Statistics for Astronomers
Solutions to Final Examination (Tuesday, 2021.02.09)

Prof. Sundar Srinivasan

Note: the script `finalexam.py` can be used to generate the output described in this document.

Note: I'm working with base-10 logarithms in this document. If you work with natural logarithms, then your results won't have the $\ln 10$ factor in your analysis. The resulting intercepts will be larger by the same factor (≈ 2.303).

1. The `q1` module in `finalexam.py` outputs

```
Performing Student's t-test assuming unequal variances
======================================================
------------------------------------------------------
Samples:  E-S0 vs.  S0/a-Sd.
H0:  samples drawn from distributions with identical mean stellar masses.
Ha:  samples drawn from distributions with differing mean stellar masses.
 The p-value (1.76e-01) is >= the significance (0.05).  H0 cannot be rejected!
------------------------------------------------------
------------------------------------------------------
Samples:  E-S0/a vs.  Sa-Sd.
H0:  samples drawn from distributions with identical mean stellar masses.
Ha:  samples drawn from distributions with differing mean stellar masses.
 The p-value (1.37e-01) is >= the significance (0.05).  H0 cannot be rejected!
------------------------------------------------------
------------------------------------------------------
Samples:  E-Sab vs.  Sb-Sd.
H0:  samples drawn from distributions with identical mean stellar masses.
Ha:  samples drawn from distributions with differing mean stellar masses.
 The p-value (2.28e-02) is < the significance (0.05).  H0 rejected!
------------------------------------------------------
======================================================
```

The subsample consisting of late-type galaxies beyond Sb appear to be drawn from a population with differing properties from the subsample of galaxies with morphologies earlier than Sb. The full sample considered in Pruzhinskaya et al. (2020) gives more robust results.

2. (a) The `q2` module in `finalexam.py` outputs

```
Number of data points:  10.
===> DOF = 8.
Expected chisq for this problem:  8.
Observed chisq:  7.11.
68% central CI: [4.19, 11.81].
```

```
Fit within tolerance!
```

(b) The central assumption in χ^2 fitting is that the data are perfectly described by the model except for Gaussian errors. That is,

$$y_{\text{data},i} = y_{\text{model},i} + \epsilon_i \qquad \text{with } \epsilon_i \sim \mathscr{N}(0, \sigma_i^2),$$

where σ_i are the measurement uncertainties in y. In order for the χ^2 distribution to be applicable, the weighted residuals (deviation of model from data, divided by the uncertainty in the data) have to be drawn from a Standard Normal distribution. We can perform a KS test to check for this. The q2 module in `finalexam.py` outputs

```
 Performing KS test to check whether residuals are Gaussian...
H0:  residuals are drawn from standard normal distribution.
p-value = 1.1e-03 < alpha = 0.05.  H0 rejected!
```

Therefore, just because a minimisation procedure results in a reasonable χ^2 doesn't automatically imply that the procedure was valid! The procedure must always be complemented with a check for Gaussian residues.

3. (a)

$$x = \log L = \frac{\ln L}{\ln 10} \Longrightarrow \frac{dx}{dL} = \frac{1}{L \ln 10} \Longrightarrow \sigma_x = \frac{\sigma_L}{L \ln 10}$$

Similarly,

$$y = \log F_8 \Longrightarrow \sigma_y = \frac{\sigma_{F8}}{F_8 \ln 10} \qquad (1)$$

(with $\ln 10 \approx 2.303$).

(b) The q3q4 module in `finalexam.py` produces the following output for this part of the question:

```
Intercept and slope from linear fitting with y uncertainties:
-8.11 and 1.4 respectively.
```

(c) `finalexam.py` produces the following output for this part of the question:

```
Uncertainties in the intercept and slope are 0.00932 and 0.00218 respectively.
```

(d) The q3q4 module in `finalexam.py` outputs

```
The correlation coefficient between the uncertainties of the intercept and the slope
is -0.9991.
```

That is, the intercept and the slope are almost perfectly anti-correlated!

(e) The q3q4 module in `finalexam.py` produces the following output for this part of the question:

```
The reduced chi-square for the standard fit is 42.37.
```

The reason this number is so high is due to the fact that we have underestimated the uncertainties. s_y only accounts for measurement error in y; one look at the plot of y versus x shows that the spread orthogonal to the linear relationship is much wider than can be explained by the measurement errors alone.

(f) The uncertainties for the parameters are very unrealistic – the spread in the data orthogonal to the linear relationship is much larger than the uncertainties in the slope and intercept can explain. The model assumes that any departure from the linear relationship can be explained based on normally-distributed measurement errors. This is clearly not the case.

