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

Lecture 4

In this lecture I'll start with a few tips and tricks that I didn't get to last time, and that are extremely useful to me in practice. I have tried to continually emphasize all of the assumptions that go into the derivations of the likelihood, χ^2 and subsequent analyses. I also want to continually caution you against overusing techniques and underthinking their applicability and assumptions. That said, the formulation of χ^2 and the likelihood that we went through last week remains incredibly useful. The fact that a linear model with Gaussian errors may be fit using linear algebra is especially remarkable. In your first homework, you have an example of an equation for χ^2 that cannot be solved analytically. In that case, you can use a grid-based approach to optimize and to search about the maximum likelihood/minimum χ^2. There is a whole literature about optimizing one-dimensional functions (like χ^2) that applies in all of these cases. Different sampling and optimization approaches work better in different cases. We will discuss some of them later in this course. For now, I want to emphasize that you should not just reach for a standard tool but should stop and think about your problem first.

Here is an example from my own work. Suppose that I have many measurements of the number of counts in a pixel, and suppose that my measurement errors are always approximately the same. To be explicit, I measure counts $\{x_i\}$ at times $\{t_i\}$, and my model is that

$$x_i = at_i + b + \delta_i \tag{1}$$

where a and b are unknown parameters to be fit and δ_i are Gaussian errors with known variance. I can write down χ^2 as

$$\chi^2 = \sum_i \frac{(x_i - at_i - b)^2}{\sigma_i^2}. \tag{2}$$

This is a linear model, and the best-fit a and its error may be derived analytically. Suppose, however, that a better model is nonlinear, for example

$$x_i = b + c_1(at_i) + c_2(at_i)^2 + c_3(at_i)^3 + \delta_i \tag{3}$$

where c_1, c_2, and c_3 are known parameters of the instrument, and I want to derive the best-fit quantity a. This model is nonlinear, and if I write down χ^2 it looks like I need to do a two-dimensional optimization. In my application, speed was critical, and a full 2D optimization wouldn't work. But look closer: if I fix a at some assumed value, the problem remains linear in b. **I can treat this as a one-dimensional nonlinear optimization, with the optimal b computed for each trial value of a.** In this case, it wouldn't matter very much even if I did have to optimize c_1, c_2, and c_3 for each value of a: the problem would still be linear in these four parameters. One-dimensional nonlinear optimizations are much, much easier than two-dimensional nonlinear optimizations, and this statement becomes even more true as the dimensionality increases. A decent initial guess plus fitting parabolas locally to χ^2 and iterating to the minimum gets convergence to machine precision almost as fast as if the whole problem were linear. Any time you can reduce the effective dimensionality of your problem like this, you can typically save a ton of computer time.

I now want to return to confidence intervals on parameters, and discuss a common way that they are calculated: contours of constant χ^2. Recall that we can approximate χ^2 about its minimum by the Taylor expansion. The first-order terms vanish because the minimum is a critical point of χ^2. If I truncate the Taylor expansion at second order, that is equivalent to approximating the likelihood about its peak by a multidimensional Gaussian. Taking $\boldsymbol{p}$ to be the array of parameters and $\tilde{\boldsymbol{p}}$ to be the best-fit parameters, we have

$$\chi^2(\boldsymbol{p}) \approx \chi^2(\tilde{\boldsymbol{p}}) + (\boldsymbol{p} - \tilde{\boldsymbol{p}})^T \boldsymbol{C}^{-1} (\boldsymbol{p} - \tilde{\boldsymbol{p}}) \tag{4}$$

where the covariance matrix $\boldsymbol{C}^{-1}$ is the inverse of one-half the Hessian. Again, remember that this has implicit assumptions:

- The errors on the data points are independent and Gaussian, which was used to generate χ^2 from the likelihood function.
- The likelihood function is approximately Gaussian within a few confidence intervals of its peak, and is negligibly small elsewhere.

If we're already approximating the likelihood function as a Gaussian about its peak, we can give thresholds of χ^2 that enclose a given fraction of the probability. Let's start with one dimension to see how this works, after first reviewing the relevant facts about the Gaussian distribution.

