Statistics, Data Analysis, and Machine Learning for Physicists
Timothy Brandt
Spring 2020

Homework 1

This homework will introduce you to a few statistical concepts and analysis techniques, hopefully at a challenging level.

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 using `numpy.random.rand`. 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 z_i and $0.5z_i$, respectively (use `numpy.random.randn`, and use 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 extra credit, see if you can prove that this estimate is unbiased (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: Supernova Neutrinos

Most of the energy in a supernova explosion is released in the form of neutrinos. Supernova 1987A, which exploded in the Large Magellanic Cloud, was the nearest supernova to the Earth in modern times, and the Kamiokande II detector in Japan observed 12 neutrinos from this explosion. We will model the neutrino event rate, $R(t)$, at Kamiokande as a function of three terms: $t_{\rm SN}$, the time when the supernova went off, τ, the exponential decay time of the neutrino signal and F_0, the product of the fluence and the "effective" cross-section of the detector:

$$R(t) = \frac{F_0}{\tau} \exp\left[-\frac{t - t_{\rm SN}}{\tau}\right] \Theta(t - t_{\rm SN}) \tag{1}$$

where $\Theta(x)$ is the Heaviside function: $\Theta(x) = 0$ for $x < 0$ and $\Theta(x) = 1$ for $x \geq 0$. The Heaviside function encodes the information that the event rate is 0 before the explosion.

We detect N events from the supernova that arrive at time t_1, t_2, ..., t_N, where $t_1 < t_2 < ... < t_N$.

a. Write down the log of the likelihood function, the probability of the data given the model (parametrized by F_0, τ, and $t_{\rm SN}$), binning the data in bins of width Δt. [*Note that you could solve the rest of the problem using binned data, but binning is never optimal. The useful method here, letting the interval size go to zero and using Poisson statistics, ends up simplifying a lot of problems; I often use it.*]

b. Now take the limit as $\Delta t \to 0$ to get an unbinned log likelihood.

c. Determine analytic expressions for the maximum likelihood values of F_0, τ, and $t_{\rm SN}$ by differentiating.

d. This is a case where the likelihood function is very asymmetric. We can see this by computing the likelihood function as a function of $t_{\rm SN}$ and τ for realistic data. For this part of the problem, assume that the detector saw 12 events that arrived at the following times: 0, 0.1, 0.15, 0.3, 0.5, 0.9, 1.55, 1.7, 3, 5, 7 and 9.15 seconds, where time is measured relative to the first neutrino. Plot a contour plot for the likelihood function as a function of $t_{\rm SN}$ and τ; show the 1, 2, and 3σ contours. Compute these contours two ways: first use standard thresholds in $\Delta\chi^2 = -2\ln\mathcal{L}$, e.g., $\Delta\chi^2 = 2.3$ for 1σ. Second, determine contours of constant likelihood enclosing 68%, 95%, and 99.7% of the total, integrated likelihood. Explain why your contours differ, and which you should actually use.

e. Finally, marginalize (integrate) over F_0 and τ assuming uniform priors to obtain a one-dimensional probability distribution for $t_{\rm SN}$. Plot this distribution.