Exam

D. Jason Koskinen
koskinen@nbi.ku.dk

Advanced Methods in Applied Statistics
Feb - Apr 2023

Photo by Howard Jackman

Info

- In submitting the solutions there is no need to rephrase the problem. For example, “Solution for 1a” is sufficient.
- The submission format for explanations and plots is a PDF file. Also, in a separate file (or files) include any and all software scripts used to establish your answers 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 prohibited.
- Some questions have multiple parts. For full credit, all parts must be done.

Info

- The exam will be graded out of 10 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.
- The exam must be electronically submitted via the Digital Exam website.
    - For catastrophic submission failures you can email the exam submission to Jason
- Look through all problems in the exam. Some problems are easier than others.
- For any concerns, questions, or comments email Jason (koskinen@nbi.ku.dk)

Starting

- On the first page of your write-up include your full name, date, name of this course, UCPH ID, 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, academic, and moral integrity by working individually on this exam and soliciting no direct external help or assistance."
- Finding help/solutions online is fine. But, for example, posting to a forum and receiving assistance is not okay.
- Good luck!!!

Problem 1 (3.0 pts.)

- There is a file posted online which has 5 columns, each representing data of interest generated from some underlying function. There are 5119 entries, i.e. rows.
  - http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2023/data/Exam_2023_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

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, f(x) or f(k) shown at right
- Note that the displayed 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

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

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

Problem 1a

- Use the separate data from first, second, and third columns to 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 solution is necessary, but needs to be **justified** as statistically compatible.
- The first and second columns have artificially truncated ranges
  - First column is only sampled in the independent variable from 20 to 27
  - Second column 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 (2.0 pts.)

- There is a file posted online (http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2023/data/Exam_2023_Problem2.txt) with data.
  - The first column is the azimuth angle of the data point
  - The second column is the zenith angle of the data point
  - There are 139 paired data points in total
  - The values are in units of radian

Problem 2a

- Correctly quantify whether the data is spherically isotropically distributed
  - Include any supporting plots, discussion, and numbers
  - A spherically isotropic distribution is uniform in the azimuth angle from 0 to 2π, and uniform in cos(zenith angle) from -1 to 1
    - Hint: you can use Monte Carlo generated pseudo-experiments to produce a test-statistic distribution of a spherically isotropic distribution.
    - Hint: isotropically distributed means ‘uniform’ **simultaneously** in azimuth and cos(zenith).

Problem 2b

- Test whether the data fits the two following alternative hypotheses better than the isotropic hypothesis:
  - Hypothesis A: That 20% of the total sample is uniformly distributed in azimuth over the range {0.225π, 0.725π} and uniformly distributed in zenith over the range {0.30π, 1π}, and the remaining 80% is fully isotropic
  - Hypothesis B: That 15% of the total sample is uniformly distributed in azimuth over the range {0π, 1π} and uniformly distributed in zenith over the range {0.5π, 1π}, and the remaining 85% is fully isotropic.
  - Report the two p-values:
    - $H_{isotropic}$ versus H_A
    - $H_{isotropic}$ versus H_B

Problem 3 (1.5 pts.)

- The following function is for this problem:

$$f\left(x|a,b\right) = \frac{\cos(a \cdot x)\cos(b \cdot x)}{x^2} + 2$$

- To normalize the function and create a probability distribution function requires the indefinite integral, which includes the sine integral "Si(x)". The indefinite integral can be expressed as:

$$0.5 * \left((b-a)\mathrm{Si}((a-b)x) - (a+b)\mathrm{Si}((a+b)x) - \frac{2\cos(ax)\cos(bx)}{x} + 4x \right)$$

  - There is a scipy special function to calculate the sine integral
  - Alternatively, instead of using the indefinite integral to get the normalization to construct a PDF, you can use trapezoidal summation or some other numerical method

Problem 3

- There is a file at https://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2023/data/Exam_2023_Prob3.txt containing Monte Carlo generated x values from the probability distribution function
  - The function is only sampled over a range of $1 \leq x \leq 3$
  - The true values are in the range of $0 \leq a_{true} \leq 15$ and $9 \leq b_{true} \leq 27$

- Using the data from the file, what are the best-fit values of a and b, i.e., $\hat{a}$ and $\hat{b}$?

- Make and submit a 2D raster scan of the test-statistic used for the fitting routine around the best-fit parameters $\hat{a}$ and $\hat{b}$.
  - Be sure to label all axes and include a color scale with an appropriate color bar
  - The raster scan should be over the range $(\hat{a} - 3) \leq a \leq (\hat{a} + 3)$ and $(\hat{b} - 3.5) \leq b \leq (\hat{b} + 3.5)$
  - The raster scan should be in steps no greater than 0.1 for both a and b

Problem 4 (1.5 pts.)