In one dimension, a Gaussian is given by

$$\frac{dp}{dx} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{5}$$

With this normalization the integral over all x is unity. I can also write down the fraction of the probability between $-\sigma$ and σ of μ:

$$\int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \approx 0.683. \tag{6}$$

These numbers are commonly tabulated: the fraction of a Gaussian's area between $\pm n\sigma$ of the mean. It is also available within python and other standard programming languages via the *error function.* The error function is defined as

$$\mathrm{erf}(x) \equiv \frac{1}{\sqrt{\pi}} \int_{-x}^{x} \exp\left[-t^2\right] = \frac{2}{\sqrt{\pi}} \int_{0}^{x} \exp\left[-t^2\right] \tag{7}$$

where the last equality is because of the symmetry of the Gaussian. The error function is closely related to the cumulative distribution function of the Gaussian:

$$\Phi(x) \equiv \int_{-\infty}^{x} p(x)\,\mathrm{d}x = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right]. \tag{8}$$

So, if I want the probability of a result at least as discrepant as x, I am after the quantity

$$\int_{-\infty}^{-x} p(x)\,\mathrm{d}x + \int_{x}^{\infty} p(x)\,\mathrm{d}x = 2\int_{x}^{\infty} p(x)\,\mathrm{d}x = 2\left(1 - \Phi(x)\right). \tag{9}$$

Equation (8) shows that I can write this probability as

$$2\left(1 - \Phi(x)\right) = 1 - \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right) = \mathrm{erfc}\left(\frac{x}{\sqrt{2}}\right) \tag{10}$$

where

$$\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) \tag{11}$$

is defined separately to avoid roundoff error when subtracting two numbers that are very close to one to get a number very close to zero. These functions are available in python as `scipy.special.erf` and `scipy.special.erfc`, and their inverses are available as `scipy.special.erfinv` and `scipy.special.erfcinv`.

So, in one dimension, for example, 99% of the time a value will be within $\sqrt{2}\,\mathrm{erfc}^{-1}(0.01) \approx 2.58\sigma$. And of course, 68.3% of the time a value will be within 1σ of the truth.

In multiple dimensions things are a bit more complicated. There are two main questions I can address in the case of n parameters, with $n > 1$:

1. What is the joint distribution of parameters within a given confidence interval?

2. What is the confidence range of a single parameter if I marginalize (integrate) over all of the other parameters?

Let's take the second question first. Earlier in this class, we discussed how in a model that is linear in some parameters but nonlinear in one, we can adopt a trial value of the nonlinear parameter (call it a) and minimize $\chi^2(a)$ with respect to all of the other parameters. In this way we can empirically compute a one-dimensional distribution $\chi^2(a)$. What if, instead of computing the optimal values for all of the other parameters given a, I had integrated over them? In the case where everything is Gaussian, these two procedures are equivalent. Take, for example, a two-dimensional Gaussian likelihood in parameters a and b:

$$\mathcal{L}(a,b) = \exp\left[-\left((a-\tilde{a})^2 C^{-1}_{aa} + 2(a-\tilde{a})(b-\tilde{b})C^{-1}_{ab} + (b-\tilde{b})^2 C^{-1}_{bb}\right)\right]. \tag{12}$$

In the equation above, C^{-1}_{aa} is a diagonal element of the inverse of the covariance matrix,

$$C^{-1}_{aa} = \frac{1}{2}\frac{\partial^2 \chi^2}{\partial a^2} = -\frac{\partial^2 \ln \mathcal{L}}{\partial a^2}, \tag{13}$$

so that Equation (12) is the (exponentiated) second-order Taylor expansion of the log likelihood. In one dimension, $C^{-1}_{aa} = 1/\sigma_a^2$, but this is not true in multiple dimensions unless the covariance matrix is diagonal. In other words, for a general covariance matrix,

$$C^{-1}_{ij} \neq \frac{1}{C_{ij}}. \tag{14}$$

For a diagonal covariance matrix, we would have

$$C_{ij} = C^{-1}_{ij} = 0 \text{ for } i \neq j \text{ and} \tag{15}$$

$$C_{ii} = \frac{1}{C^{-1}_{ii}}. \tag{16}$$

At fixed a, I can write Equation (12) as:

$$\mathcal{L}(a,b) = f(a)\exp\left[-\left((b-g(a))^2 C^{-1}_{bb}\right)\right] \tag{17}$$

where $f(a)$ depends only on a and $g(a)$ is chosen to complete the square at that value of a. The likelihood at this value of a marginalized over all values of b is then

$$\mathcal{L}(a) = f(a)\int_{-\infty}^{\infty} \exp\left[-\left((b-g(a))^2 C^{-1}_{bb}\right)\right]. \tag{18}$$

