Statistics, Data Analysis, and Machine Learning for Physicists
Tejaswi Nerella
Winter 2023

Homework 2

This homework will introduce you to issues in fitting for parameters given data, building on ideas that we talked about in class. Upload your solutions on Gauchospace by EOD Sunday 02-19-2023.

Problem 1: Modeling of Noisy Data

Suppose you are measuring two quantities (call them y and x) related by $y = ax$. This problem will explore how even well-behaved measurement errors can make it difficult to recover a.

a. First, you will generate some fake data from this model, with $a = 0.5$. This type of procedure is known as *Monte Carlo*, after the casinos, and is often a good place to start when you are trying to understand more complicated real data. Generate 10,000 values $\{z_i\}$ from a uniform distribution from -1 to 1. Pseudo-random numbers are an important research topic, and the subject of Chapter 7 in Numerical Recipes. Now generate your measured values x_i and y_i by adding independent Gaussian noise to both z_i and $0.5z_i$, respectively, with unit variances. Finally, make a scatter plot of your x and y values, and overplot the correct model $y = 0.5x$ for comparison.

b. Now use χ^2 to recover the value of a assuming a model $y = ax$. Write down the χ^2 expression for this model, taking into account the errors in y, but ignoring errors in x for now. Minimize this expression with respect to a and use it on your fake data to calculate the best-fit value of a and its standard deviation. Did you recover $a = 0.5$? By how many standard deviations are you off? What is the Gaussian probability of being off by at least this many standard deviations?

c. Numerical Recipes has a discussion on fitting data with errors in two coordinates (Section 15.3). Briefly, you define a new variable $w = y - ax$ with variance $\sigma_w^2 = \sigma_y^2 + a^2\sigma_x^2$. The sum of n independent w_i^2, which your data are, will therefore be χ^2 distributed. Write down the new χ^2 expression and minimize it with respect to a. The algebra is a bit tougher, but with $\sigma_x = \sigma_y = 1$, you can (and should) do everything analytically. What value of a do you recover? Don't worry about computing its standard deviation for now. For an extra challenge, see if you can show that this estimate is consistent, meaning that it gives the right answer for a very large number of data points (assuming you got your errors right!).

d. Now we will explore what happens if we don't get our errors right. Let's say, for example, that we overestimate our errors in x by 15% and underestimate our errors in y by 30%. Use the same fake data, but replace σ_x with $1.15\sigma_x$ and σ_y with $0.7\sigma_y$ in your χ^2 estimator. What value of a do you recover? What value of χ^2 per degree of freedom do you obtain? By this measure, is the model a good fit to the data?

People work very hard to understand their errors for reasons like this. Simple estimators for parameters tend not to be robust, in the sense that improperly accounting for even well-behaved measurement errors can introduce biases. Real data are often much worse than the example here, where a linear, one-parameter model describes the relationship between two quantities, and where our errors really are Gaussian and free of systematics.

e. Now create a new set of fake data from the same model, $y = 0.5x$. As before, draw 10,000 z_i from a uniform distribution from -1 to 1, and corresponding y_i by adding noise with unit variance to $0.5z_i$. This time, generate your x_i by adding noise to your z_i with variance proportional to $|z_i|$ (standard deviation proportional to $\sqrt{|z_i|}$). Thus, σ_x will be different for each point. Even worse, you can only estimate its value, since z_i (from which the error is derived) is not measurable! Write down χ^2 as in part c, first using

$$\sigma^2 = \sigma_y^2 + a^2\sigma_x^2|x_i| = 1 + a^2|x_i|$$

as your variance, and then using

$$\sigma^2 = \sigma_y^2 + a^2\sigma_x^2|z_i| = 1 + a^2|z_i|$$

(the true variance). Note that with real data, you would not be able to carry out the last case. Minimize each χ^2 expression with respect to a. Because you can't do this analytically, compute χ^2 on an appropriately spaced grid of a values and find the values of a that minimize χ^2. In more than one dimension, you would have to think more carefully about this minimization problem (Chapter 10 of Numerical Recipes). What best-fit values of a do you recover for each model? You should also compute the standard deviations of a by calculating the range in a over which $\chi^2(a) \leq \chi^2_{\rm min} + 1$. This range is equal to $2\sigma_a$. By how many standard deviations are you off from the true model in each case?

