Fitting a Straight Line to Data Part 2: Uncertainties in Both Parameters

In this lecture we will continue our discussion of how to fit a straight line to data. We'll use the same data set as last time, but now we will take into account that there are uncertainties in both variables. This will also give us an opportunity to discuss two additional statistical topics: the correct Bayesian approach when one of your parameters can vary from data point to data point, and propagation of uncertainties. To be self-contained, we'll give our data subset as before:

$\log_{10} M_{\rm bary}(M_\odot)$	$\sigma_{\log_{10} M_{\rm bary}}$	$\log_{10} v_{\rm rot}({\rm km\ s}^{-1})$	$\sigma_{\log_{10} v_{\rm rot}}$
8.68	0.06	1.76	0.02
9.45	0.09	2.03	0.02
10.62	0.11	2.19	0.02
11.03	0.13	2.45	0.02
10.65	0.28	2.27	0.02
9.94	0.15	2.12	0.01
11.13	0.12	2.38	0.01
10.09	0.14	2.10	0.02
11.06	0.10	2.33	0.01
10.38	0.27	2.19	0.02

As a reminder, the columns are: (1) $\log_{10}$ baryonic mass ($M_\odot$), (2) Gaussian uncertainty on $\log_{10}$ of baryonic mass, (3) $\log_{10}$ of the rotation speed (km s^{-1}), and (4) Gaussian uncertainty on $\log_{10}$ of the rotation speed.

Now that we allow the rotation speed, as well as the baryonic mass, to have uncertainties, we have to confront another issue. We assume that the slope and intercept of the relation between baryonic mass and rotation speed are the same for every galaxy. But the rotation speed is certainly *not* the same for every galaxy. Thus our analysis has to take into account that there are two types of parameters: one type (the slope and intercept of the relation) which are assumed to have the same value for every data point, and another type (the rotation speed) that can be different from one data point to the next. When we assumed that the rotation speed had zero uncertainties, we dealt with this by assuming explicitly that the prior on the rotation speed was independent of the slope and intercept of the relation, and then used the rotation speed as just an input. But now we have to be more careful. At the end of this lecture, we give more technical detail regarding a proper Bayesian treatment, but this is optional.

Propagation of uncertainties

You've probably seen something about how to propagate uncertainties, e.g., how you can find the uncertainty on a quantity that depends on more than one uncertain quantity. The problem is that the standard formula makes some assumptions, which are often not stated. To unearth those we will follow closely the as-usual excellent Wikipedia treatment of the subject. You can find that webpage at https://en.wikipedia.org/wiki/Propagation_of_uncertainty.

Say that you have some function f of n variables x_i, where i runs from 1 to n. Then we can write the Taylor approximation to f as

$$f = f_0 + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} x_i + \ldots \tag{1}$$

Note the ... part! It includes all the combinations of products of first derivatives, second derivatives, third derivatives, and so on. The usual discussion of propagation of uncertainty makes the *assumption* that these higher-order terms can all be neglected. We will now make that assumption, but note that this presumes that we are very close to some reference point, and that might not be true.

Now if we think like frequentists instead of Bayesians, we can imagine that f is some quantity that we measure over and over again. We'd like to know the variance of those many measurements. Recall that the variance is $\langle f^2 \rangle - \langle f \rangle^2$. Because f_0 is a constant, it does not contribute to the variance (remember our derivation of the variance in Lecture 2).

For the rest of the expression, there will in general be cross-terms between each pair x_i and x_j. *Only* when the uncertainties in the variables are uncorrelated with each other can these cross-terms be neglected (because individual measurements of $(x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle)$ will be negative as often as positive, and thus will average to zero). *Only* in that case can we then write the variance of f as

$$\sigma_f^2 = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i} \right)^2 \sigma_{x_i}^2 \ . \tag{2}$$

This is the standard expression you often see, but it's wrong unless both (1) higher-order terms in the Taylor expansion can be ignored, and (2) the uncertainties in all the variables on which f depends are uncorrelated with each other. You need to check carefully in your specific case whether this is true. In the case of our present data set, (1) is very likely *not* true; the standard uncertainties we are given suggest that the measured values of the log baryonic mass could often be many tens of percent different from the best value, and that is not a small deviation. Whether (2) is true depends on details that we don't have; for any analysis that we do from the beginning, we'd want to check this out!

