How to fit Data, and estimate parameter precisions.

Bahram Houchmandzadeh

January 30, 2019

1 Introduction.

Given a set of data (x_i, y_i), a major task of the experimentalist is to fit them into a model, *i.e.* find parameters $a, b, c, ...$ such that $f(x; a, b, c...)$ best describes the data. Of course, to do this, he has to have in mind *a model*. The first task of the experimentalist is to be reasonably sure that this model is good. We will not go so much in this direction and we will assume that the model is sound, even though what we'll say later can be extended to assess the goodness of fit. The second task is, once the model is assumed, to find the best set of parameters which minimize the difference between the model and the data. The third task is to estimate the uncertainties Δa, Δb, ... of these parameters. We are concerned here with the second and third task. This last one is often either neglected or made unnesserily difficult to understand.

What we'll say here will be mostly focused on *linear* regression, *i.e.* when parameters $a, b, ...$ appears linearly in the function ; for example $y = ax + b$ or $y = a\sin(x^2/2) + b\Gamma(\sqrt{x})$ are linear in their parameters a, b. The extension to non-linear fit, such as $y = \sin(ax^2/2)$ will be treated only approximately, but estimating their uncertainties follows the same rule.

2 A short reminder of quadratic functions.

Finding the best fit is mostly the art of manipulating quadratic forms, *i.e.* functions of the form $y = Ax^2 + 2Bx + C$. A small complication arises because we may have more than one dimension ; the general quadratic function read

$$y = \sum_{i,j=1}^{n} a_{ij} x_i x_j + 2\sum_{i=1}^{n} b_i x_i + c \tag{1}$$

Let us massage a little our one dimensional quadratic form $f(x) = Ax^2 + 2Bx + C$. Obviously, it can be put under the canonical form (see Fig. 1)

$$f(x) = A(x - x^*)^2 + f(x^*) \tag{2}$$

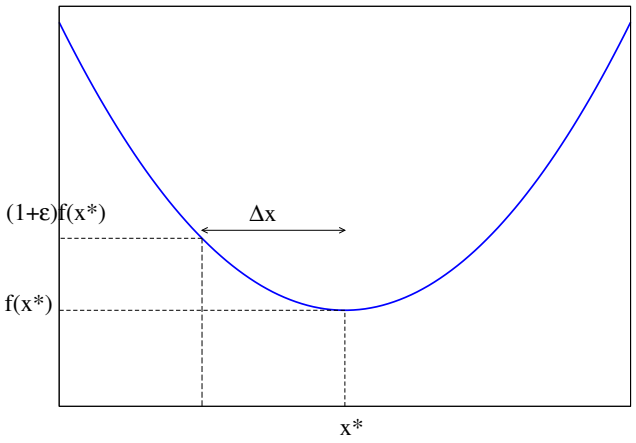

Figure 1: a simple quadratic function. Displacement Δx^* is such that it doubles the value of the function at its minimum.

where x^* is the minimum of the function f (we suppose $a > 0$). As $f'(x^*) = 0$,

$$x^* = -A^{-1}B. \tag{3}$$

Coefficient A is (twice) the curvature at x^*. Now, in order to get a better feeling of our quadratic form, we could ask : how much do we have to move away from x^* in order to multiply the value of our function by an $(1+\varepsilon)$ factor, compared to its minimum value ? The answer obviously is

$$\Delta x^2 = \varepsilon A^{-1} f(x^*) \tag{4}$$

Note that the answer combines both the curvature and the minimum value of the function.

