The year range of 540 to 550 before present (BP)
  - Scales/widths defined as 2^{N-2} for N in integers from 1 to 10, i.e. [0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256]
  - The input data for the wavelet should have the mean of the dD over years 540-550 subtracted off. This should adjust the pink line in the upper plot from a few slides back to have dD values bounce around a value of 0 instead of approx. -200