**Fitting a Straight Line to Data (repeat of previous lecture)**

Thanks for your patience. Finally we'll take a shot at real data! The data set in question is baryonic Tully-Fisher data from http://astroweb.cwru.edu/SPARC/BTFR_Lelli2016a.mrt, which was suggested by Liz Tarantino and which is from Lelli, McGaugh, and Schombert 2016, ApJ, 816, L14 (you might want to read this paper – it's just six pages long – to see what the authors did). This will give us an opportunity to go through our recommended procedure of deciding on a statistical approach in which each decision is conscious. The set is given in full on the website, but to be self-contained we reproduce part of it here:

| $\log_{10} M_{\rm bary}(M_\odot)$ | $\sigma_{\log_{10} M_{\rm bary}}$ | $\log_{10} v_{\rm rot}$(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 |

The columns are: (1) $\log_{10}$ baryonic mass ($M_\odot$), (2) standard error on $\log_{10}$ of baryonic mass, (3) $\log_{10}$ of the rotation speed (km s$^{-1}$), and (4) standard error on $\log_{10}$ of the rotation speed. Here the "standard error"s, represented by $\sigma$s, mean a Gaussian uncertainty (not an error!) in the measurement. Thus it seems like we're ready to go: just toss this into some kind of $\chi^2$ minimization and we're set, right?

Nope. Remember that one of the guiding principles of this course is that we *think* about what we are doing! Thus our first step will be to look at the data carefully to consider what stands out to us. Then we'll determine what we *would* do if we had unlimited time and resources. Finally we'll decide what we *can* do given our limited time, and in this case, the limited data that we have.

So what stands out about these data? Here are some thoughts:

1. The data in the file are not the raw data. Instead, the baryonic mass and the rotation speed are quantities derived from the raw data. Therefore, as certainly will happen sometimes, we will not be able to perform the ideal analysis of raw data.
2. The derived data have uncertainties in *both* quantities! Moreover, the quantity that we might be tempted to consider as the more fundamental, independent, quantity is

the baryonic mass, which has a fractional uncertainty that is greater than that of the rotation speed (recall that these are logarithmic quantities). Often it is assumed that the independent variable has no uncertainties at all, and thus that we can treat its measurements as perfect. Clearly we can't really do that here.

3. Those can't be the actual uncertainties on $\log_{10} v_{\rm rot}$. They're quantized! This means that the authors of the table decided to use only two significant figures, or maybe something else was going on (for example, it could be that the actual measurements have much smaller uncertainties, but that they add some rough estimate of systematic errors). We'd have to check the original sources of the data to know for sure.
4. Because the uncertainties are in the logs of quantities, we have to wonder whether they are really symmetric. For example, $10^{0.28} = 1.9$, so in the fifth row an assumption of symmetry would mean that, for example, it is equally likely that the baryonic mass is a factor of 1.9 below the best estimate, as it is that the baryonic mass is a factor of 1.9 above the best estimate. Is that true?
5. More generally, if we use $\chi^2$ on these data, we have to realize that we are making the implicit assumption that the probabilities are Gaussian *to arbitrary numbers of standard deviations*. This is almost never the case, and it could affect the analysis in a large data set like this one because there could be points that are "$3\sigma$" off of the line but are either more or less probable than they would be for a Gaussian.

Is there anything else you notice about the data?

Now let's take another obvious simple step: we'll look at the data, which we have plotted in Figure 1.

We are thinking about fitting a straight line to these data. Because the data we have involve the logs of two things, we are therefore considering the possibility that the two quantities are related by a power law. From Figure 1, does it appear that a straight line could more or less go through the data? Sure. If the lower left point weren't in the set then we might be tempted to try a fit with a quadratic, but even then it wouldn't be silly to try a straight line.

Now that our eyeball test tells us that a straight-line fit is plausible, we need to think about how we should do our fit. As noted before, we can't apply the ideal procedure (of folding a model through the response of our detector and comparing the predictions with the actual, raw, data). Therefore, we need to make some compromises.

As our first compromise, we'll pretend that only one of the derived variables has uncertainties, whereas the other is known perfectly. In the next class we'll use this same data set without that assumption, so that we can see the differences in our inferences. As we say above, we might think of the baryonic mass as the independent quantity, but because the