Now, all these concepts generalize to higher dimension. All we have to do is to use nice notations. Instead of manipulating one by one each coordinate, we'll pack them into a vector x with component x_i . Our quadratic form (1) can now be written

$$f(x) = x^T Ax + 2x^T B + c \tag{5}$$

The T exponent denotes the transposition operation (line vectors become colon ones and vice versa) ; A is the matrix of component a_{ij} ; B the vector with component b_i. We assume matrix A to be symmetric (why?). We can now repeat the same operations and put the quadratic function in its canonical form (Fig **??**)

$$f(x) = (x - x^*)^T A(x - x^*) + f(x^*)$$

As before, the function f reaches its minimum for the vector x^*; this is equivalent to say that $\partial f/\partial x_i = 0$ for all i. In matrix notation, we'll say $\partial f/\partial x = 0$[1], *i.e.*

$$2Ax^* + 2B = 0$$

[1]The notion of derivation is not limited to scalars, and can be extended to much more complex objects. Let say

or, equivalently

$$x^* = -A^{-1}B$$

where A^{-1} is the reverse of matrix A (curvature at the minimum). You see how nicely matrix notations extend our knowledge of simple quadratic forms. Again, we can ask how much do we have to move away from x^* to multiply f by $(1+\varepsilon)$? The answer is not a scalar, because we have the choice of the direction : we can move only in the $x_1$direction, or first in $x_1$and then x_2, ... If we form the matrix $\Delta x \Delta x^T$(which is square , $n \times n$ and whose elements are $\Delta x_i \Delta x_j$), the equation $(x-x^*)^T A(x-x^*) + f(x^*) = (1+\varepsilon) f(x^*)$ solves into

$$\Delta x \Delta x^T = \varepsilon A^{-1} f(x^*) \tag{6}$$

(recall that $f(x^*)$ is a scalar).

OK, we are know well equipped to do serious business and fit our data.

3 Linear Regression.

Suppose we have a set of data (x_i, y_i) and we want to model them with the function $y = f(x; a, b, c, ...)$. As we said before, we are interested in linear regression, *i.e.* when parameters $a, b, c, ...$ enter the function linearly. For example, $y = ax + b$ or $y = ax^2 + bx + c$, or $y = a \sin x + b \cos x$. In all these examples, the function is linear for the parameters, even though it can be non-linear for the independent variable x. From here on, we will use the straight line $y = ax + b$ as the model, but all other linear regressions are similar and can be deduced from it.

Estimating the parameters.

Our desire is to get the line $y = ax + b$ as close to the data as possible, by wisely choosing a and b (Fig 2). For each point, the difference between the data y_i and where we expected the data $ax_i + b$ is $e_i = ax_i + b - y_i$. In order to minimize the gap between data and model, we can choose to minimize

$$\chi^2(a,b) = \sum_{i=1}^{N} (ax_i + b - y_i)^2 \tag{7}$$

The function χ^2 is the sum of the square of residual[2]. We could have used an other evaluation function, like the sum of the absolute value of the residuals, or the maximum residual or ... But

---

we have a function $f(x)$, where x can be a scalar, vector, matrix, and so on (it can even be a function) and same thing for f itself. If we can compute small variation of the function as a *linear functions* of small variations of the variable, then we say we have a derivative :

$$f(x+dx) = f(x) + L.dx + O(dx^2)$$

Obviously, we suppose that in the space in which x is defined, we possess operations such as additions and multiplications and we can define the distance between two points. Now, it is not difficult, by writing $f(x+dx)$ in expression (5) and extracting the linear terms in dx to see that the derivative is the one we have given.

[2]Why $N-2$ and not N ? Normally, you have many data points, and the difference between N and $N-2$ is very small. We will see however that there is a small advantage at using $N-2$. For practical purpose of evaluating a, b, the prefactor has no importance.

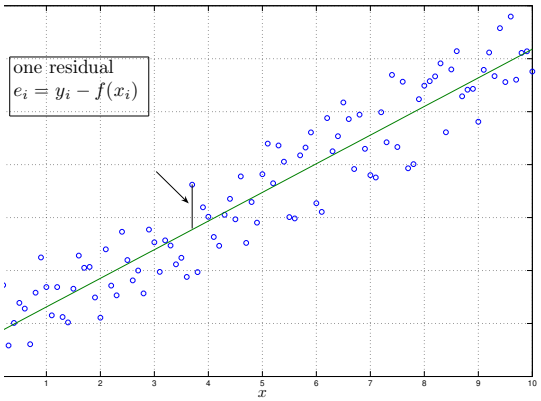

Figure 2: Example of data fitted to a line.