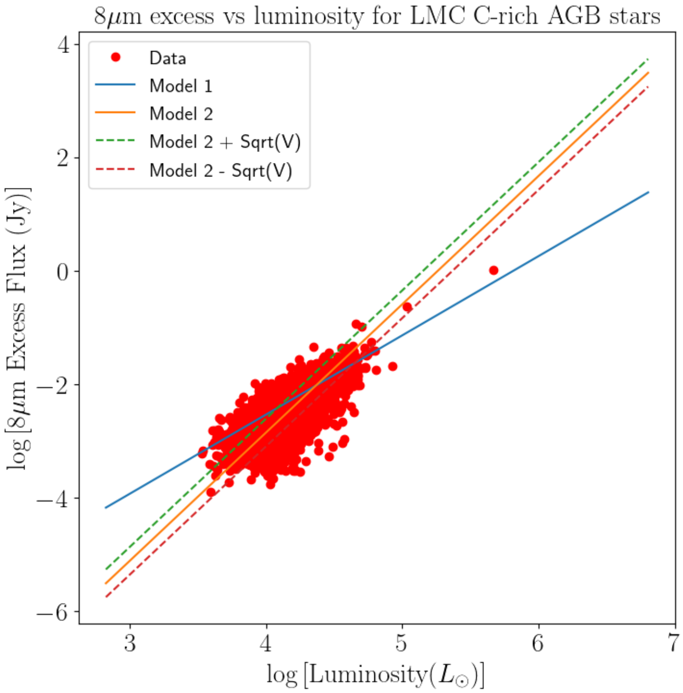

Figure 1: Comparison of the lines generated using the standard procedure introduced in Section 1 of Hogg et al. (*blue*) and the orthogonal least-squares method from Sections 7 and 8 (*orange*). We also show relations produced by shifting the latter line by $\pm\sqrt{V}$ (*dashed lines*).

.

(g) The `q3q4` module in `finalexam.py` produces the following output for this part of the question:

```
Best-fit intercept:  -8.113
Standard deviation in intercept:  0.343
Best-fit slope:  1.395
Standard deviation in slope:  0.0188
```

4. (a) Question 3c had $b = -8.11, m = 1.4$. Using these values, $\theta = \arctan m = 0.951, b_\perp = b\cos\theta = -4.714$.

(b) By trail-and-error, I settled on `invar` $= V = 0.25$.

(c) `orthofit` is called in `q3q4` module of `finalexam.py`, resulting in the following output:

```
The improved intercept and slope are -11.89 and 2.26 respectively.
```

The slope is higher, which means that the index of the power-law relationship is higher. The fit using uncertainties in both coordinates as well as an intrinsic scatter orthogonal to the linear relation does a reasonable job of reproducing the trend of the excess w.r.t. the luminosity. In addition to plotting both lines in Fig. 1, we also show the effect of the intrinsic scatter. The two dashed lines shifted up/down by $\sqrt{V}$ from the `orthofit` best-fit relationship show the equivalent of a 1σ range of intrinsic scatter orthogonal to the line. The overall scatter is due to a combination of this and the measurement uncertainties.

Discussion

Note that we have three free parameters in the problem – the intercept, the slope, and the intrinsic variance orthogonal to the line.

For a proper Bayesian analysis of the problem, we would first choose reasonable priors for these three parameters (while we could spend some time and derive the Jeffreys priors using the likelihood, this is not really required since the problem is **heavily data-dominated**) – for example, a uniform prior for θ and $b_\perp$ (remember, a uniform prior in m results in uneven coverage in 2D), and a uniform prior for $\log V$. Next, we compute the likelihood for the parameters given that the uncertainties and the intrinsic variance are Gaussian. The (log-)likelihood is given by Equation 35 in Hogg et al., which we have maximised in `orthofit` (note that `orthofit` uses the `scipy.optimize.minimise` method, so the code actually **minimises the negative of the log-likelihood** instead of maximising the log-likehood).

Finally, we compute the posterior probability for the parameter triplet. The posterior is proportional to the product of the likelihood with the priors. In order to proceed, we first have to normalise this posterior. Depending on its functional form, it may be possible to identify the underlying distribution; this may not be possible in general, in which case we may have to fall back on MCMC methods. Once the posterior is specified, we can again use MCMC methods to marginalize over the nuisance parameter V and produce the joint posterior for (m, b). We may further marginalise over one of these to obtain the conditional posterior distribution of the other.