Since I don't know of a good reference that lays out model-fitting etc. in astronomy the way I like to think about it, I've started writing up notes. They are currently a work in progress and will periodically be updated/extended.

The problem of "what model fits my data" is as old as experimental data [1]. An important category is when measurement errors have Gaussian distributions. Not only are Gaussians some of the easiest distributions to work with, but due to the central limit theorem, if we have enough data we'll end up at a Gaussian distribution anyways. The aim of this note is to show, starting from the probability distribution function (PDF) of a Gaussian, that we can compare the relative probabilities that different models would have given rise to our observed data.

1. From Gaussians to χ^2

If we have a Gaussian random variable with mean μ and standard deviation σ, then we want to know how likely we are that a realization of this variable gives us a value x. While the probability of a single value is zero for any continuous definition, we instead use the *probability density function* or PDF, where the probability of observing an event between x and $x + \delta_x$ is $\mathrm{PDF}(x)\delta_x$. Properly normalized PDFs should always be non-negative and integrate to unity. For a Gaussian, the PDF is:

$$\exp(-\frac{(x-\mu)^2}{2\sigma^2})/\sqrt{2\pi\sigma^2}$$

where the factor of $\sqrt{2\pi\sigma^2}$ is needed so the PDF integrates to one.

In traditional model fitting, we are usually interested in figuring out that the data "should have been," which mean we want to figure out μ. We also usually take the errors σ as given. While we can certainly try to estimate the errors at the same time as a model, but it's quite a bit more complicated and so we don't try to do that here. We also usually have many data points, which may have their own errors and expected means. One simple case would be fitting a line to data taken at different times t, where the expected value of each data point would be $at+b$. In this case, if we measure the i^{th} data point at time t_i, then we have $\mu_i = at_i + b$ for a slope a[2] and intercept b . Each data point would have its own value, taken at its own time, but if a line is a good description of the data, they should all agree on the values for the slope and intercept.

If we have many data points, with *uncorrelated* errors, then the probability density of observing several data values is just the product of the probability density of the individual data points. In otherwords, if d_i are our n data points, then the PDF is

$$\prod_{i=1}^{n} \exp(-\frac{(d_i-\mu_i)^2}{2\sigma_i^2})/\sqrt{2\pi\sigma_i^2}.$$

Of course, the product is rather awkward to work with, so let us instead work with the log of the

[1] citation needed

[2] We are using a here and not the usual m to avoid confusion down the line.

PDF:

$$\log \text{PDF} = \sum \frac{-(d_i - \mu_i)^2}{2\sigma_i^2} - \sum \frac{1}{2} \log(2\pi\sigma_i^2)$$

We can also pull a factor of $-\frac{1}{2}$ to the left hand side to further simplify matters. The PDF is also often referred to as the likelihood $\mathcal{L}$, so we make that change as well, giving:

$$-2\log(\mathcal{L}) = \sum \frac{(d_i - \mu_i)^2}{\sigma_i^2} + \sum \log(2\pi\sigma_i^2)$$

With this expression for the likelihood in hand, we can now say how likely it is that given a model μ_i and errors σ_i would have given rise to the observed data d_i. Of course, what we really want to ask is "given the data, what is the true model." Unfortunately, that question is rather ill-posed (see *e.g.* the long literature on Bayesian statistics). Instead, the question we know how to answer is "given a model, how likely is it to have given rise to the data we got." We can ask this question about many different models, and can compare the relative probabilities of two different models giving rise to the observed data. That is a well-defined question and the answer is just the ratio of the likelihoods/PDFs. In log space, the ratio becomes a difference, and we have for models μ_i and $\mu_i^\dagger$

$$-2\delta\log(\mathcal{L}) = \sum \frac{(x_i - \mu_i)^2}{\sigma_i^2} - \sum \frac{(x_i - \mu_i^\dagger)^2}{\sigma_i^2}$$

Happily, the second term in the likleihood has gone away, because it doesn't depend on the model and so gets cancelled in the difference. Note that if we ever try to fit for the σ's, though, we would need to keep that term around.

Looking at the likelihood difference expression, we can see that the full information we need to know for a given model is just the sum, which is called χ^2

$$\chi^2 \equiv \sum \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

Once I have the value for χ^2 for a set of models in-hand, I can quickly compare the relative likelihoods by noting that $-2\delta\log(\mathcal{L}) = \delta\chi^2$, which gives a likelihood ratio for the two models of $\exp(-\frac{\delta\chi^2}{2})$. If the difference between two models is large, the model with the larger χ^2 is exponentially less likely to have given arise to the observed data. Similarly, is $\delta\chi^2$ is small (where small means much less than 1), then the relative probabilities are very similar.

In short, we use χ^2 because for Gaussian random variables, everything we need to know about the relative probailities of two models producing the observed data is encapsulated in χ^2.

2. Searching for the "Best" Model

While it is useful to be able to compare two different models, usually what we want to do is find the model that does the "best" job fitting the data. In more general cases, even defining what we mean by this question can get quite complicated, and reasonable people can disagree. We can, however, get to an answer by defining the "best" model as the one that had the highest likelihood

of producing the observed data, which is usually described as the frequentist approach. Since we know that low values of χ^2 mean those models have relatively higher probabilities of producing our observed data, we can say that the "best" model is the one that minimizes χ^2. Very often, we have models that depend on one or more continuous parameters. One example of this that we have arleady seen is fitting a line to data, where we can set the slope and intercept to be whatever we want. The problem then becomes, of all the infinite possible values, how do we find the ones that minimize χ^2?