Back to our problem...

For our analysis of the $\log M_{\rm bary}$ vs. $\log v_{\rm rot}$ data, we will now make all of the assumptions above. This is what Lelli, McGaugh, and Schombert (2016) do. The model whose parameters we are estimating is of the type

$$y = ax + b \; . \tag{3}$$

We can think of this as looking at the distribution of $y - (ax + b)$. If we set $f = y - (ax + b)$ then given all the assumptions above, we expect

$$\sigma_f^2 = \sigma_y^2 + a^2\sigma_x^2 \; , \tag{4}$$

and thus our equivalent of χ^2 when there are uncertainties in the measurements of both quantities might be

$$\chi^2 = \sum_i \frac{[y_i - (ax_i + b)]^2}{\sigma_y^2 + a^2\sigma_x^2} \; . \tag{5}$$

For our specific analysis task, $y = \log_{10} M_{\rm bary}(M_\odot)$ and $x = \log_{10} v_{\rm rot}({\rm km\ s^{-1}})$.

This is essentially the same as equation (6) of Cappellari et al. 2013 (MNRAS, 432, 1709), which is used by Lelli et al. 2016. In that equation, however, they add one more term to the denominator: ϵ_y^2, which they use to represent *intrinsic* scatter in the relation (i.e., taking into account that the actual physical relation is not a precise, zero-width line). That term is also used by Lelli et al., but for simplicity we won't use that in our analysis. Remember, however, that adding intrinsic scatter requires yet another choice: the choice made by Cappellari et al., which is also made by Lelli et al., is that the intrinsic scatter is the same for every point, but there are many other possibilities. Cappellari et al. note that their equation (6) is "only rigorously valid when the errors in x and y are Gaussian and uncorrelated", and thus that if there are correlations the expression must be modified.

With all this in mind, we can redo the analysis we did in the last lecture, but now using the quoted uncertainties in both variables. As expected, we see in Figures 1 through 5 that the credible regions are larger than they were when we ignored the uncertainties in $\log v_{\rm rot}$. The minimum chi squared for the 10-point fit is now 17.5 rather than 25.9, for dof = 8 as before. This has a formal probability of 0.025 rather than 0.001, so we've improved. However, we'd want to keep an eye on this for the full data set. We can also determine the marginalized probability distributions as before, but I'll leave that to you.

Now it's your turn: do the same analysis as we have done here, using the full data set. What do you find? If you were to add a constant uncertainty ϵ_y^2 to each measurement, what value of ϵ_y^2 is needed so that the minimum $\chi^2 = {\rm dof}$?

Debriefing: what have we learned?

In this lecture and the previous lecture we have gone over many aspects of a "simple" linear fit to data. What I hope you take away from our discussions is that there are many assumptions that go into a standard fit of this type. Sometimes you can't get around them, or sometimes you want to do just a quick analysis, but please be aware of the compromises you have made and their consequences, and report them honestly!

In some sense it comes down to how much you care about your results. If you are doing a simple analysis, without major implications, then making some standard assumptions might not be bad (although you need to check!). If you think your data imply something really special, then you need to be very careful indeed. But always, always, you need to tell your reader precisely what you did.

In more technical detail: Bayesian treatment of the ideal case

Let's begin with some general statements and then treat our problem more specifically. Suppose that there is one set of parameters, which we represent by the vector α, which we expect to be the same for every observation (such as the slope and intercept), and another set of parameters, which we represent by the vector β_k for the kth observation, which can be different in every observation. We will marginalize over β, but we need to be careful. For example, it would be wrong to multiply the marginalized posteriors:

$$P(\alpha) \neq \prod_k \left(\int \mathcal{L}_k(\alpha, \beta_k) p(\alpha) p(\beta_k|\alpha) d\beta_k \right) \ . \tag{6}$$

Here $p(\beta_k|\alpha)$ is the prior probability of β_k given particular values of the parameters α. The reason this is wrong is that we could take $p(\alpha)$ out of all the integrals, which would mean that the α dependence would end up being

$$\prod_k p(\alpha) \ . \tag{7}$$

Thus if $p(\alpha)$ is even somewhat peaked, and there are many observations k, then the result would be to raise $p(\alpha)$ to a large power, which would make it extremely sharply peaked independent of the data. That's certainly not right!

