Some notes on fitting data

Markus Deserno
Max-Planck-Institut für Polymerforschung, Ackermannweg 10, 55128 Mainz, Germany
(Dated: September 21, 2004)

Fitting some functional form to some set of data can be a more subtle exercise than one normally expects. These notes illustrate problems and possible solutions for a few typical cases.

I. POWER LAWS

Power laws are very often found in measured data. Some variable y is proportional to some power of another variable x. Such a relation is best visualized by plotting y versus x on a doubly logarithmic scale, since then one will see a "straight" line, and its slope indicates the scaling exponent.

Fig. 1 gives an example of such data. The points follow the power law $y = 2\,x^3$, but there's additional noise. It would seem the most straightforward thing to fit these data to a functional form $f_1(x) = a\,x^b$ and then determine the scaling exponent b from this fit. This, however, can go badly wrong: The dashed line in Fig. 1 has been obtained by precisely this prescription, and it is evidently so much off the data that there's not even a point in looking at the fitted values of a and b.

What has gone wrong? Was the fitting routine bad? Not so. The situation is more subtle and has something to do with the precise nature of the *noise* which is present in the data. If we look a bit more closely at the figure, we see that over the entire range the data points scatter evenly about the average power law trend. Notice, however, that the vertical axis is also on a logarithmic scale! Some particular deviation on the right hand side of the figure (say, one centimeter off the average trend line), corresponds to a far bigger *absolute* difference between data point and trend than the same deviation on the left hand side of the figure. But standard fitting routines minimize χ^2, the mean square average of the absolute deviations! Consequently, χ^2 will be vastly dominated by whatever goes on at the right end of the figure, while the entire left part is completely irrelevant, even though the power law trend continues all the way through. One might object that since even the points at the right part of the figure follow the power law (well, at least in this case!), we ought to get basically the same fitting function as if we had "properly" managed to fit to all data points. This objections overlooks that if one is fitting to far less relevant points, the statistical noise is of course going to be far more relevant, and this might well lead to a fit which outside the sensitive range is completely off – as it did here.

Once one has realized this problem, it becomes quite immediately clear what has to be done to resolve it: If a standard fitting routine minimizes the mean square average of the absolute deviations, we should give it the data in such a form that the scatter about the trend assumed by the fit is in fact even. Looking back at Fig. 1 we see that evidently the logarithm of the y values scatters in that sense evenly. Hence, we're far better off to fit a functional form $f_2(x) = a + b\log x$ to the *logarithm* of the data points. Indeed, this gives the solid line in Fig. 1, which not only reproduces the trend very well, the particular value of b obtained in this case, $b \approx 2.87$, is rather close to the value of the exponent which underlies this set of data, namely 3.

Notice that the nature of the noise is something which pertains to the *data*, so it has to be looked at in each individual case. Thus, whether the noise is additive (like one usually assumes) or multiplicative (as in this case) has to be checked first. On the other hand, if one has power law data, then there is usually some deeper physical reason behind that, and the same reason often makes the noise multiplicative, and thus the situation comparable to the case studied here.

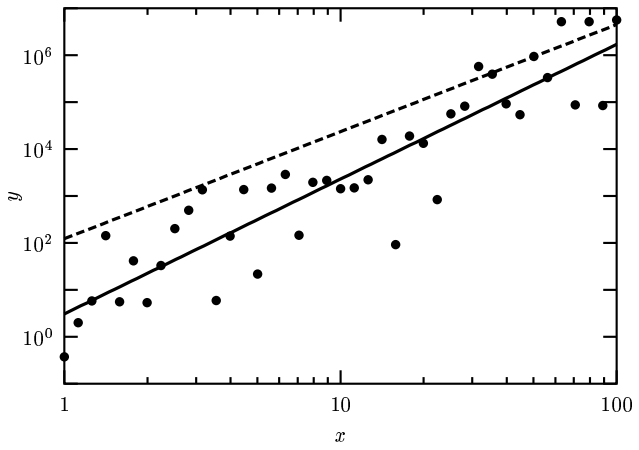

FIG. 1: The data correspond to a "noisy" power law, the dashed line is a direct nonlinear fit to a functional form $f_1(x) = a\,x^b$, while the solid line is a fit of the *logarithm* of the data to a functional form $f_2(x) = a + b\log x$.