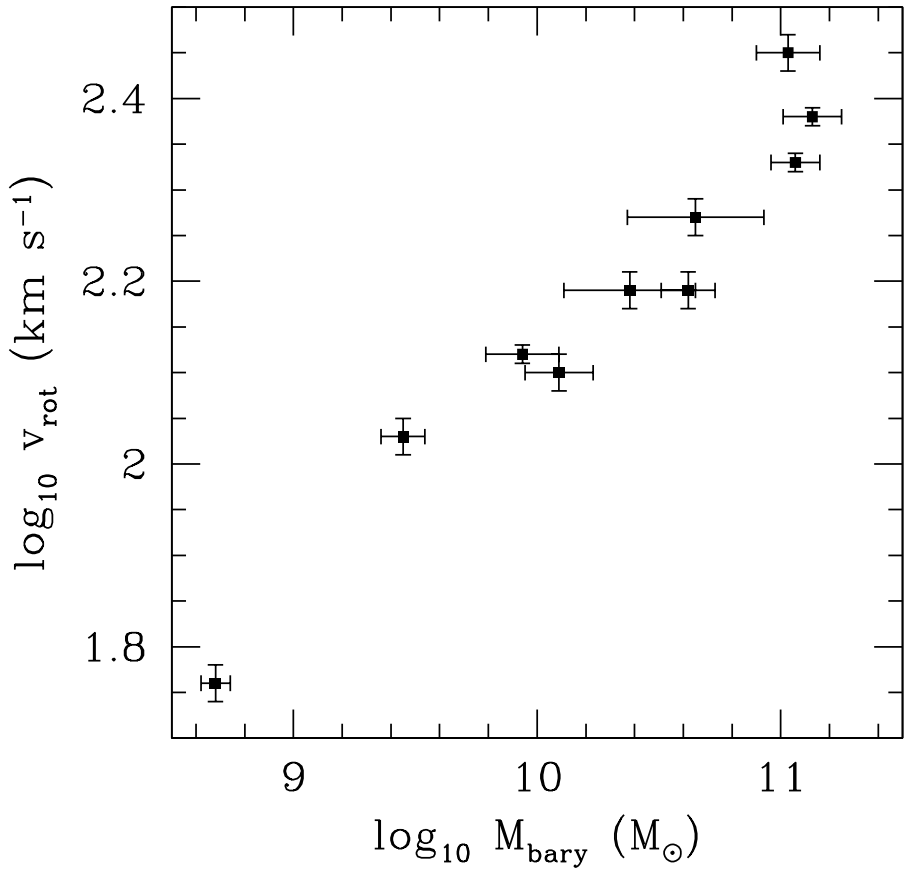


Fig. 1.— Ten selected points from a baryonic Tully-Fisher data set, along with their quoted standard uncertainties. Original data set from Lelli, McGaugh, and Schombert 2016, ApJ, 816, L14, tabulated at http://astroweb.cwru.edu/SPARC/BTFR_Lelli2016a.mrt

rotational speed has smaller fractional uncertainties it is more reasonable to treat *that* as the independent variable, with no uncertainties.

Our next step as Bayesians is to make a conscious decision about whether we wish to do parameter estimation, or to perform a model comparison. Let's say that we want to do parameter estimation, which means that we cannot say anything about whether the straight-line fit (in log-log space) is preferred over other possible fits.

In our particular parameter estimation the model parameters we care about are the slope and intercept in $\log_{10} M_{\rm bary} - \log_{10} v_{\rm rot}$ space. But another parameter in the model is the rotation speed; remember, the data space we're considering (which still isn't the true space of raw data, but it's what we have) has *two* observables, the rotation speed and the baryonic mass. Thus for our parameter estimation we need to (1) specify the prior probability distribution for all three parameters (and note that in general we need to specify the *joint* prior, because it will not always be the case that the three probability distributions are independent), (2) perform our analysis using the data to get a normalized posterior probability distribution in all three parameters, and then (3) marginalize over the rotation speed to get our final joint posterior probability distribution for the slope and intercept.

What should we choose for our priors? We'll begin with the rotation speed, which is our nuisance parameter for this problem (because we don't care about it directly). Should we assume that all rotation speeds are, a priori, equally probable? Should we instead assume that all *logs* of rotation speeds are, a priori, equally probable? Maybe we should take our prior from observations. All would lead to different weightings. Our choices could, in fact, lead to different results for our estimation of the slope and intercept if the prior for the rotation speed depends on the slope and/or intercept.

This suggests a simplification that we can make consciously. We can assume that the prior on the rotation speed, whatever it is, does *not* depend on either the slope or the intercept. Thus if our model is $\log_{10} M_{\rm bary} = a \log_{10} v_{\rm rot} + b$, the assumption of independence means that the prior $p(a, b, v_{\rm rot})$ becomes $p(a, b)p(v_{\rm rot})$. Then, because we are assuming in our current analysis that $v_{\rm rot}$ is measured without any uncertainty, the probability of $v_{\rm rot}$ given the model is something that does not depend on $a$ or $b$, and so we don't have to worry about it.

This is what we need to assume to get to the point where many people would automatically start: for a given $a$ and $b$ we can simply compute the likelihood, for each point in our data set, that we would measure the observed $M_{\rm bary}$ given the $M_{\rm bary}$ that is expected in our linear model at the exactly measured $v_{\rm rot}$. The product of those likelihoods, over all of the points in our data set, is the probability of the data given the model. Because we have been handed the data with uncertainties attached to the data (rather than having a probability distribution for data given a particular model), and because we are for the same reason

forced to assume that the uncertainties are Gaussian to an unlimited number of standard deviations, this means that our likelihood is the exponential of minus one half of the chi squared.