So instead we need to do something else. We can derive the right expression by remembering that the posterior is the product of the prior with the likelihood:

$$P(\alpha, \beta_1, \beta_2, \ldots) = p(\alpha, \beta_1, \beta_2, \ldots) \mathcal{L}(\alpha, \beta_1, \beta_2, \ldots) \ , \tag{8}$$

and the marginalized posterior for α is

$$P(\alpha) = \int P(\alpha, \beta_1, \beta_2, \ldots) d\beta_1 d\beta_2 \cdots \ . \tag{9}$$

We are assuming that the observations are independent of each other, so we can write this as

$$\begin{aligned} P(\alpha) &= \int p(\alpha, \beta_1, \beta_2, \ldots) \mathcal{L}(\alpha, \beta_1, \beta_2, \ldots) d\beta_1 d\beta_2 \cdots \\ &= \int p(\alpha) p(\beta_1|\alpha) p(\beta_2|\alpha) \ldots \mathcal{L}_1(\alpha, \beta_1) \mathcal{L}_2(\alpha, \beta_2) \ldots d\beta_1 d\beta_2 \cdots \\ &= p(\alpha) \prod_k \left[\int p(\beta_k|\alpha) \mathcal{L}_k(\alpha, \beta_k) d\beta_k \right] \ . \end{aligned} \tag{10}$$

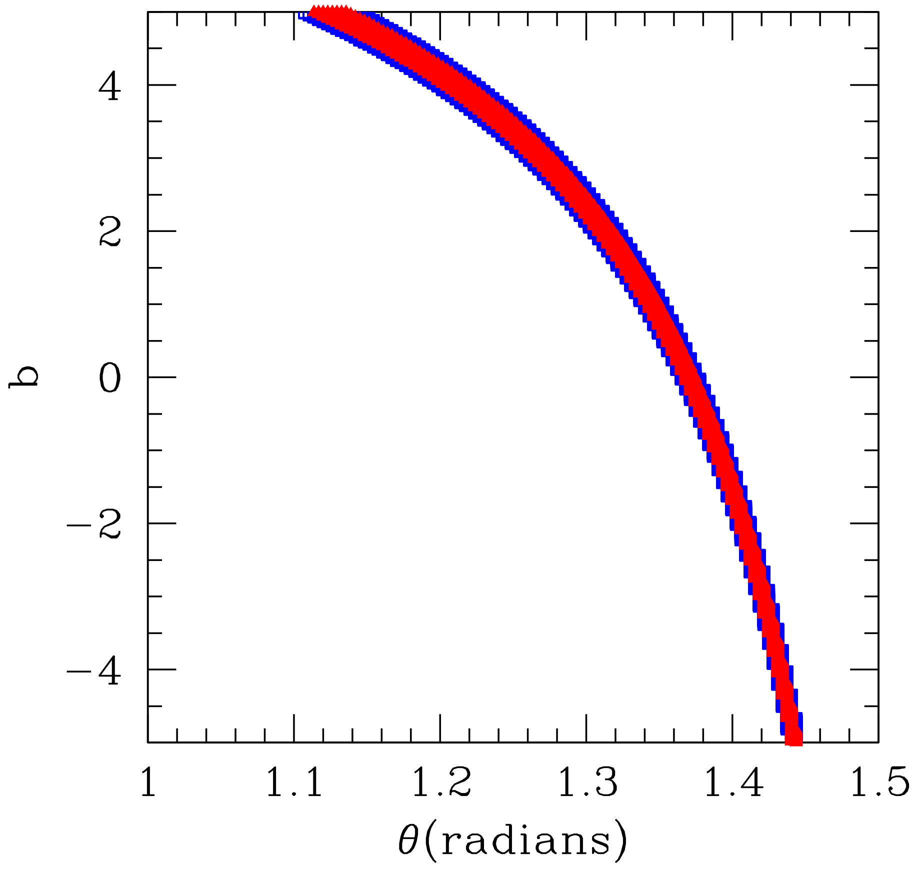

Fig. 1.— Same as Figure 2 in Lecture 10, but using the quoted uncertainties in both variables. Notice that as expected for this figure, and the rest of the figures, the 68.3% credible region and the 95.4% credible region are both larger than they were when we did not incorporate the uncertainties in $\log v_{\rm rot}$ in our analyses.