χ^2 has very good analytical behavior and is the most popular metric. As you see, χ^2 is a function of the parameters of the straight line a,b ; choosing wisely a,b means finding the point (a^*,b^*) which minimizes χ^2. Expanding the parentheses of (7), we get

$$\begin{aligned}\chi^2(a,b) &= \left(\sum_{i=1}^{N} x_i^2\right) a^2 + Nb^2 + 2\left(\sum_{i=1}^{N} x_i\right) ab \\ &- 2\left(\sum_{i=1}^{N} x_i y_i\right) a - 2\left(\sum_{i=1}^{N} y_i\right) b + \left(\sum_{i=1}^{N} y_i^2\right)\end{aligned}$$

All the expressions we have written between parentheses are computed by summing various combination of our data point and therefore are known. Let us rename them : $S_{xx} = \sum_{i=1}^{n} x_i^2$; $S_x = \sum_{i=1}^{n} x_i$; $S_{xy} = \sum_{i=1}^{n} x_i y_i$; $S_y = \sum_{i=1}^{n} y_i$; $S_{yy} = \sum_{i=1}^{n} y_i^2$. Rewriting in these new notation,

$$\chi^2(a,b) = S_{xx}a^2 + Nb^2 + 2S_x ab - 2S_{xy}a - 2S_y b + S_{yy} \tag{8}$$

which is a quadratic form in a and b. Noting $u = (a,b)^T$,

$$A = \begin{pmatrix} S_{xx} & S_x \\ S_x & N \end{pmatrix} \tag{9}$$

and

$$B^T = -(S_{xy}, S_y) \tag{10}$$

We can use the matrix notation

$$\chi^2(u) = u^T A u + 2u^T B + S_{yy} \tag{11}$$

The best fit (the vector u which minimizes χ^2 is therefore given by solving the linear system $Au^* = -B$. For a linear regression, the inverse matrix computation is trivial

$$A^{-1} = \frac{1}{\Delta}\begin{pmatrix} N & -S_x \\ -S_x & S_{xx} \end{pmatrix} \tag{12}$$

where $\Delta = NS_{xx} - (S_x)^2$ is the determinant of the matrix A. The explicit solution is

$$a^* = \frac{NS_{xy} - S_x S_y}{\Delta} \quad (13)$$

$$b^* = \frac{S_{xx} S_y - S_x S_{xy}}{\Delta} \quad (14)$$

If you have more parameters, say $(a_1, a_2, ...a_n)$, matrix A and vector B get bigger ($n \times n$ matrix and n vector), and the computation of their elements get longer ; but the derivation process follows exactly the same rules and by the end, you have to solve a linear system $Au^* = -B$ where the vector $u^* = (a_1^*, ...a_n^*)$ is the best estimator of your parameters. Big ($n > 2$) linear systems are not solved by computing A^{-1} and then multiplying it by B, this would be too costly. They are instead directly solved by using such methods as the Gaussian elimination or other means. Any basic mathematics's package on your computer will perform such a task[3]

What is an average and its uncertainty ?

Let us forget for a moment all the fitting buiseness and come back to a much simpler question. Suppose you have a collection of N data y_i ; for example, you have made many measurments of your weight in the morning, but your brand new scale gives you each time a slightly different result. What is the best estimation of the *true* value ? Well, obviously, you would use the average $\bar{y}$ of your measurements as a good estimator of the true value μ :

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \quad (15)$$

In fact, you could have asked the follwing : how can I choose y in order to minimize

$$\chi^2(y) = \sum_{i=1}^{N} (y - y_i)^2 \quad (16)$$

Of course, minimizing χ^2 with respect to y would give you $y = \bar{y}$, the result (15). As you see, the numerical average (15) is precisely the value which minimizes the residual errors. Finding the average is a fit with an horzontal line.