Are we ready to do the calculation? Not yet. We still need to think about our priors for $a$ and $b$. First, we can choose $b$. We may not have good reason to restrict $b$ particularly, so as long as the prior is basically flat over the possible range then we'll be fine. Looking at our plot, maybe the prior could be that $b$ is constant from $b = -5$ to $b = +5$ (remember, $b$ has units of $\log_{10} M_{\rm bary}$).

But $a$ is another story. Our first inclination might be to say that $a$ could be anything, with equal probability. That sounds great except that $a$ could go to $+\infty$ or $-\infty$ if horizontal lines are allowed in Figure 1. Thus if we allow $a$ to be "anything, with equal probability", almost all of the weight of the prior will be on nearly horizontal lines, which would then overwhelm any evidence we have from the data!

With that in mind, it would be more reasonable to, for example, have as a prior that the line in question could make any angle with equal probability. Then our new parameterization would be

$$\log_{10} M_{\rm bary} = \tan\theta \log_{10} v_{\rm rot} + b \tag{1}$$

where the prior is $p(\theta, b) = p(\theta)p(b)$, where $p(\theta) = \frac{1}{\pi}$ from $\theta = 0$ to $\theta = \pi$ and 0 otherwise (because $\pi$ to $2\pi$ would duplicate the lines we have from $\theta = 0$ to $\theta = \pi$), and $p(b) = \frac{1}{10}$ from $b = -5$ to $b = +5$ and 0 otherwise. Note that in making this choice we have explicitly assumed that the priors for the slope and intercept are independent of each other, but such independence will not be the case in general.

With that in mind, we can compute the posterior probability distribution as a function of $\theta$ and $b$:

$$P(\theta, b) \propto \frac{1}{\pi}\frac{1}{10}\prod_{i=1}^{N} \exp[-(m_i - d_i)^2/2\sigma_i^2] \ , \tag{2}$$

where $\frac{1}{\pi}\frac{1}{10}$ is the prior, $d_i$ is the $i$th measurement of $\log_{10} M_{\rm bary}$, $\sigma_i$ is the reported Gaussian uncertainty in the $i$th measurement of $\log_{10} M_{\rm bary}$, and $m_i = \tan\theta \log_{10} v_{\rm rot,i} + b$, where $v_{\rm rot,i}$ is the $i$th measurement of the rotational speed. Note that the natural log (i.e., ln) of the product is $-\chi^2/2$. We normalize the posterior probability density by multiplying it by a constant factor such that

$$\int_0^{\pi}\int_{-5}^{5} P(\theta, b) db d\theta = 1 \ . \tag{3}$$

We found earlier that even the probability distribution for a single parameter does not give us a unique definition for (say) the 68.3% credible region, and as you'd expect the situation is not improved for multiple parameters. For the moment, let's do a standard $\chi^2$

thing and assume that for two parameters the "$1\sigma$" (68.3%) credible region includes all points within $\Delta\chi^2 = 2.3$ of the minimum, and that the "$2\sigma$" (95.45%) credible region includes all points within $\Delta\chi^2 = 6.18$ of the minimum. Because computers are fast, we'll do this using brute force: we will compute the total chi squared at each of $10^6$ points in a grid: 1000 in $\theta$ and 1000 in $b$. The results are shown in Figures 2 through 5 (which give the $\chi^2$ surfaces for, respectively, 1, 2, 5, and all 10 points), and Figure 6 shows a zoomed-in version of the 10-point $\chi^2$ surfaces.

We can double-check that the $\Delta\chi^2 = 2.3$ region has 68.3% of the total probability by integrating the normalized posterior probability distribution in that region. Note that because we have assumed a constant-probability prior, the posterior probability density is proportional to the likelihood. When we do this, we find that 68.2% of the probability is within $\Delta\chi^2 = 2.3$ of the minimum. Similarly, we find that 95.4% of the probability is within $\Delta\chi^2 = 6.18$ of the minimum. It is unsurprising that we get the expected probabilities; this follows from our *assumptions* of Gaussianity in our probabilities. One way to convince yourself of this is to reassign $v_{\rm rot}$ randomly with $M_{\rm bary}$ in our data set and redo the analysis; you should find that the fraction of the probability within $\Delta\chi^2 = 2.3$ or $\Delta\chi^2 = 6.18$ is very similar to what it is in the real case.