In searching for the parameter values that minimize χ^2, which from now on we will refer to as the "best-fit" values, we could directly differentiate χ^2 as we have written it, and possibly find a solution. However, it is usually simpler to understand what is going on, and how best to proceed, if we re-write χ^2 in the language of linear algebra. We can start by putting the observed data points and their expected values into column vectors d and μ. Furthermore, we can define a diagonal matrix N where $\mathrm{N}_{i,i} = \sigma_i^2$. In this case, N^{-1} is just trivially $\mathrm{N}_{i,i}^{-1} = 1/\sigma_i^2$. If we then take $\mathrm{N}^{-1}(d-\mu)$ we can see that this makes a column vector, with i^{th} element just $(d_i - \mu_i)/\sigma_i^2$. Recalling that a row vector times a column vector is just a dot product, we can take $(d-\mu)^T\left(\mathrm{N}^{-1}(d-\mu)\right)$. This becomes $\sum \frac{(d_i-\mu_i)^2}{\sigma_i^2}$, which was just our original expression for χ^2. So, in linear algebra notation, we have

$$\chi^2 = (d-\mu)^T \mathrm{N}^{-1}(d-\mu)$$

. Of course, we usually have some form for what the data values should be, and usually that depends on some set of model parameters. In this case, $\mu_i =< d_i >= A_i(m)$ where A_i is some function that depends on both the model parameters and which data point we're looking at. Again, in the case of fitting a line to data, we need to know both the slope/intercept *and* the x-value of a point in question before we can predict the y-value. Of course, we can as usual turn these things into vectors, giving us

$$\chi^2 = (d - A(m))^T \mathrm{N}^{-1}(d - A(m))$$

where if we knew the true model parameters, we'd have $< d >= A(m_{true})$. If we were arbitrarily good at math, we could differentiate χ^2 with respect to m, and find the global minimum, which would give us our best fit parameters. Unfortunately, this is usually very hard to do.

3. Linear Least Squares

Fortunately, there is a very broad class of models where we can analytically solve for our best-fit parameters. If we make the important simplification that we'll restrict ourselves to models that depend *linearly* on the model parameters, it turns out we can indeed solve for the global minimum of χ^2. What do we mean by a linear dependence? We mean

$$< d >= \mathrm{A}m$$

for a potentially arbitrary matrix A and fit parameters m. Generally, we decide on A, which corresponds to selecting which class of models we're going to look at, and then finding the parameters that give us the best model in that whole class. Going back to the line-fitting example where $< d_i >= at_i + b$, we could write A as an n-data by 2 matrix where the first column is t_i and the

second colum is just a vector of ones. Our fit parameters m (for model parameters) is then just a column vector consisting of the slope and intercept $m = [ab]$. If you carry one the multiplication Am you can see that the individual elements are $at_i + b$, so line-fitting fits nicely into this formalism. It's important to note that once we said we were going to fit a line to our data, that specified A. We could have instead said we'd fit a quadratic or cubic polynomial, or sum of exponentials, or sum of sines and cosines, and that each of the choices would have specified a different version of A. For now, we will not worry about how to select between different choices of A, but just worry about how to pick the best parameters once we have selected A.

We can rewrite χ^2 to reflect the fact we have restricted ourselves to the linear case:

$$\chi^2 = (d - \mathrm{A}m)^T \mathrm{N}^{-1} (d - \mathrm{A}m)$$

. We now have to find the values of m that minimize χ^2. As usual, we do this by taking the derivative of χ^2 with respect to m. As shown in the appendix, this is just

$$\nabla\chi^2 = -2\mathrm{A}^T \mathrm{N}^{-1} (d - \mathrm{A}m)$$

. At the minimum of χ^2, we know this must be equal to zero and so we have

$$-2\mathrm{A}^T \mathrm{N}^{-1} (d - \mathrm{A}m) = 0$$

or

$$\mathrm{A}^T \mathrm{N}^{-1} \mathrm{A}m = \mathrm{A}^T \mathrm{N}^{-1} d$$

.

Calculus with Linear Algebra': It is likely that many people reading this note have taken linear algebra at some point but perhaps not have done calculus with it. To first order, doing calculus with linear algebra is just like doing regular calculus, as long as one is careful not to switch the order of any matrices. For instance, the derivative of the linear least-squares equations $\chi^2 = (d - \mathrm{A}m)^T \mathrm{N}^{-1} (d - \mathrm{A}m)$ is just

$$\nabla\chi^2 = -\mathrm{A}^T \mathrm{N}^{-1} (d - \mathrm{A}m) + (d - \mathrm{A}m)^T \mathrm{N}^{-1} A$$

. We've played a bit loose here since we've ignored the difference between taking the derivative of m and m^T, but remember that underneath we can think about this as taking the derivative of χ^2 with respect to the individual m_i's, each of which is a scalar. Since the two terms in the sum are just conjugates of each other, the individual elements are the same and so we can just add their elements and pick one of the transposes to use. The usual choice is to then say

$$\nabla\chi^2 = -2\mathrm{A}^T \mathrm{N}^{-1} (d - \mathrm{A}m)$$

.