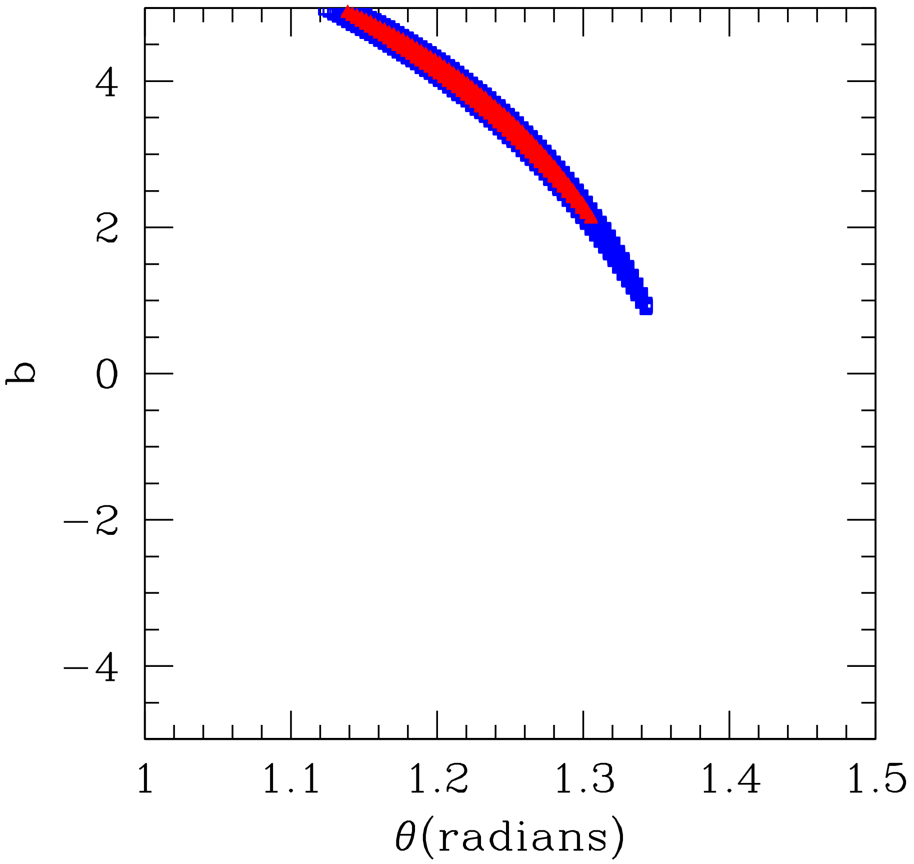

Fig. 2.— Same as Figure 3 in Lecture 10, but using the quoted uncertainties in both variables.

