UNIVERSITY OF CAMBRIDGE

MPhil in Data Intensive Science

Michaelmas Term 2024

S1: Statistical Methods Coursework

Matt Kenzie

*Attempt **all** parts of the coursework.*

The anticipated number of marks allocated to each part of a question is indicated in the right margin, to be assessed from the both the code repository and your report

This coursework should be submitted via a GitLab repository which will be created for you. Place all of your code and your report into this repository. The report should be in pdf format and placed in a folder named `report`*. You will be provided access to your repository until 23:59pm on Wednesday the 18th of December, after which we will assess whatever the repository contains at that time.*

The report should be laid out with separate sections for the answers to parts a) – g). The expectation is that parts a) – d) have a relatively small amount of text. Part a) will be marked on the mathematical proof, including written justifcation of the steps taken. Parts b) – d) will be marked on the quality and efficiency of the code as well as the accuracy and quality of the plots. In parts e) – g) you are expected to write some text which explains and justifies the use of the methods. Part g) should be mostly text and provide a clear and coherent comparison between the two methods.

There is more than one way to approach this problem. It is more important that you explain and justify your approach, and submit suitable code to perform the task, than it is to get the answer "right".

You are expected to submit code and associated material that demonstrates good software development practices as covered in the C1 Research Computing module.

Your report should not exceed 3000 words (including tables, figure captions and appendices but excluding references); please indicate its word count on its front cover. You are reminded to comply with the requirements given in the Course Handbook regarding the use of, and declaration of use of, autogeneration tools.

Failure to include the wordcount and a declarations of use of autogeneration tools will result in automatic loss of marks.

1 This question aims to make a comparison between the statistical power of a multi-dimensional likelihood fit and a weighted fit exploiting *sWeights*. By the end you will perform a parametric bootstrapping, or a "toy study", by generating an ensemble of samples according to a two-dimensional model and then fitting those samples back to determine the model parameters. You will compare the performance of this with a contrasting method, *sWeights*, in which you fit only one dimension of the model and project the relevant component in the other dimension.

The *Crystal Ball* probability distribution is defined in the following way

$$p(X;\mu,\sigma,\beta,m) = N \cdot \begin{cases} e^{-Z^2/2} & \text{for } Z > -\beta \\ \left(\frac{m}{\beta}\right)^m e^{-\beta^2/2} \left(\frac{m}{\beta} - \beta - Z\right)^{-m} & \text{for } Z \leq -\beta \end{cases} \tag{1}$$

where $Z = (X - \mu)/\sigma$ is the standard normal transformation for a random variable X at location, μ, with scale, σ. The distribution has a "Gaussian core" with a lower-side power-law tail, where β and m are additional shape parameters that describe where the power-law tail transitions to the Gaussian (in units of σ), β, and what the slope of the power-law tail is, m. The distribution is only valid when $\beta > 0$ and $m > 1$.

(a) Show by a mathematical proof that the inverse of the normalisation constant, N, can be written

$$N^{-1} = \sigma \left[\frac{m}{\beta(m-1)} e^{-\beta^2/2} + \sqrt{2\pi}\Phi(\beta) \right], \tag{2}$$

where $\Phi(x)$ is the cumulative density of the standard normal distribution. [4]

(b) Consider a statistical model of two random variables, $X \in [0, 5], Y \in [0, 10]$ which consists of a signal component and a background component. The fraction of the total probability density which is signal is defined by the paramter f, in other words the signal-to-background ratio within the domain in which X and Y are defined is $f/(1 - f)$.

Both the signal and background models are independent in X and Y, in other words their probability densities factorise and can be written in the following way

$$\begin{aligned} f(X,Y) &= fs(X,Y) + (1-f)b(X,Y) \\ &= f g_s(X) h_s(Y) + (1-f) g_b(X) h_b(Y). \end{aligned} \tag{3}$$

The signal model as a function of X, $g_s(X)$ is defined by the Crystal Ball function given in Eq. (1) but now truncated over the narrower ranges of X and Y. The signal model as a function of Y, $h_s(Y)$, (also truncated) is defined by an exponential decay with decay constant λ,

$$h_s(Y) = \lambda e^{-\lambda Y}. \tag{4}$$

The background model as a function of X, $g_b(X)$, is defined by a uniform distribution. The background model as a function of Y, $h_b(Y)$, is defined by a (truncated) normal distribution with mean μ_b and width σ_b.

Please note that the expectation is that all of the probability distributions described above are appropriately normalised in the region for which the random variables are valid. You can refer to Sec. 2.2.8 of the lectures notes for more details on truncated distributions.

The total model should consist of eight free parameters which you can assume have the following true values:

$$\begin{aligned} \mu &= 3, \\ \sigma &= 0.3, \\ \beta &= 1, \\ m &= 1.4, \\ f &= 0.6, \\ \lambda &= 0.3, \\ \mu_b &= 0, \\ \sigma_b &= 2.5. \end{aligned} \tag{5}$$

Write a small module in `python` which defines these p.d.f.s (you should make use of common libraries such as `scipy.stats` or `numba-stats` which provide analytic descriptions of the p.d.f.s, c.d.f.s and p.p.f.s for you) and demonstrate by means of a numeric integration that your definitions for $g_s(X)$, $h_s(Y)$, $g_b(X)$, $h_b(Y)$, $s(X,Y)$, $b(X,Y)$ and $f(X,Y)$ from Eq. (3) are properly normalised in the range $X \in [0,5], Y \in [0,10]$. In order to check your definitions are correct it may be worth checking the normalisation with different values of the parameters in Eq. 5. [4]

(c) Plot the one dimensional projections of these distributions in both the variables X and Y. Ensure that your plots contain the total pdf as well as a breakdown of the signal and background components. You should also make a two-dimensional plot of the joint probability density. [4]

(d) Generate a high-statistics sample from the joint distribution, containing a total of $100,000$ events, and then perform an extended maximum likelihood fit to estimate the nine parameters (notice there is an extra free parameter because the fit is extended). You should also determine an estimate of the uncertainty on these estimates. Please provide an evaluation of the execution time, using the `timeit` library, averaged over 100 calls for the following processes:

(i) Calling `np.random.normal(size=100000)` (this simply sets a standard benchmark for your machine)

(ii) The call which generates your sample of $100,000$ events

(iii) The call which performs the fit to the sample to estimate the parameters

Please present the last two numbers relative to the first [8]

(e) Now run a simulation study using parametric bootstrapping (with an ensemble of at least 250 samples) from the true probability distribution. You should trial sample sizes of 500, 1000, 2500, 5000 and 10000, including a Poisson variation on the sample size. Determine whether you see any bias on the λ parameter, describing the decay constant of the signal in Y, as a function of the sample size. Also determine the expected uncertainty on λ as a function of the sample size. [12]

(f) Using the samples produced above, perform an extended maximum likelihood fit in just the X variable and use it to produce *sWeights* which project out the signal density in Y. Then use an estimation method of your choice to determine the decay constant λ based on the weighted sample in Y. Compare the bias and the uncertainty to your findings from the previous part. [12]

(g) Compare and constrast these two methods. Explain the potential drawbacks and disadvantages of one over the other. State which you think is most appropriate and why. Explain in which scenarios one method might be preferred over the other. [8]

END OF PAPER