What if we are interested in only one of our two parameters? Then, as we discussed in an earlier class, we marginalize the posterior probability distribution over the other parameter. Those distributions are displayed in Figure 7 and Figure 8. The peak probability density in two dimensions has $\theta = 1.32088$ and $b = 1.758$. The peak of the marginalized distribution in $\theta$ is $\theta = 1.32088$, but the peak of the marginalized distribution in $b$ is $b = 1.782$, i.e., larger by a small but significant amount than the value of $b$ at the peak in the two-dimensional distribution. It is, in fact, common that the peak in the marginalized distribution is not the same as the peak in the two (or larger) dimensional distribution. The reason in our case is that the two-dimensional region bends. In particular, at larger $b$ there is a larger range of $\theta$ that gives a good fit, which means that when we integrate over $\theta$ the integrated probability is larger at large $b$. This is clearer in the two-point analysis Figure 3, but also applies to the less obviously curved posterior probability distribution for ten points. Sometimes people will *assume* that the peak in the full distribution coincides with the peak in the marginalized distribution, but in general they won't be the same.

In any case, this is all looking pretty good! Just one last check: we can compute the total $\chi^2$, and compare it with what we would expect given the number of degrees of freedom. This procedure is non-Bayesian; we should really just compare precisely specified models, but as I said before I think it's useful to get a sense for whether our fit is okay. The number of degrees of freedom (abbreviated as dof) is the number of data points minus the number of model parameters, so dof=10−2=8 in our case. Thus we expect that for a good fit, the minimum $\chi^2$ would be around 8. Instead, it's 25.9. Looking this up in a $\chi^2$ table we find

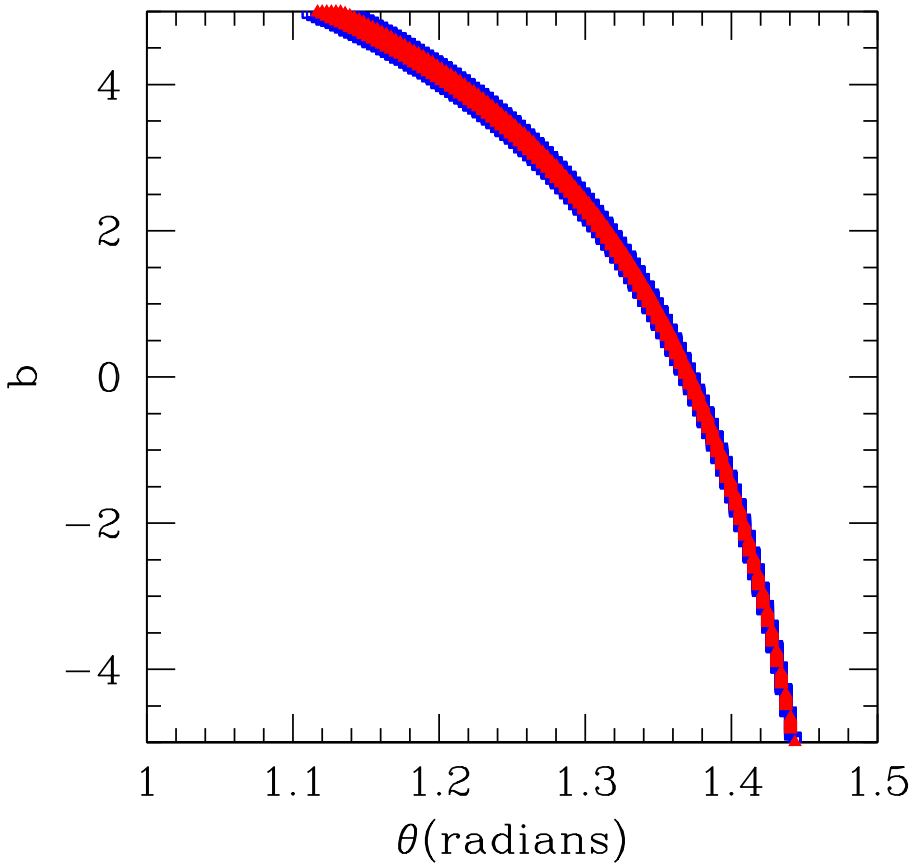


Fig. 2.— Representation of the posterior probability density in $\theta - b$ space using only the first point in our data set. The model we employ is that $\log_{10} M_{\rm bary}(M_\odot) = \tan\theta \log_{10} v_{\rm rot}(\mathrm{km\ s^{-1}}) + b$, and we do our analysis assuming that $\log_{10} v_{\rm rot}$ is measured with no uncertainty. The red points indicate where in our $\theta - b$ grid the $\chi^2$ is within 2.3 of the minimum (which corresponds to $1\sigma$, or the 68.3% credible region, for two parameters if $\chi^2$ analysis is valid). Similarly the blue points are where $\chi^2$ is within 6.18 of the minimum, which corresponds to $2\sigma$, or the 95.4% credible region, for two parameters. With only one point there are obviously an infinite number of lines that go straight through the point, which means that there is a strong correlation between the two parameters.