Now I really make use of my assumptions of Gaussianity. The integral is from $-\infty$ to ∞, so it doesn't matter what $g(a)$ is: the integral evaluates to $\sqrt{\pi/C^{-1}_{bb}}$. I then have

$$\mathcal{L}(a) = f(a)\int_{-\infty}^{\infty} \exp\left[-\left((b-g(a))^2 C^{-1}_{bb}\right)\right] \tag{19}$$

$$= f(a)\sqrt{\frac{\pi}{C^{-1}_{bb}}}. \tag{20}$$

The likelihood isn't normalized anyway, and $\sqrt{\pi/C^{-1}_{bb}}$ doesn't depend on a if the model is linear in b. So, if I just ignored the integral over b entirely, I would get the same answer as doing the integral! But what about $f(a)$? How do I calculate that? Well, (modulo that factor of $\sqrt{\pi/C^{-1}_{bb}}$) it is the likelihood for which $b = g(a)$. This sets the exponential term in the integral to unity, which is its maximum value. In other words, $g(a)$ is the value of b for which the likelihood at a is maximized, the best-fit b given a trial value of a: $\tilde{b}(a) = g(a)$. So setting $b = \tilde{b}(a)$ in the likelihood is equivalent to marginalizing over b at that value of a.

This argument generalizes to multiple dimensions. In that case, $g(a)$ becomes a function of many parameters, but it doesn't matter since the integral runs from $-\infty$ to ∞. The one potentially important difference is that instead of the integral evaluating to $\sqrt{\pi/C_{bb}^{-1}}$ as before, when there are many parameters (say, b, c, and d), it is

$$\mathrm{norm} = \sqrt{\pi^n/\det(\mathbf{C}^{-1})} = \sqrt{\pi^n \cdot \det(\mathbf{C})} \tag{21}$$

where $\mathbf{C}$ is the covariance matrix for the parameters being marginalized out, n is its dimensionality, and $\det(\mathbf{C})$ is its determinant. For a linear model, this covariance matrix still shouldn't depend on a. In that case, since your likelihood isn't normalized, it still doesn't matter whether you include it, and marginalizing over all of the linear parameters is equivalent to just plugging in their best-fit values.

So, just like $\Delta\chi^2 = 1$ gives the 1σ confidence intervals for a one-dimensional fit, $\Delta\chi^2 = 1$ also gives the 1σ confidence intervals for a single parameter within a many-dimensional fit, *provided that I always choose the other parameters to minimize χ^2 given all values of my parameter of interest.* While this particular statement is only true in the case where everything is Gaussian, I can in general integrate out parameters in my distribution to obtain a *marginalized* distribution, or a marginalized likelihood. This is a common and standard thing, and something that you will do in Problem 2 of the first homework.

Back to the first question: what is the joint distribution of parameters within a given confidence interval? I will define my confidence intervals as the area of parameter space enclosed within a surface of constant likelihood, or constant χ^2. In one dimension I can always, for example, define a 90% confidence interval such that 5% is excluded on the low end and 5% is excluded on the high end. I can't do this in multiple dimensions. I can't even conceive of how to define such a constrained region in two dimensions. However, a likelihood threshold such that the integral of the likelihood above this threshold is a given fraction of the total integrated likelihood is always well-defined. In the case of a multi-dimensional Gaussian, those likelihood thresholds can be computed and are often tabulated. For example, in two dimensions, 68.3% of the area of a Gaussian is above the threshold of 31.7% of its peak. In the language of χ^2, 68.3% of the χ^2 values satisfy $\chi^2 < 2.3 + \chi^2_{\rm min}$. You can look this up for other values, or compute it directly. If you compute it directly the easiest way is to use the χ^2 distribution itself with n degrees of freedom, where n is the number of parameters of interest. To do this in python, try the inverse of the survival function, `scipy.stats.chi2.isf`. For example, for the 90% confidence interval with 3 degrees of freedom, use

