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

Lecture 3

We'll begin this class with a lightning review of χ^2 as a maximum likelihood estimator: the assumptions that went into it, how something as simple as χ^2 came out, and how to derive confidence intervals. We will also go over another way of deriving confidence intervals that I did not discuss last time. First recall that I could use the assumption of independent, Gaussian errors on measurement values x_i to obtain the likelihood, the probability of the $\{x_i\}$ given an assumed value for y. This led me to a definition of χ^2,

$$\chi^2 = -2\ln\mathcal{L} + \text{constant} = \sum_i \frac{(x_i - y)^2}{\sigma_i^2}. \tag{1}$$

We found the maximum likelihood value of y by finding the critical point of χ^2, solving

$$\frac{d\chi^2}{dy} = 0. \tag{2}$$

The result turned out to be

$$y = \left(\sum_i \frac{x_i}{\sigma_i^2}\right)\left(\sum_i \frac{1}{\sigma_i^2}\right)^{-1}. \tag{3}$$

This procedure generalizes to fitting any linear model. The fully general case (Gaussian errors but with a model that is nonlinear in fitted parameters) is similar, but does not admit the same nice linear algebra solution. For a general *linear* model, each measurement x_i might be at a certain time t_i, or have certain conditions, or something like that. The model for the value I should measure might depend on all of these things. When I say I have a linear model, I mean that there are a number of conditions α_i, β_i, γ_i, etc. associated with each x_i, and that my model is

$$x_i = a\alpha_i + b\beta_i + c\gamma_i + \ldots + \delta_i \tag{4}$$

where δ_i, as before, is my measurement error. I can write down χ^2 (or, if you prefer, $-2\ln\mathcal{L}$) exactly as before:

$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - a\alpha_i - b\beta_i - c\gamma_i - \ldots)^2}{\sigma_i^2}. \tag{5}$$

Just as before, I can differentiate χ^2 with respect to each of the parameters of my model, a, b, c, etc.

$$\frac{\partial\chi^2}{\partial a} = -2\sum_{i=1}^{N} \frac{(x_i - a\alpha_i - b\beta_i - c\gamma_i - \ldots)\alpha_i}{\sigma_i^2}. \tag{6}$$

I would get something basically identical for all of the other partial derivatives. Something magical happened here: this derivative is linear in all of the parameters a, b, c, etc. The set of partial derivatives is nothing but a set of coupled linear equations, which I can solve by simple linear algebra. In practice, it is usually better to avoid inverting the matrix formed by the coefficients of the partial derivatives. For example, maybe the system of equations has many solutions (because, for example, $\alpha_i = 2\beta_i$, so that some of the parameters in my model are duplicates of one another). Matrix inversion is also vulnerable to round-off error. To avoid these, the standard method is to use the *singular value decomposition*. In python, it's especially easy using the function `numpy.linalg.lstsq`. This will be a topic for another class (soon!).

For now, we return to the simplest model, $x_i = y + \delta_i$. Last class, we found the confidence intervals on y by adding up the variance contributed by each measurement x_i,

$$\sigma_y^2 = \sum_{i=1}^{N} \left(\frac{dy}{dx_i}\right)^2 \sigma_i^2 = \left(\sum_{j=1}^{N} \frac{1}{\sigma_j^2}\right)^{-1}. \tag{7}$$

We can also derive this a different way, which will provide some insight. The idea is to approximate $\chi^2(y)$ by its Taylor expansion around the point of interest, in this case the maximum likelihood value of y, which I will denote by $\tilde{y}$. The Taylor expansion is given by

$$\chi^2(y) = \chi^2(\tilde{y}) + \left.\frac{d\chi^2}{dy}\right|_{\tilde{y}} (y - \tilde{y}) + \frac{1}{2}\left.\frac{d^2\chi^2}{dy^2}\right|_{\tilde{y}} (y - \tilde{y})^2 + \ldots \tag{8}$$

Now, if $\tilde{y}$ is the maximum likelihood value of y, the first derivative vanishes, and we have

$$\chi^2(y) \approx \chi^2(\tilde{y}) + \frac{1}{2}\left.\frac{d^2\chi^2}{dy^2}\right|_{\tilde{y}} (y - \tilde{y})^2. \tag{9}$$

Remember the relationship between the likelihood and χ^2? It was

$$\chi^2 = -2 \ln \mathcal{L} + \text{constant}, \tag{10}$$

or

$$\mathcal{L} \propto \exp\left[-\frac{\chi^2}{2}\right]. \tag{11}$$

The likelihood is not normalized, so this equation does not give you a probability. However, let's now assume that the first nonzero term in the Taylor expansion of χ^2 about its minimum provides a good approximation to the likelihood. In that case, the actual minimum value of χ^2 just multiplies the equation by a constant (it was a proportionality anyway). So, I can write

$$\mathcal{L} \propto \exp\left[-\frac{(y - \tilde{y})^2}{2}\left(\frac{1}{2}\left.\frac{d^2\chi^2}{dy^2}\right|_{\tilde{y}}\right)\right]. \tag{12}$$

Well, that derivative evaluated at the minimum χ^2 (maximum likelihood) is just a constant; it has some value. I can therefore simply define

$$\sigma_y^2 \equiv \left(\frac{1}{2}\left.\frac{d^2\chi^2}{dy^2}\right|_{\tilde{y}}\right)^{-1}. \tag{13}$$