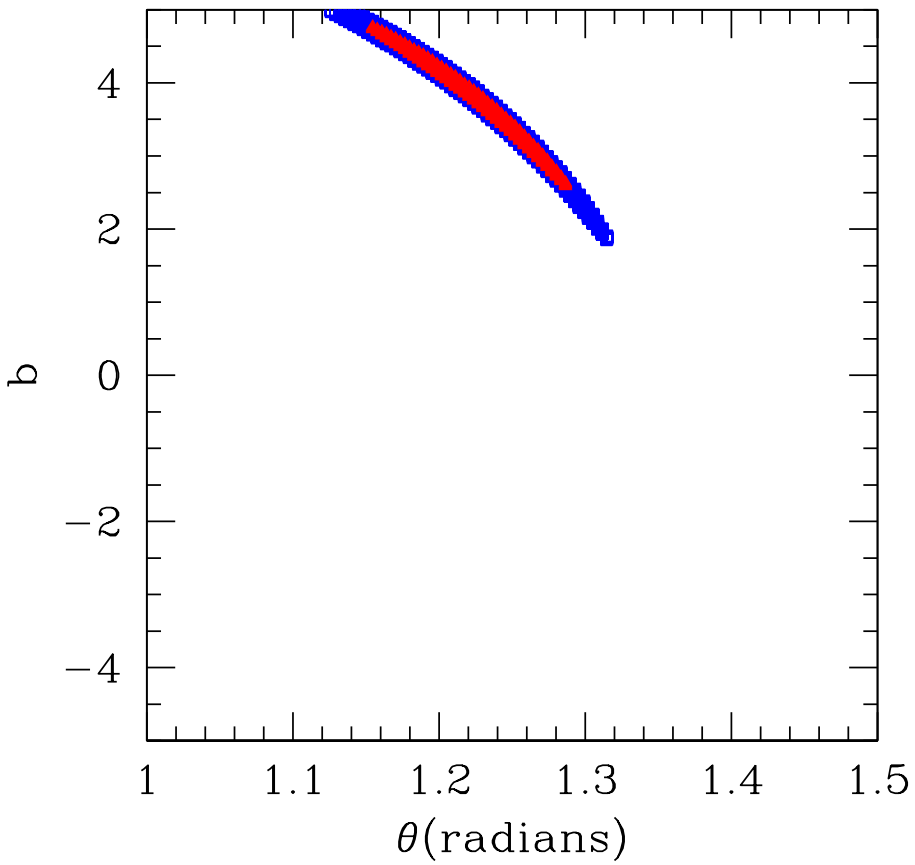


Fig. 3.— Same as Figure 2, but using the first two points in our data set. Now there is a unique line that goes perfectly through both points, so the range of decent fits is finite.