$$\Delta\chi^2 = \texttt{scipy.stats.chi2.isf}(0.1, 3) \approx 6.25. \tag{22}$$

So far, we've been assuming Gaussians left and right. Now I want to step back a bit. What do we really mean, in general, when we say we have a 2σ result, to use a very common terminology? For example, I measure the existence of the Higgs boson with 5σ significance. It's not the clearest wording, but I would argue that the most sensible interpretation is the following:

- My experimental results favor the existence of the Higgs boson at 5σ, in the (frequentist) sense that if the Higgs boson did *not* exist I would obtain such suggestive data as rarely as I obtain $\geq 5\sigma$ outliers in Gaussian statistics. This would happen in a fraction $\mathrm{erfc}(5/\sqrt{2}) \approx 6 \times 10^{-7}$ of hypothetical repeated experiments.

People often quote a result in terms of σ, but the identification of a number of σ with a false positive probability or a p-value is thoroughly identified with the Gaussian distribution. It's easy to get slippery about the definition of σ, but unless your distribution really is Gaussian, it's often better (and safer) to report confidence intervals, for example an interval containing 90% of your likelihood.

If you can't assume that everything is Gaussian, then the thing that you should really care about is the significance of your detection, not some magic quantity called σ. Be careful about the distinction: there have been plenty of claims in the literature of impossibly high significance measurements because σ took on a new meaning (i.e. the distributions ceased to be Gaussian). In frequentist statistics the significance is something like a p-value, the probability under the null hypothesis of getting a data set as suggestive as the one you observed. In Bayesian statistics it might be the fraction of the posterior probability distribution that covers the null hypothesis. The thing to do, then, is to decide what significance threshold or confidence interval you want (which may correspond to the significance corresponding to a given number of Gaussian σ) and then to integrate the likelihood function directly to find the threshold of constant likelihood containing a given fraction of the total likelihood. This can be a computationally expensive procedure, and integrating the likelihood or posterior probability is the subject of techniques that we will discuss later in the course.

Let's have a little bit of a philosophical debate on confidence intervals or credible regions. In Section 5.3.1 of your astrostatistics book, the authors define a *credible region estimate* (a, b) at a significance α such that

$$\int_{-\infty}^{a} f(t)\,\mathrm{d}t = \int_{b}^{\infty} f(t)\,\mathrm{d}t = \alpha/2. \tag{23}$$

Conceptually, this means that a fraction $1-\alpha$ is within your confidence interval, $\alpha/2$ is below your confidence interval, and $\alpha/2$ is above it. I actually don't like this convention, for two reasons:

- It does not handle endpoints well, and
- It does not generalize beyond one dimension.

For endpoints, suppose that we have an exponential distribution for a parameter,

$$p(t) = e^{-t} \tag{24}$$

for $t \geq 0$. If I adopt the definition above, *the maximum likelihood value is outside of any nontrivial credible interval.* It also introduces a huge distinction between model comparisons where t is forced to be zero and t is allowed to float (something that we'll discuss soon), and the simple endpoint of our distribution. Intuitively, it would be nice if these methods gave the same answer, or at the very least similar answers. For all of these reasons, I would favor the definition that we use for multiple dimensions: the confidence interval (a, b) at a threshold α is defined by

$$\int_{a}^{b} f(t)\,\mathrm{d}t = 1 - \alpha \tag{25}$$

such that

$$f(t) \geq f(x) \quad \forall t \in (a, b), x \notin (a, b) \tag{26}$$

Note that this has a different drawback: it behaves badly in the case of a multimodal distribution. However, I would argue that if your distribution is multimodal, you shouldn't be using a single confidence interval to describe it. Also, as a general rule, you should always be as explicit as possible about the numbers you are using to characterize your distribution.