Credit: Luca Galuzzi - www.galuzzi.it

- Data from the World Glacier Inventory* includes information from over 130k glaciers on Earth. The elevation of a subsample of 1000 glaciers is included at https://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2023/data/Exam_2023_Prob4.txt

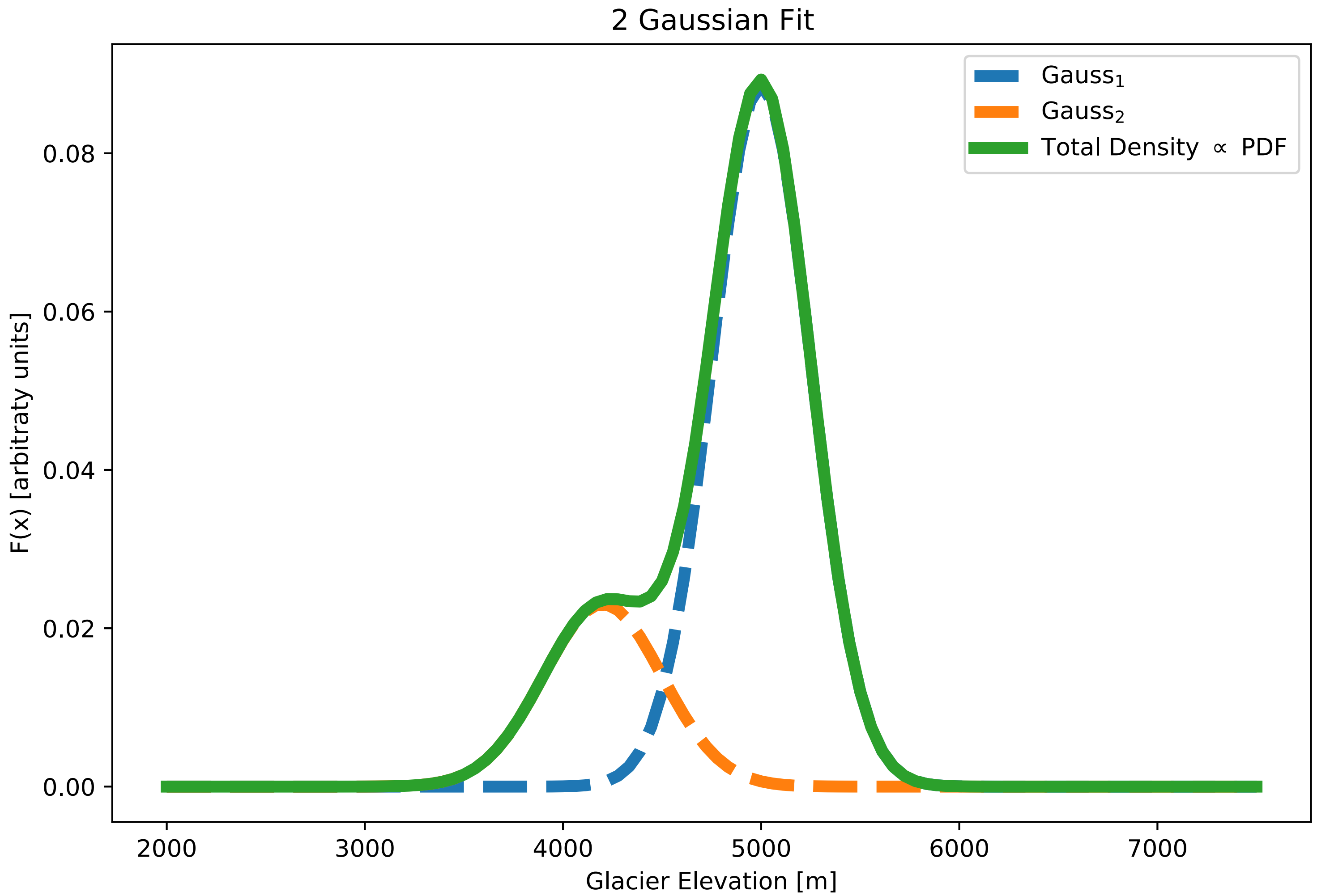
*https://nsidc.org/data/g01130/versions/1/documentation

Problem 4a

- Fit two different hypotheses to the glacier elevation data:
  - Hypothesis A — Double gaussian fit
  - Hypothesis B — Triple gaussian fit
  - Plot the total probability distribution function for each hypothesis

Problem 4a (illustration)

- Below is an example illustration of a double-gaussian fit which uses a function which is proportional to the actual PDF and built from 2 gaussian distributions. This illustration is not fit to any data, and is only an illustration.



Problem 4b

- At what statistical significance can Hypothesis A be excluded/rejected when compared to Hypothesis B for the data?
  - Provide a quantitative number(s) calculated from the data and your reasoning for the hypothesis testing acceptance/rejection.
  - Hint: a p-value is a useful way to report hypothesis testing results.

Problem 5 (2 pts.)

- Small problems

Problem 5a (1.5 pts.)

- The following data file has a list of test statistic values. For this test-statistic higher values are always associated with worse agreement than lower values.
  - The file is at: http://www.nbi.dk/~koskinen/Teaching/AdvancedMethodsInAppliedStatistics2023/data/Exam_2023_Problem5a.txt
- The file is a list of 3000 bootstrap test-statistic samples. What is the critical value, i.e. threshold, of the test-statistic that corresponds to a one-sided p-value of 4.55%?
- If the true distribution for the test-statistics in the file is chi-squared distributed, does the test-statistic threshold established with the bootstrap samples match the expected critical value from a chi-squared distribution with 5 degrees-of-freedom (k=5)?
  - Quantitatively and qualitatively justify your answer.

Problem 5b (0.5 pts)

- Make a linear and a Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) spline using the following (x, y) data: (1, 3.4), (1.7, 3.9), (1.9, 2.6), and (2.2, 3.1)
- What is the interpolated y-value for x=2.0 from the linear spline and the PCHIP spline?