Problem 2: Cosmic Dawn

The EDGES experiment is searching for a signal of the cosmic dawn, when the first stars turned on. The expected signal is a small drop in the brightness of the radio sky at a frequency of around 80 MHz. The problem is that this drop is about 10^{-4} of the foreground at that frequency. The foreground is composed of emission by the sky and by synchrotron radiation from free electrons in the Galaxy. In this problem set, you will analyze data from the EDGES experiment yourself to constrain the cosmic dawn. The data file provided includes the sky temperature as a function of frequency. The original paper is available at `https://www.nature.com/articles/nature25792`.

Part 1

a. I will define ν to be the frequency divided by the central frequency of observations, i.e., 75 MHz; ν is then dimensionless. The published paper fit the sky temperature using the following model:

$$T[\nu] \approx a_0\nu^{-2.5} + a_1\nu^{-2.5}\log\nu + a_2\nu^{-2.5}(\log\nu)^2 + a_3\nu^{-4.5} + a_4\nu^{-2}. \quad (1)$$

Perform this fit yourself using least-squares with unit weights for each point (σ^2 is a constant independent of ν) and plot the residuals. This should match Figure 1 of the paper exactly.

b. The model that you fit in Part (a) is a linearization of a more physically motivated model,

$$T[\nu] = b_0\nu^{-2.5+b_1+b_2\log\nu}e^{-b_3\nu^{-2}} + b_4\nu^{-2}. \quad (2)$$

Perform the nonlinear fit using the optimization routine of your choice (check out `scipy.optimize.minimize` or `scipy.optimize.curve_fit`). Verify that you closely reproduce the residuals from Part (a).

c. It is often possible to make an optimization problem much, much simpler with a bit of cleverness, and Part (b) is a great example. Subtract $b_4\nu^{-2}$ from each side, multiply both sides by $\nu^{2.5}$, and take the logarithm. You should see that the fit is now linear in all variables except for b_4. To maintain the same weights for each point as before, you should set

$$\sigma_\nu^2 \propto \left(\frac{d\log\left[T - b_4\nu^{-2}\right]}{dT}\right)^2 = \frac{1}{(T - b_4\nu^{-2})^2}. \quad (3)$$

Use least-squares to optimize the solution at each trial b_4, and solve the 1-D nonlinear optimization using the method of your choice. Plot your residuals, and verify that they look almost the same as for Part (a) and, hopefully, identical to Part (b). Your fit should be very slightly better than you got for Part (a) as measured by χ^2.

Part 2

Light from the first stars is expected to cause a small dip in T around an unknown frequency, very roughly 80 MHz (don't worry about the basic physics reason why if you don't want to, but feel free to ask!). The published paper fit a flattened Gaussian and reported a central frequency of 78 MHz, a width of around 15 MHz, and an amplitude of 0.5 K. It reported a very high signal-to-noise ratio and small errors. In this part, you will reproduce this result and then do a series of more rigorous fits to judge its reliability.

a. Begin by adding the following function to your fit (not quite what is in the paper, but close and easier to write):

$$c_0 \exp\left[-\left|\frac{\nu - c_1}{c_2}\right|^5\right]. \tag{4}$$

For now you may stick with the polynomial model from Part (a); it will make the fitting a bit easier. Find the best-fit c_0 and c_1, and compare your amplitude with the published 0.53 K and 78 MHz (dimensionless $\nu = 1.04$). Also, compare your root-mean-square residuals with the published value of 25 mK. Hint: the model is linear in all but two parameters.

b. If you take the error on each measurement to be 25 mK (and uncorrelated), compute χ^2 before and after fitting this extra component. Then compute the best-fit χ^2 as a function of c_0, holding c_0 fixed at a range of possible values and varying all of the other parameters to optimize χ^2 at each trial c_0. The 1-σ interval is approximately the range of c_0 for which $\chi^2 \leq \chi^2_{\rm min} + 1$. What is the signal-to-noise of your detection using this definition to compute σ of c_0? Take as your definition of signal-to-noise the best-fit c_0 divided by its standard error, and compare to the value of 37 given in the paper (the model you are fitting is slightly different, so don't expect to get exactly the same answer). Note that the errors of 25 mK are just from setting $\chi^2_{\rm dof} = 1$.

