Data Analysis Techniques

Problem Sheet 2: Maximum Likelihood and Least Squares Fitting

Submission deadline: 4pm Wednesday 6 November
Final submission opportunity: 4pm Thursday 7 November
Upload your solutions **as a single PDF file in portrait orientation** via the Canvas submission point for this assessment
Marked scripts will be made available via Canvas within two weeks

1. Straggling of charged particles in matter.

   In pion therapy, a monoenergetic beam of charged particles enters a homogeneous medium (the human body); they lose energy (by ionisation), and eventually come to a stop. The average distance they travel is called the *range,* λ_1. The actual stopping distances are distributed around this average, with a standard deviation called the *straggling coefficient,* λ_2. It is desirable to have all the particles stop in as small an interval as possible, so the energy deposit can be localised, e.g. in a tumour. Suppose that we do an experiment to measure λ_1 and λ_2.

   (a) Assuming the stopping depth is distributed as

   $$f\left(x;\lambda_1,\lambda_2\right) = \frac{1}{\lambda_2\sqrt{2\pi}} \exp\left\{-\frac{\left(x-\lambda_1\right)^2}{2\lambda_2^2}\right\},$$

   i.e., a normal distribution with mean $\mu = \lambda_1$ and variance $\sigma^2 = \lambda_2^2$, give an expression for the log likelihood l for the stopping distances of N incident particles. Show that the likelihood is maximised when λ_1 has a value λ_1^* given by

   $$\lambda_1^* = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

   as you would expect, and that λ_2 has a maximum-likelihood value λ_2^* given by

   $$\lambda_2^* = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i-\lambda_1^*\right)^2}{N}}.$$

   [6]

   (b) Since the expectation value of the estimator λ_1^* is equal to the expectation value of λ_1, it is unbiased. Is this also true of the estimator λ_2^*? (Hint: look at the variance). What happens as $N \to \infty$? [2]

   (c) From the second derivatives of l, find the Fisher information, or curvature, matrix $\underline{\underline{G}}$ defined by

   $$G_{ij} = \left(\left[\frac{-\partial^2 l}{\partial\lambda_i\partial\lambda_j}\right]_{\lambda=\lambda^*}\right).$$

   Hence show that the covariance matrix for the parameters λ_1^*, λ_2^* is

   $$\underline{\underline{V}} = \begin{pmatrix} \frac{\lambda_2^{*2}}{N} & 0 \\ 0 & \frac{\lambda_2^{*2}}{2N} \end{pmatrix}.$$

   From this, we see that the range and the straggling coefficient are uncorrelated, and that

   $$\begin{aligned} \sigma_{\lambda_1}^* &= \frac{\lambda_2^*}{\sqrt{N}} \\ \sigma_{\lambda_2}^* &= \frac{\lambda_2^*}{\sqrt{2N}}. \end{aligned}$$

   [6]

2. Suppose you are measuring a decay lifetime τ. You measure N decay times t_i, but you cannot measure any times longer than T. The probability function (normalised to 1) is then

$$\frac{e^{-t/\tau}}{\int_0^T e^{-t/\tau}dt} = \frac{1}{\tau}e^{-t/\tau}\left(1-e^{-T/\tau}\right)^{-1}.$$

By differentiating the log likelihood l and setting it to zero, show that the best estimator of the time is given by

$$\widehat{\tau} = \frac{1}{N}\sum t_i + \frac{Te^{-T/\widehat{\tau}}}{\left(1-e^{-T/\widehat{\tau}}\right)},$$

a nasty implicit equation which would have to be solved numerically. Notice that if the time T is sufficiently long, the second term becomes negligible, and the expected solution – the average of measured values – is found. [5]

3. Referring to slides 6–9 on least squares for definitions of variables, and using the given expressions for the variance and covariance of slope and intercept (which you need not prove), show that the error σ_{y_0} on the interpolated value of $y_0 = mx + c$ is given by

$$\sigma_{y_0}^2 = \frac{1}{S_1} + \frac{S_1}{\Delta}\left(x - \frac{S_x}{S_1}\right)^2.$$

[5]

4. Fit the following data set with a straight line, and find the gradient (with error), the intercept (with error), the full covariance matrix of gradient and intercept, the total χ^2, and the number of degrees of freedom. Would you regard this as a good fit? Plot the data along with your fit.

The data are available on the Canvas page as lin_fit.csv, a file of comma-separated values. Please show some of your calculation; if you do it on a computer, include your source code (do not using an existing fitting package).

x	y	σ_y
-4.06	-44.37	25.19
3.91	31.40	23.98
-5.70	-11.56	19.81
-9.76	-37.03	25.35
6.90	25.92	20.72
-14.09	-56.86	14.31
-13.31	-57.54	28.80
4.37	51.62	25.27
5.31	25.66	15.18
-13.25	-43.26	18.39
7.26	14.56	13.73

[12]

5. The most accurate way to find the average frequency ω of a sinusoidal signal is to find the phases ϕ_1, ϕ_2 at points t_1, t_2 near each end of the measurement period. Taking into account the number n of complete cycles between these points, the total phase difference is $\phi_2 - \phi_1 + 2\pi n$ in time $t = t_2 - t_1$, giving a frequency of $\omega = (\phi_2 - \phi_1 + 2\pi n)/t$. In reality, to find the phases ϕ_1 and ϕ_2, it is necessary to fit a small sample of data in each of the regions around $t = t_1$ and $t = t_2$ to sine curves.

Suppose, then, that we have such a data sample, consisting of a set of measurements $y_i \pm \sigma_i$ (taken at precisely-known times t_i) that we wish to fit to

$$y = A\cos(\omega t - \phi)$$

in order to extract the phase ϕ. The amplitude A is a parameter that will come out of the fit "for free". The frequency ω is known well enough for a first iteration by counting the number of whole cycles between the start and end of the measurement, and dividing by the measurement time.

(a) Rewrite the above expression in the form

$$y = \sum_k a_k f_k\left(x\right),$$

where $x = \omega t$, so that A and ϕ are within the parameters a_k, and the functions f are independent of them. (Hint: don't think too hard; there are only 2 terms.) We can then perform a linear least-squares fit. [2]

(b) Write down an exact expression for χ^2 in terms of a_1, a_2, x_i and the measured y_i with their errors σ_i. [2]

(c) Show that the phase $\phi \pm \sigma_\phi$ is given by

$$\begin{aligned} \phi &= \arctan\left(\Theta\right) \\ \sigma_\phi^2 &= \left(\frac{1}{1+\Theta^2}\right)^2 \cdot \sum_i \left(\frac{\partial\Theta}{\partial y_i}\right)^2 \sigma_i^2 \end{aligned}$$

where

$$\Theta = \frac{\sum_i y_i \sin x_i/\sigma_i^2 \cdot \sum_i \cos^2 x_i/\sigma_i^2 - \sum_i y_i \cos x_i/\sigma_i^2 \cdot \sum_i \sin x_i \cos x_i/\sigma_i^2}{\sum_i y_i \cos x_i/\sigma_i^2 \cdot \sum_i \sin^2 x_i/\sigma_i^2 - \sum_i y_i \sin x_i/\sigma_i^2 \cdot \sum_i \sin x_i \cos x_i/\sigma_i^2}.$$

[10]

This procedure is used in the neutron edm experiment to measure the Lamour precession frequency of polarised Hg atoms in a weak magnetic field. In this way changes in the field strength can be tracked at the nanogauss level — close to a billionth of the strength of Earth's field.

[:50]