Note that if I do this, the likelihood function is just a Gaussian with mean $\tilde{y}$ and variance σ_y^2 (and I can easily normalize it and interpret it as a probability). So, truncating the Taylor expansion of χ^2 at its second order term is equivalent to approximating the likelihood as a Gaussian near its peak. Let's first demonstrate that this approach gives the same result for our confidence interval on y that we got previously. We obtained our best-fit y by differentiating χ^2 with respect to y:

$$\frac{d\chi^2}{dy} = -2\sum_i \frac{x_i - y}{\sigma_i^2}. \tag{14}$$

It's easy to take one more derivative and multiply by $\frac{1}{2}$:

$$\frac{1}{\sigma_y^2} = \frac{1}{2}\frac{d^2\chi^2}{dy^2} = \sum_i \frac{1}{\sigma_i^2}. \tag{15}$$

I get the same result as before. Have you heard the rule that the confidence intervals for a parameter may be derived by setting χ^2 equal to its minimum value plus 1? In other words, if my best-fit χ^2 is 100, then

the range of possible values for the parameter y for which $\chi^2 \leq 101$ defines my 1-σ, or 68% confidence region? This is pretty standard, and it all comes from approximating the likelihood around its peak by a Gaussian. Equivalently, χ^2 gets approximated as a Taylor expansion throwing away all but the first and second order terms (the first-order terms vanish anyway at the maximum likelihood). There are similar rules for values to add to χ^2 where you are looking for simultaneous confidence intervals on more than one parameter. These still come from the Gaussian approximation, but in this case you can derive those values from the multivariate Gaussian. For example, with two parameters, $\Delta\chi^2 = 2.3$. For a 2-D Gaussian, 68% of the probability is within $\sqrt{2.3}\sigma$ (assuming σ to be the same in each dimension).

In two dimensions, remember that the Taylor expansion has cross terms. If I am fitting two parameters, a and b, and am expanding χ^2 about the point $(\tilde{a}, \tilde{b})$ which maximizes the likelihood, then the Taylor expansion is given by

$$\chi^2 \approx \chi^2(\tilde{a},\tilde{b}) + \frac{1}{2}\frac{\partial^2\chi^2}{\partial a^2}\bigg|_{\tilde{a},\tilde{b}}(a-\tilde{a})^2 + \frac{1}{2}\frac{\partial^2\chi^2}{\partial b^2}\bigg|_{\tilde{a},\tilde{b}}(b-\tilde{b})^2 + \frac{\partial^2\chi^2}{\partial a\partial b}\bigg|_{\tilde{a},\tilde{b}}(a-\tilde{a})(b-\tilde{b}). \tag{16}$$

The last term encodes the *covariance* between a and b. In general, a and b will not be independent: the parameter which is measured best is neither a nor b, but some linear combination of a and b. The likelihood function around the peak is approximately Gaussian but it is not necessarily symmetric about these two quantities. It is possible, however, to write down new quantities that are linear combinations of a and b and which do not have any covariance. To see this, let's write these derivatives as a matrix:

$$\chi^2 \approx \chi^2(\tilde{a},\tilde{b}) \begin{bmatrix} a-\tilde{a} & b-\tilde{b} \end{bmatrix} \begin{bmatrix} \frac{1}{2}\frac{\partial^2\chi^2}{\partial a^2}\big|_{\tilde{a},\tilde{b}} & \frac{1}{2}\frac{\partial^2\chi^2}{\partial a\partial b}\big|_{\tilde{a},\tilde{b}} \\ \frac{1}{2}\frac{\partial^2\chi^2}{\partial b\partial a}\big|_{\tilde{a},\tilde{b}} & \frac{1}{2}\frac{\partial^2\chi^2}{\partial b^2}\big|_{\tilde{a},\tilde{b}} \end{bmatrix} \begin{bmatrix} a-\tilde{a} \\ b-\tilde{b} \end{bmatrix}. \tag{17}$$

Remember that in one dimension the variance in my measurement of a parameter was the inverse of the second derivative of χ^2. That remains true here; the covariance matrix of my fitted parameters is the inverse of the matrix given above:

$$\boldsymbol{C}^{-1} = \begin{bmatrix} \frac{1}{2}\frac{\partial^2\chi^2}{\partial a^2}\big|_{\tilde{a},\tilde{b}} & \frac{1}{2}\frac{\partial^2\chi^2}{\partial a\partial b}\big|_{\tilde{a},\tilde{b}} \\ \frac{1}{2}\frac{\partial^2\chi^2}{\partial b\partial a}\big|_{\tilde{a},\tilde{b}} & \frac{1}{2}\frac{\partial^2\chi^2}{\partial b^2}\big|_{\tilde{a},\tilde{b}} \end{bmatrix} \tag{18}$$

Diagonalizing the covariance matrix will give the appropriate transformation to independently constrained parameters. This covariance matrix is closely related to the *Fisher information matrix*. In the Gaussian case (which we have been assuming all along), the Fisher matrix is the expected value of $\boldsymbol{C}$ for a proposed experiment, and it forecasts how well parameters and combinations of parameters will be constrained.

Keep in mind as you absorb all of this that all of the preceding analysis assumed the likelihood function to be locally Gaussian. In your first homework you will check this assumption explicitly for a case where the likelihood function is analytic.