c. Extra credit: repeat the previous part but drop the assumption that the 1σ confidence interval corresponds to $\Delta\chi^2 = 1$. Do this by computing the probability distribution of c_0 marginalized over the other seven model parameters. Hint: take fixed values for the nonlinear parameters c_1 and c_2. If the problem is linear in the remaining six parameters (including c_0), what is the likelihood for c_0 at these fixed c_1 and c_2 marginalized over the five a_i (what is its maximum value and what is its functional form)? You can then add these likelihoods for c_0 over a large grid of c_1 and c_2 (but not too large–this computation will be a little bit expensive) and, finally, normalize this marginalized likelihood to compute a new σ. Compare this to the value you got from $\Delta\chi^2 = 1$.

Part 3

There are a few serious problems in the preceding analysis. First, we assumed that the errors were uncorrelated between frequencies and simply determined errors by setting $\chi^2_{\rm dof} = 1$. Second, we ignored the fact that we chose a functional form of Equation (4) just to match the data; that parametrization wasn't physically motivated. Finally, the foreground model we fit was a linearization of a physically motivated model, and the coefficients you derived for the best-fit physical model in part (b) were unrealistic (again, don't worry about why). In this section you will begin to address these shortcomings.

a. First, we will deal with the fact that errors are likely to be correlated, for example, that there is some extra foreground or beam correction or something that isn't in the model but that is correlated bewteen nearby frequencies. The EDGES team has an earlier paper (`https://arxiv.org/pdf/1708.05817.pdf`), and from fitting the same model from Part 1a to a nearly identical experimental setup at other

frequencies (Figure 4 of that paper), we can estimate that the covariance might be approximately given by

$$\mathrm{Cov}(T_1, T_2) \approx A\delta_{12} + B\cos\left[\frac{0.3}{\mathrm{MHz}}(\nu_1 - \nu_2)\right]\exp\left[-\left|\frac{\nu_1 - \nu_2}{60\ \mathrm{MHz}}\right|\right], \tag{5}$$

where δ_{12} is the Dirac delta function; it is one if $\nu_1 = \nu_2$ and zero otherwise. This covariance implies that there is some sinusoidal correlation in the residuals (I do not have great confidence in this particular estimate of the covariance; getting this right is hard and important!). Note that I am now using ν to represent frequency in MHz rather than a normalized, dimensionless frequency. For the data set provided from the Nature paper, take $A = (25\,\mathrm{mK})^2$ and $B = 2A$ (this is roughly consistent with the earlier data set), and assume this estimate to match the true covariance. In other words, assume that Equation (5) is the true covariance with A and B given above. Construct the covariance matrix for the actual data under these assumptions. Note that in Part 2 we assumed $B = 0$.

b. Redo the fit from Part 1a minimizing

$$\chi^2 = (\boldsymbol{T} - \boldsymbol{m})^{\mathrm{T}}\,\boldsymbol{C^{-1}}\,(\boldsymbol{T} - \boldsymbol{m}) \tag{6}$$

rather than the diagonal form you used earlier, and recalculate χ^2. In this expression, $\boldsymbol{m}$ is the model, Equation (1), and $\boldsymbol{T}$ is the array of measured brightness temperatures. Next, redo Part 2 minimizing this expression for χ^2. It's still a linear problem in all of the parameters from Equation (1), so you should be able to do this mostly with matrix operations.

c. Now redo the fit from Part 2 with the full covariance matrix. Furthermore, let one additional parameter float: the exponent inside the exponential of Equation (4). To avoid letting the exponent be negative you can write it as $(\sqrt{c_3})^2$ and optimize $\sqrt{c_3}$. Your fit should end up being linear in all but *three* parameters now, so you'll want to combine nonlinear optimization in those three with a linear fit at each set of values of those three. What is your best-fit amplitude c_0 now, and what is its standard error (derived from $\chi^2 = \chi^2_{\mathrm{min}} + 1$)? How many sigma is the detection under these assumptions?

d. Note that the model still is not physical, and a physically motivated model would produce a vastly inferior fit to this one. Try an exponent of 2 in Equation (4), just for fun, if you have the time. This is probably what the team fit to their data first before they went looking for models that fit better (maybe there should be a trials factor penalty here that we haven't even assessed yet). What is the best-fit value of c_0 and what is its standard error, using the full covariance matrix for the data?