G. Cowan / RHUL Physics
Statistics Exercises
Summer 2023

Exercise on Maximum-Likelihood Fitting

Exercise 1: This exercise concerns maximum-likelihood fitting with the minimization program MINUIT using either its python implementation iminuit. The exercise is carried out by modifying and running the program `mlFit.py` (or the jupyter notebook `mlFit.ipynb`). These can be found on the web page

`https://www.pp.rhul.ac.uk/~cowan/stat/exercises/fitting/`

To use python on your own computer, you will need to install the package iminuit (should just work with "pip install iminuit"). See:

`https://pypi.org/project/iminuit/`

Alternatively, you can run the code in google colaboratory. Please see the school's website for further information on how to access and run the software.

The program provided generates a data sample of 200 values from a pdf that is a mixture of an exponential and a Gaussian:

$$f(x;\theta,\xi) = \theta \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} + (1-\theta)\frac{1}{\xi} e^{-x/\xi} \ , \tag{1}$$

The pdf is modified so as to be truncated on the interval $0 \leq x \leq x_{\rm max}$. The program Minuit is used to find the MLEs for the parameters θ and ξ, with the other parameters treated here as fixed. You can think of θ as representing the fraction of signal events in the sample (the Gaussian component), and the parameter ξ characterizes the shape of the background (exponential) component.

1(a) By default the program `mlFit.py` fixes the parameters μ and σ, and treats only θ and ξ as free. By running the program, obtain the following plots:

- the fitted pdf with the data;
- a "scan" plot of $-\ln L$ versus θ;
- a contour of $\ln L = \ln L_{\rm max} - 1/2$ in the (θ, ξ) plane;
- confidence regions in the (θ, ξ) plane with confidence levels 68.3% and 95%.

From the graph of $-\ln L$ versus θ, show that the standard deviation of $\hat{\theta}$ is the same as the value printed out by the program.

From the graph of $\ln L = \ln L_{\rm max} - 1/2$, show that the distances from the MLEs to the tangent lines to the contour give the same standard deviations $\sigma_{\hat{\theta}}$ and $\sigma_{\hat{\xi}}$ as printed out by the program.

1(b) Recall that the inverse of the covariance matrix variance of the maximum-likelihood estimators $V_{ij} = {\rm cov}[\hat{\theta}_i, \hat{\theta}_j]$ can be approximated in the large sample limit by

$$V_{ij}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right] = -\int \frac{\partial^2 \ln P(\mathbf{x}|\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} P(\mathbf{x}|\boldsymbol{\theta})\, d\mathbf{x} \,, \tag{2}$$

where here $\boldsymbol{\theta}$ represents the vector of all of the parameters. Show that V_{ij}^{-1} is proportional to the sample size n and thus show that the standard deviations of the MLEs of all of the parameters decrease as $1/\sqrt{n}$. (Hint: write down the general form of the likelihood for an i.i.d. sample: $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta})$. There is no need to use the specific $f(x;\boldsymbol{\theta})$ for this problem.)

1(c) By modifying the line

```
numVal = 200
```

rerun the program for a sample size of $n = 100, 400$ and 800 events, and find in each case the standard deviation of $\hat{\theta}$. Plot (or sketch) $\sigma_{\hat{\theta}}$ versus n for $n = 100, 200, 400, 800$ and comment on how this stands in relation to what you expect.

1(d) By modifying the line

```
parfix  = [False, True, True, False]           # change these to fix/free parameters
```

find $\hat{\theta}$ and its standard deviation $\sigma_{\hat{\theta}}$ in the following four cases:

- θ free, μ, σ, ξ fixed;
- θ and ξ free, μ, σ fixed;
- θ, ξ and σ free, μ fixed;
- θ, ξ, μ and σ all free.

Comment on how the standard deviation $\sigma_{\hat{\theta}}$ depends on the number of adjustable parameters in the fit.

1(e) Consider the case where θ and ξ are adjustable and σ and μ are fixed. Suppose that one has an independent estimate u of the parameter ξ in addition to the $n = 200$ values of x. Treat u as Gaussian distributed with a mean ξ and standard deviation $\sigma_u = 0.5$ and take the observed value $u = 5$. Find the log-likelihood function that includes both the primary measurements $(x_1, \ldots, x_n)$ and the auxiliary measurement u and modify the fitting program accordingly. Investigate how the uncertainties of the MLEs for θ and ξ are affected by including u.