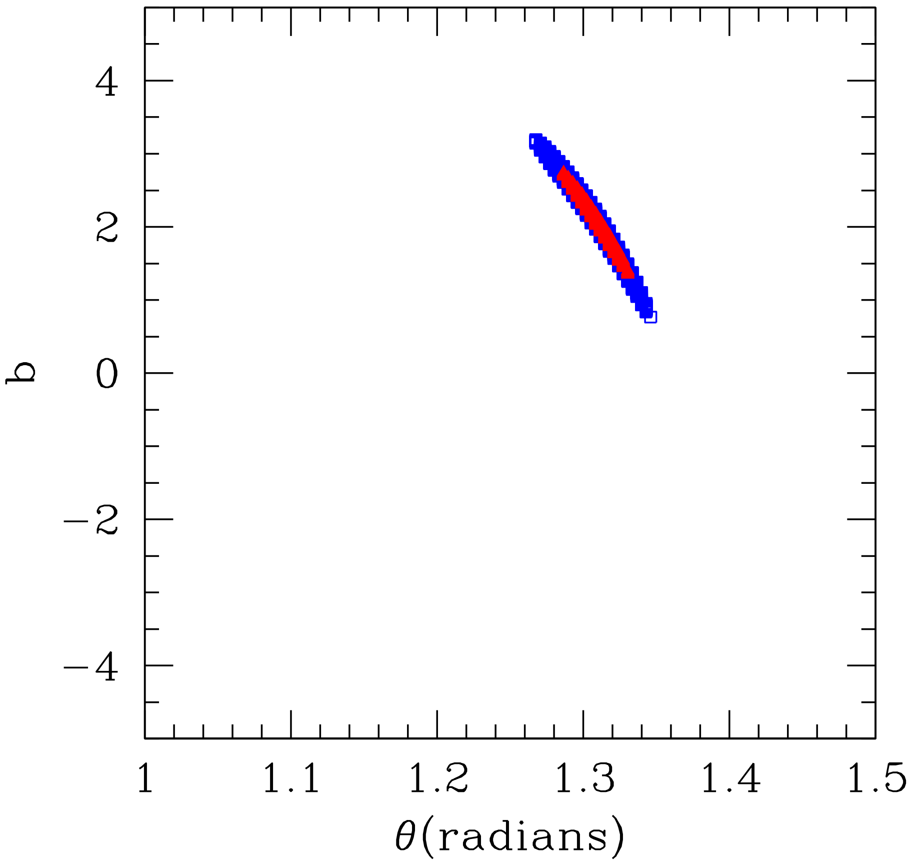

Fig. 3.— Same as Figure 4 in Lecture 10, but using the quoted uncertainties in both variables.

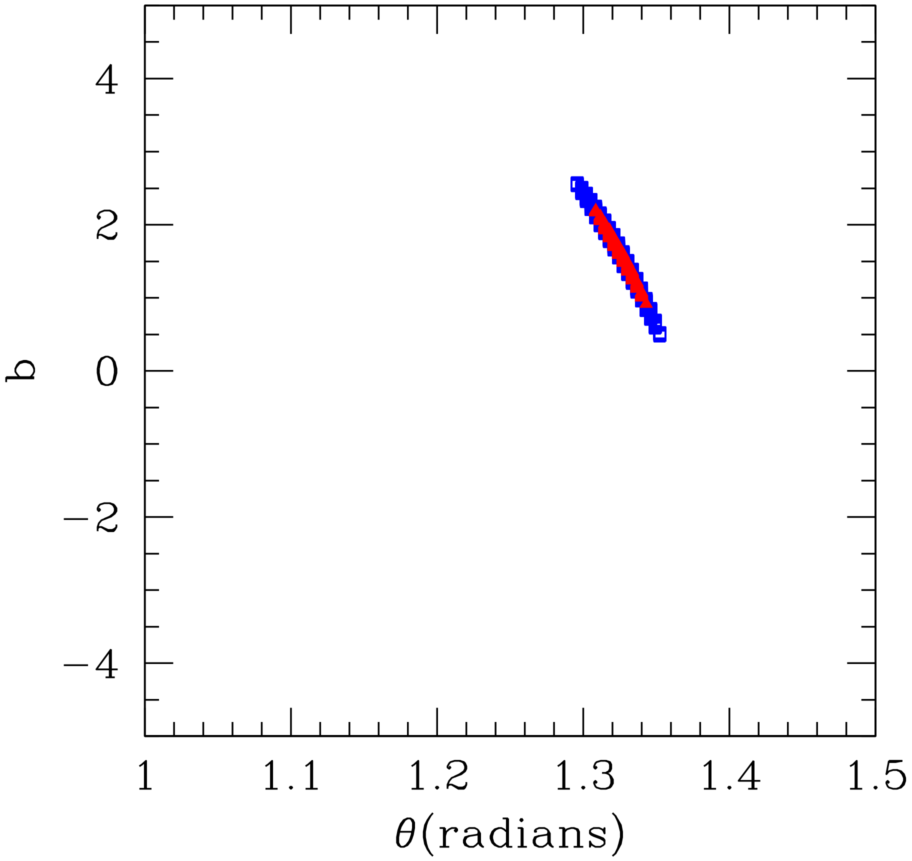

Fig. 4.— Same as Figure 5 in Lecture 10, but using the quoted uncertainties in both variables.