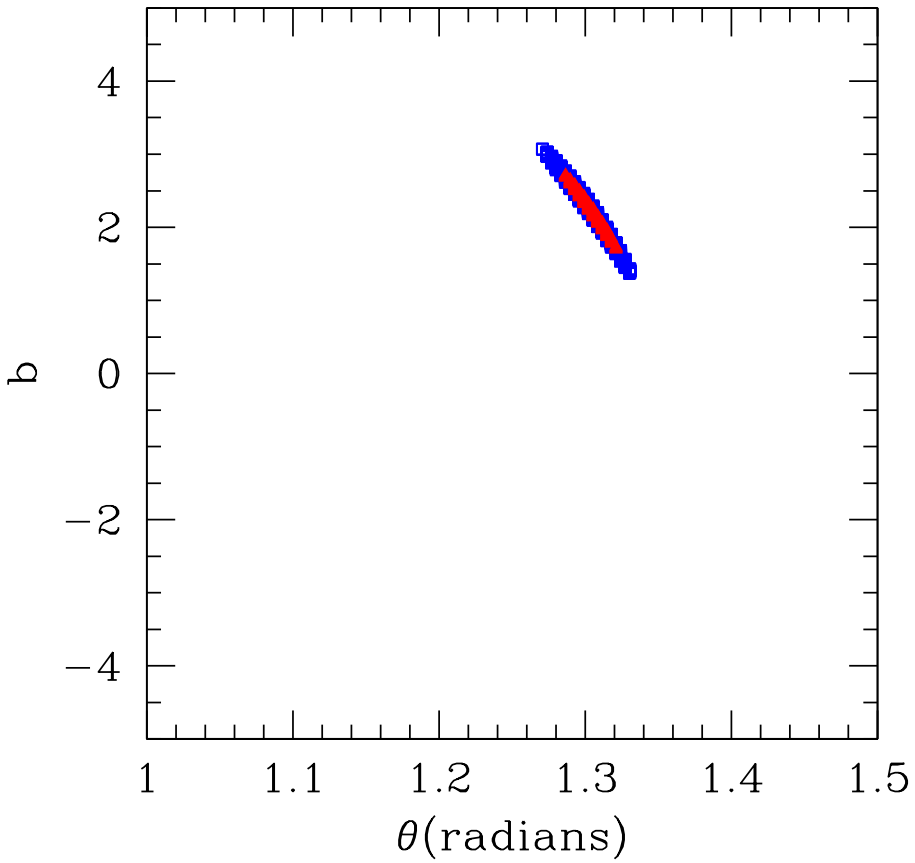


Fig. 4.— Same as Figure 2, but using the first five points in our data set. Now the constraints are obviously much tighter. Note that there are parts of the 68.3% credible region that are well outside the corresponding region in the two-point plot. This is reasonable and expected; we still don't have many points, so fluctuations play a big role.

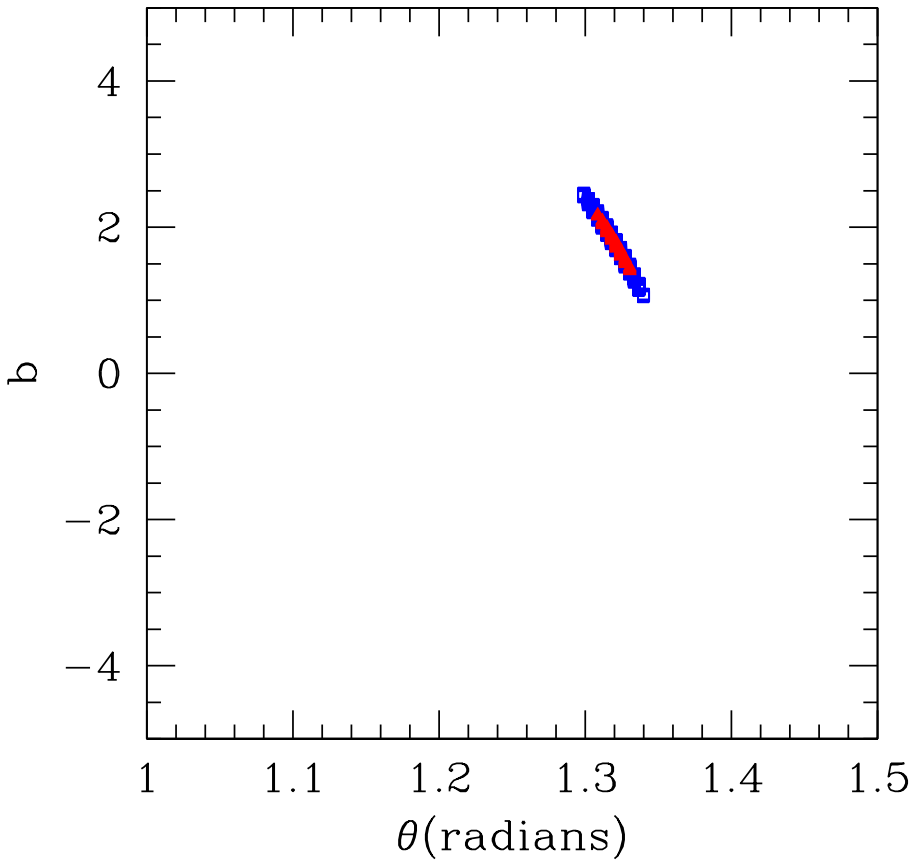


Fig. 5.— Same as Figure 2, but using all ten points in our data set (which is a subset of the whole data set). The constraints have improved further, but at this stage our grid is too coarse to represent the region well.