Now, what confidence could you have in your estimation ? Is it 75 ± 5kg or 75.3 ± 0.1 ? You can see the scale as a random variable generator : each time you measure youre weight, the scale produces the number $\mu + \delta_i$. μ is the true value and δ_i is a random variable with average 0 and variance σ^2 (which for the moment you ignore). How can you estimate σ^2 by just looking at your data ? As you may know from probability lectures, the best estimator for σ^2 is the variance of data : *once* you have determined the best estimation of the true value μ from eq.(15), the best estimation of the true variance σ^2 is

$$V = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2 \quad (17)$$

[3] a very (very) good numerical math package is the open source julia.

But we are not finished about the confidence in $\bar{y}$. How does the variance of $\bar{y}$ relates to the variance of the scale ? The answer is[4]

$$\Delta\bar{y}^2 = \frac{1}{N}\sigma^2 = \frac{1}{N(N-1)}\sum_{i=1}^{N}(y_i - \bar{y})^2 \tag{18}$$

The more data point you have, the smaller the uncertainties of your average.

We can formulate the confidence problem still another way. Expanding the parenthesis of eq.(16), we can write

$$\chi^2(y) = Ay^2 + 2By + C$$

where $A = N$, $B = -\sum_i y_i$ and $C = \sum_i y_i^2$. The y-value which minimizes $\chi^2(y)$ is $\bar{y} = -A^{-1}B$ as we said in eq.(3) and is just the average given by eq.(15). Now we can ask : How much Δy do we have to move away from $\bar{y}$ in order to multiply the minimum χ^2 by $(1 + 1/(N-1))$. In other words, how can we choose Δy to get a relative $1/(N-1)$ increase in χ^2 ?

$$\frac{\chi^2(\bar{y}+\Delta y) - \chi^2(\bar{y})}{\chi^2(\bar{y})} = \frac{1}{N-1}$$

The answer, from what we said in the introduction (eq. 4) is

$$(\Delta y)^2 = \frac{1}{N-1}A^{-1}\chi^2(\bar{y}) = \frac{1}{N(N-1)}\sum_{i=1}^{N}(y_i - \bar{y})^2 \tag{19}$$

which is exactly the uncertainty of the average as we saw in eq.(18). Let us reformulate that : $N-1$ is the degree of freedom of our sample ; we have N data points, but have already computed the average from them, so we are let with $N-1$ effective data points. The uncertainty in the estimation of the fit parameter $\bar{y}$ is a Δy which make a relative increase of $1/(N-1)$ in the minimum χ^2.

Estimating the uncertainties of the parameters.

Well, now we can come back to the fit our data $\{x_i, y_i\}$ by a straight line $y = ax + b$ and ask the question of the confidence we can have in their estimation. What we said above can be repeated word by word. Once we have determined $u^* = (a^*, b^*)$ which minimzes the χ^2, how much do we have to move away in order to make a relative increase of $1/(N-2)$ in χ^2 ? We already know, from eq.(6) that the answer is

$$\Delta u \Delta u^T = \frac{1}{N-2}\chi^2(a^*, b^*)A^{-1} \tag{20}$$

[4]This result can be obtained easily without a call to the central limit theorem. Note that for independent random variables X, Y, $Var(aX) = a^2 Var(X)$and $Var(X+Y) = var(X) + Var(Y)$. thus, if a random variable Y has variance σ^2, its average over N realisation

$$\bar{Y}_N = \frac{1}{N}\sum Y_i$$

has a variance of σ^2/N.

or, explicitly,

$$\begin{aligned} \Delta a^2 &= (N/\Delta)\chi^2(a^*,b^*)/(N-2) \\ \Delta b^2 &= (S_{xx}/\Delta)\chi^2(a^*,b^*)/(N-2) \\ \Delta a \Delta b &= -(S_x/\Delta)\chi^2(a^*,b^*)/(N-2) \end{aligned}$$

It is not hard to show that as computed above, the uncertainties are good estimator of the variances and covariances of the parameters (see appendix 4.2).Well, that's it, we are done ! Note that $\Delta \sim N^2$ and thus $\Delta a^2 \sim 1/N$: the more data you have, the better your fit, and the variance decreases as $1/N$ as it should (remember central limit theorem ?).