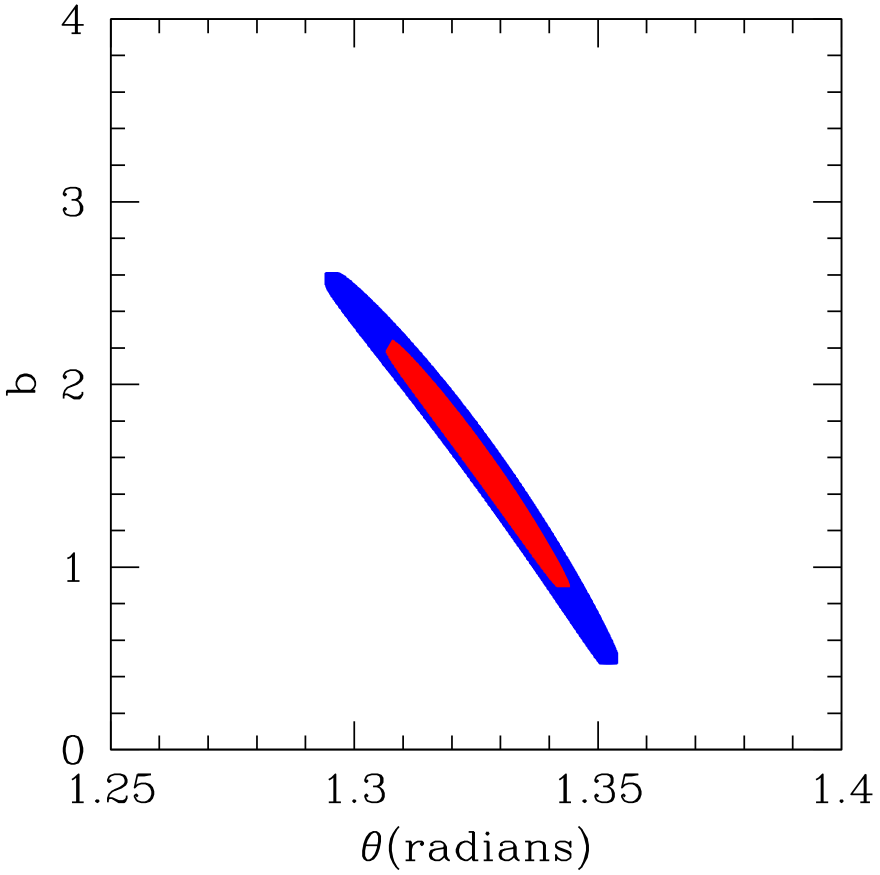

Fig. 5.— Same as Figure 6 in Lecture 10, but using the quoted uncertainties in both variables.

A couple of notes about this:

Note 1: because the likelihood is the likelihood of observing the data given the model and a specific set of parameters, this is where the details of the observations come in. For example, a shallow survey cannot detect very faint sources, so the likelihood of seeing a faint source is much less than the likelihood of seeing a bright source, all else being equal. This is how selection effects are taken into account. This is another reason why it is often not a good assumption that the likelihood is a symmetric Gaussian.

Note 2: If you look at the formalism above you can see that the nuisance parameter sets β_1, β_2, etc. do not have to be the same as each other. Thus as long as α is fixed between the different data sets, you can have *completely different observations* that constrain your interesting parameters.

As you might expect, my favorite example of this type involves neutron stars. A key question is: what is the equation of state of the matter in the cores of neutron stars? The "equation of state" relates the pressure to some other thermodynamic variable, such as the energy density. The temperature is not expected to play an important role, and the matter is expected to be in its ground state (i.e., the lowest energy possible given the energy density), which simplifies things a lot. All neutron stars should have the same equation of state; although heavier neutron stars get to higher energy densities at their center, at a given energy density the matter in the core of any neutron star should have the same pressure. We can't figure out the equation of state from terrestrial experiments, because the physical situation (very high density, far more neutrons than protons) can't be replicated in laboratories. Thus we need to do astronomical observations instead. Those observations can be varied: for example, the maximum mass possible for a slowly rotating neutron star, or the radius for a given mass, or the moment of inertia for a given mass, or the tidal deformability for a given mass, can all be derived from the equation of state. Therefore measurements of any of these relations can be used to constrain the equation of state. Although the nature of those measurements is highly varied (e.g., analysis of X-rays from rotating neutron stars, or radio signals from double neutron star binaries, or gravitational waves from coalescing double neutron star or neutron star – black hole binaries), the Bayesian framework above can be used to put them all on the same footing to constrain a parameterized model of the equation of state.