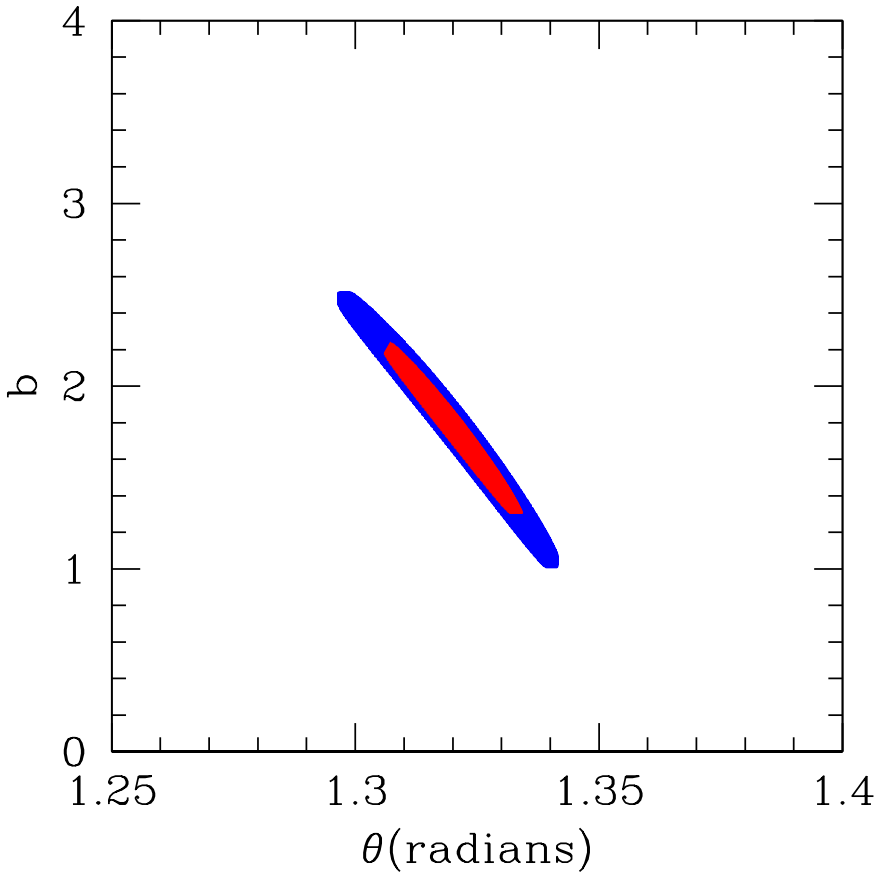


Fig. 6.— Same as Figure 5, but using a finer grid to focus on the best-fit region.

that we'd expect $\chi^2 \geq 25.9$ for dof $= 8$ only 0.1% of the time.

Oops! Maybe we just got one-in-a-thousand unlucky, but more likely something is wrong. In our particular case we have a good candidate for what that is: we ignored the uncertainties in $v_{\rm rot}$. But in other cases, we'd have to pause here and recognize that our model does not describe the data as well as would be expected if we had chosen the correct model. What would we do then? A common hack is to arbitrarily multiply all the uncertainties by a large enough factor that $\chi^2 \approx$ dof. However, this isn't justified. Among other things, in our actual data set, how do we know that all uncertainties are underestimated by the same factor?

Another variant of this approach is to assume that there is an intrinsic scatter in the relationship. Phrased in a Bayesian way, we would say that if there is some intrinsic linear relationship between $\log_{10} v_{\rm rot}$ and $\log_{10} M_{\rm bary}$, then the probability of measuring an actual $v_{\rm rot}$ and $M_{\rm bary}$ depends on true deviations from the linear relationship as well as measurement uncertainties. That is, even if we had absolutely perfect measurements of $v_{\rm rot}$ and $M_{\rm bary}$, they would not fall on a perfect log-log relationship. This is the approach adopted by Lelli et al. (2016). The challenge here is to determine how to combine the intrinsic scatter with the measurement uncertainties. A standard approach is to assume that the intrinsic scatter as well as the measurement uncertainties are (1) both Gaussian, and (2) add "in quadrature", which means that if the measurement variance is $\sigma^2_{\rm meas}$ and the intrinsic variance is $\sigma^2_{\rm int}$, then the total variance is $\sigma^2_{\rm tot} = \sigma^2_{\rm meas} + \sigma^2_{\rm int}$ (this fundamentally assumes that the measurement uncertainty and the intrinsic scatter are *uncorrelated*). Neither assumption (1) nor assumption (2) need be correct. In addition, even if the assumptions are correct, it would make sense that the intrinsic scatter depends on, say, the rotation speed, whereas a typical analysis would make the simpler assumption that the intrinsic scatter, in log space, is a constant. A better approach, in general, is to look at a formally poor fit as an opportunity to evaluate your fit: is the model actually a poor description of the data?

This comes back to the philosophy of this course. There is a right way to do things, but in practice it may not be possible to realize that ideal. All I ask is that you think carefully about what you're doing, understand the consequences of your compromises as best as possible, and report those compromises and their likely consequences honestly in the papers you write.

So there it is! Even an apparently simple task such as fitting a straight line to data actually involves many choices. Now it's your turn: look at the whole data set, which is given on the website, and do the analysis we've discussed here. What do you learn?