The principle of this example can be easily extended to any other linear regression. The more parameters you get, the more cumbersome is it to solve the system $Au = -B$. For the computer however, this is no burden except if you have more than 1000 parameters, and in this case, you should really worry about the relevance of your model.[5]

4 Important notes.

4.1 Goodness of fit.

We did not say anything about the goodness of fit : Is it reasonable to use the function $f(x;a,b,c,...)$ to fit the data ? And how good is the fit anyway ? A first answer is, well, look at the $\Delta a/a$. If it is much smaller than unity, you can have *some* confidence in your model. If not, you may think of something else , but don't throw your model quickly, (see below).

Many experimentalists use this answer ; even better, and this a well kept secret, many even don't look at the uncertainties, but at the plot of the data and the model : if they stick reasonably well, OK, the fit is not bad.

This answer will shock *serious* experimentalists. By serious experimentalists I mean people who don't have tons of data and have to assess their goodness of fit (GoF) very seriously.

The first thing you must have in order to evaluate the GoF is to know the source of uncertainties in your measurements and be able to reliably note $y_i \pm \Delta y_i$. The knowledge of Δy_i gives you the scale against which evaluate the fit.

Now, because we are dealing with probabilities, nothing is *sure*, we can just give a probability for an event to occur. If you play National Lottery and win the big prize twice in a row, other people may assume that something was wrong. There is *some* probability for you to win twice in a row ; other people however won't accept such rare event, because they have put their acceptance threshold for "not suspicious" at say, 10^{-3}: the probability for *somebody* to win is around 1 (one person in average wins the lottery each week), the probability for a *specific individual* to win is around $10^{-6} - 10^{-8}$.

The acceptance threshold is human affair. Let us put it in more probabilistic terms. Suppose you suspect your random variable X (your measurements) to have a normal distribution of width

[5] In fact, you should not wait for 1000 and begin to worry if you have more than, say, 4. As the saying goes, give me 10 parameters and I fit an elephant ; give me 12 and I make it dance !

σ, *i.e.* to have a distribution law

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right).$$

Now, you make one realization (one measurement) and measure $x = 10.1\sigma$. Is it valid to suppose that your random variable X follow a normal distribution ? The probability for one measurement to fall outside the $[-10\sigma, 10\sigma]$ window is 10^{-23}; if you had made 10^{23} measurements, it would not seem suspicious to have one data point so far. But making just one measurement and find it so far ? May be it is wiser to revise our supposition that X follow the above normal law.

Evaluating the goodness of fit is just that. Knowing the uncertainties of your measurements, what is the probability to obtain $\chi^2(a^*, b^*)$ in the range you measure ? Obviously, if $\chi^2(a^*, b^*)$ is much higher than $\sum_i (\Delta y_i)^2$, something is wrong[6]. How much wrong ? The answer is not more difficult than what we wrote above. It will take us however farther that this short paper intended to, and we will have to visit some probability law such as, surprise surprise, the χ^2_k law. So, we refer the reader to more advanced text on that.

Let us come back to the case of large $\Delta a/a$. Can that rule out your model ? Not necessarily. If the uncertainties of your data point are large, even if your model were good, you'll obtain large $\Delta a/a$. But look the other way around : if $\Delta a/a \ll 1$, then you can have some confidence that your model stick well to your data.

4.2 Estimating Uncertainties.

Giving the uncertainties of the parameters as we did in eq.(20) is like providing a recipe for cooking. Let us see here why this is indeed a good estimator. Suppose the random variable Y is related to X by

$$Y = aX + b + \delta \tag{21}$$

where δ is a centered random variable : $\langle \delta \rangle = 0$, $\langle \delta^2 \rangle = \sigma^2$. We do not suppose that X is random, this can be added at a small cost by the reader.