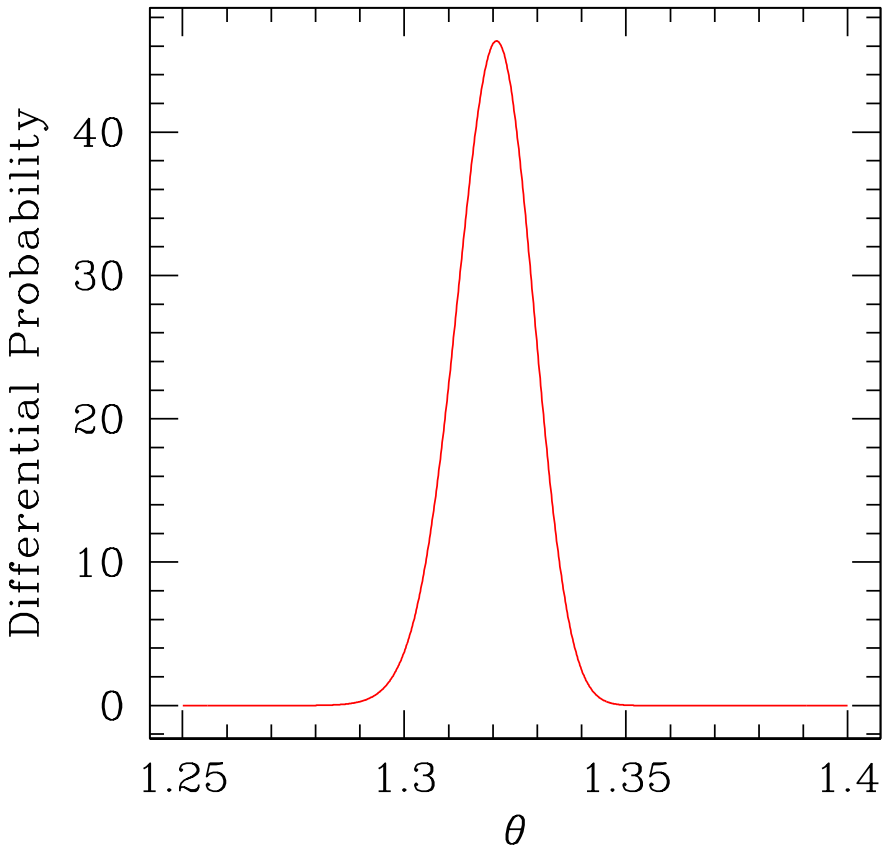


Fig. 7.— Marginalized posterior probability density for $\theta$ alone. Recall that the probability *density* can be anything; it is the total probability that must integrate to unity.

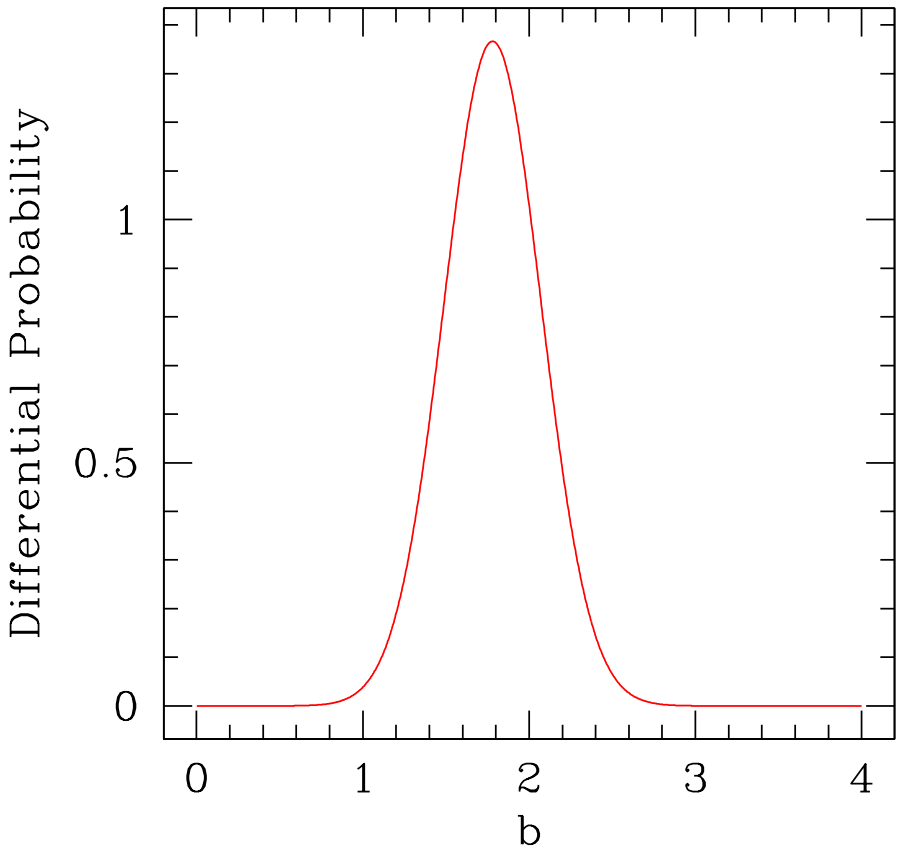


Fig. 8.— Marginalized posterior probability density for $b$ alone. The peak in the marginalized posterior for $b$ is *not* the same as the peak in the two-dimensional $\theta - b$ distribution. It's close, but not identical. This is, in fact, the most general case; strict agreement between the value of a parameter at the peak of a multidimensional distribution, and the value of that parameter at the peak of its marginalized distribution, only happens in idealized cases.