Now, we have a set of measurements $\{x_i, y_i\}$ and we are convinced that they follow the model (21). How do we best estimate parameters $p = (a, b)^T$? Let us call $u = (\alpha, \beta)^T$ our estimation of p and define

$$u = A^{-1}B$$

where A, B are the same matrix and vector as we defined in the preceding section of the text. Note that A is given only in term of S_{xx}, S_x and N, so there is nothing random in it. On the other hand,

$$S_{xy} \quad = \quad \sum x_i y_i$$

[6] We are supposing here that all points have the same uncertainties. If this is not the case, we have to add the uncertainties as weights in the evaluation function

$$\chi^2 = \sum_i (f(x_i; a, b, c, ...) - y_i)^2 / \Delta y_i^2$$

when looking for the best set of parameters.

$$\begin{aligned} &= \sum x_i(ax_i + b + \delta_i) \\ &= aS_{xx} + bS_x + \sum x_i\delta_i \end{aligned}$$

so the only random part of S_{xy} is $\sum x_i\delta_i$. By the same token,

$$\begin{aligned} S_y &= \sum y_i \\ &= \sum (ax_i + b + \delta_i) \\ &= aS_x + Nb + \sum \delta_i \end{aligned}$$

Therefore, B can be written as

$$\begin{aligned} B &= \begin{pmatrix} aS_{xx} + bS_x + \sum x_i\delta_i \\ aS_x + Nb + \sum \delta_i \end{pmatrix} \\ &= Ap + Z \end{aligned}$$

where all the randomness is captured in the vector $Z = (\sum x_i\delta_i, \sum \delta_i)^T$. Therefore, our estimation u is simply

$$u = A^{-1}B = p + A^{-1}Z.$$

Note that $\langle Z \rangle = 0$, so the expectation for u is

$$\langle u \rangle = p$$

This was expected. What is now the variance of u ?

$$\begin{aligned} Var(u) &= \left\langle uu^T \right\rangle - \langle u \rangle \left\langle u^T \right\rangle \\ &= \left(A^{-1}\right)^2 \left\langle ZZ^T \right\rangle \end{aligned}$$

It is very easy to show that $\left\langle ZZ^T \right\rangle = \sigma^2 A$: for example, $\left\langle (\sum x_i\delta_i)^2 \right\rangle = \sum x_i^2\sigma^2 = \sigma^2 S_{xx}$ and so on. As we said, a good estimator of σ^2 is $\chi^2(u)/(N-2)$ and hence our formula for the uncertainties

$$Var(u) = \frac{1}{N-2}\chi^2(a^*, b^*)A^{-1}$$

Don't be mystified by all these matrix vector multiplications. This formula is just a generalization of the "average random variable" : if $Y = (1/N)\sum X_i$, then its variance is just the variance of the original variable X divided by N

$$Var(Y) = \frac{1}{N}Var(X)$$

Here, A^{-1} plays the role of $(1/N)$, where $\chi^2(u)/(N-2)$ is the best estimation we have for the variance of original data.

4.3 Non-linear curve fitting.

The principle of non-linear curve fitting is not different from what we said above. As an example, think of the exponential function $f(x) = a + b\exp(x/c)$. Given the data set (x_i, y_i) $(1 \leq i \leq N)$ and the function $f(x; a_k)$ we want to find the best set of parameters a_k^* $(1 \leq k \leq K)$ which minimizes the evaluation function

$$\chi^2(a_k) = \sum_{i=1}^{N} (f(x_i; a_k) - y_i)^2$$

As before, χ^2 is a function of K variables, but this time, it is not quadratic anymore. The task of finding the minimum is therefore slightly more complicated and is carried out numerically by iterative methods ; there are a number of methods to do that, the most widely used is called Levenberg-Marquardt.

What we said above generalizes to the computation of uncertainties. Their matrix is given as before by

$$A^{-1}\chi^2(a_k^*)$$

where A is the curvature matrix at the minimum :

$$A_{ij} = \frac{\partial^2 \chi^2}{\partial a_i a_j}$$

Usually, the same program that find the minimum also compute numerically the curvature matrix and its inverse.