N.M.S.U. Astronomy Department: Cumulative Exam #350
13^{th} February 2010 – Nicole Vogt

Please start a new page for each problem, and when you are done staple the pages together in order. Be clear and state all assumptions. You may use your calculators only as simple calculating machines (do not access constants other than π and e, and do not access stored formulas, including any forms of sums or fits).

There are 42 points distributed between the four questions, with question 2 being the most important (and worth 50%). I anticipate a passing score for papers marked above 75%.

1. The Tully-Fisher relation is a well-known tool used in galaxy studies. [**7 pts total**]
   (a) What type of galaxy is featured on a Tully-Fisher plot? (1 pt)
   (b) What two quantities are plotted against each other to create a Tully-Fisher relation, and what are their approximate ranges? (2 pts)
   (c) It is not unheard-of to derive different results from studies comparing the Tully-Fisher relation and the Faber-Jackson relation for a particular cluster of nearby galaxies. Why? (2 pts)
   (d) For distant galaxy samples, what does a Tully-Fisher plot tell us? (2 pts)

2. Least-squares fitting is commonly used to derive Tully-Fisher relations. Let's examine how to perform an unweighted linear least-squares fit on the following four data points: [**22 pts total**]

$$(x, y) = (1, 1), (2, 3), (3, 2), \text{ and } (4, 5).$$

   (a) Sketch the data points, and show graphically what quantity is being minimized. (3 pts)
   (b) Define χ^2 algebraically, and relate it to the plot in question 2(a). (3 pts)
   (c) Is this an over-determined or an under-determined system? (1 pt)
   (d) Motivate qualitatively the relationship between χ^2 and the least-squares fit parameters (the slope, m, and the y-intercept, b of the fitted line) – explain how we will get from χ^2 to m and b. (3 pts)
   (e) Use this relationship to derive the normal equations which define m and b in terms of various combinations of x and y. (4 pts)
   (f) Calculate m and b for this data set. (5 pts)
   (g) Prove that an unweighted linear least-squares fit will always pass through the point (x_{av}, y_{av}), where x_{av} and y_{av} are average values of the input data arrays. (3 pts)

3. One often performs both direct (y-on-x) fits and indirect (x-on-y) fits for a Tully-Fisher sample, in order to examine biases in the distribution of the observables caused by sample selection effects. Consider what would happen if you swapped your x and y arrays ($y' = x$, $x' = y$), and redid your least-squares fit. [**8 pts total**]
   (a) Begin with the expression $y = mx + b$, and massage it to express y' as a function of x', with a new slope and y-intercept m' and b'. Express m' and b' algebraically in terms of some subset of x, y, N, m, b, and χ^2. (2 pts)
   (b) Though we did not calculate them in question (2), there are errors σ_m and σ_b associated with each fit component. Express σ'_m and σ'_b algebraically in terms of some subset of x, y, N, m, σ_m, b, σ_b, and χ^2, by propagation of errors. (2 pts)
   (c) If you now performed a least-squares fit on your primed (swapped) arrays, would you expect to recover the values m' and b' derived in question 3(a)? Describe the relationship, or lack thereof, between m' and b' and the values derived from your inverted fit. Under what conditions might they agree, and under what conditions might they differ? (4 pts)

4. Let's examine 240 galaxies imaged by the SDSS, and divide them into six groups based on morphological properties. The distribution of galaxies within the six groups is 20, 46, 35, 45, 42, and 52. (You may visualize this as rolling a six-sided die 240 times, if that makes it easier for you.) **[5 pts total]**

   (a) If the galaxies are distributed randomly between groups, how many do you expect to find within each group (what is the expected number, E)? (1 pt)

   (b) Calculate χ^2 for the distribution. (3 pts)

   (c) Could the galaxies in fact be distributed randomly? (1 pt)

N.M.S.U. Astronomy Department: Cumulative Exam #350
13^{th} February 2010 – Nicole Vogt

Please start a new page for each problem, and when you are done staple the pages together in order. Be clear and state all assumptions. You may use your calculators only as simple calculating machines (do not access constants other than π and e, and do not access stored formulas, including any forms of sums or fits).

There are 42 points distributed between the four questions, with question 2 being the most important (and worth 50%). I anticipate a passing score for papers marked above 75%.

1. The Tully-Fisher relation is a well-known tool used in galaxy studies. **[7 pts total]**

   (a) What type of galaxy is featured on a Tully-Fisher plot? (1 pt)
   A Tully-Fisher plot contains spiral galaxies (and occasionally S0s).

   (b) What two quantities are plotted against each other to create a Tully-Fisher relation, and what are their approximate ranges? (2 pts)
   We plot the log of the circular velocity, measuring disk rotation from spatially resolved optical or infrared emission line widths or from 21-cm line profiles, as a function of absolute magnitude. For normal galaxies, these range from roughly 100 to 350 km sec^{-1} (extending down to even lower velocities for smaller galaxies), with absolute magnitudes of -16 to -22 (varying with bandpass).

   (c) It is not unheard-of to derive different results from studies comparing the Tully-Fisher relation and the Faber-Jackson relation for a particular cluster of nearby galaxies. Why? (2 pts)
   The Tully-Fisher relation derives a cluster distance from the population of spiral galaxies in or near a cluster, while the Faber-Jackson relation is used similarly to extract a mean cluster distance from the elliptical population. These two galaxy populations may not be virialized, or share a common dynamical center, as is the case with substantial spiral infall along a particular filament.

   (d) For distant galaxy samples, what does a Tully-Fisher plot tell us? (2 pts)
   With distant galaxies, peculiar velocities become negligible and redshifts-derived distances are used. Galaxies are placed on a Tully-Fisher relation, and offsets are used to study redshift-dependent galaxy evolution in fundamental parameter space, as a function of evolutionary epoch, galaxy type, and suchlike.

2. Least-squares fitting is commonly used to derive Tully-Fisher relations. Let's examine how to perform an unweighted linear least-squares fit on the following four data points: **[22 pts total]**

$$(x, y) = (1, 1), (2, 3), (3, 2), \text{ and } (4, 5).$$

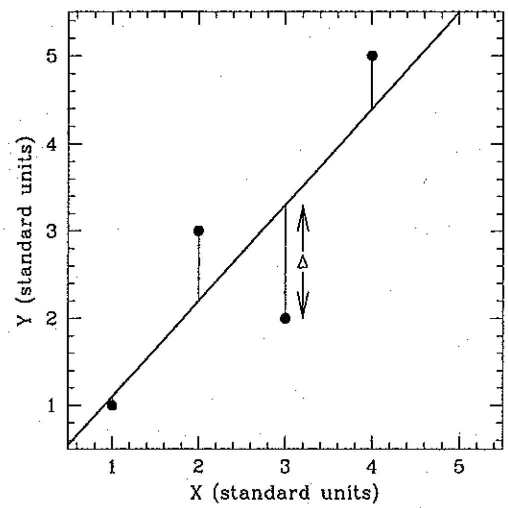
   (a) Sketch the data points, and show graphically what quantity is being minimized. (3 pts)
   A least-squares fit minimizes the sum of Δ^2, the deviation in y of each point from the best-fit line.

   (b) Define χ^2 algebraically, and relate it to the plot in question 2(a). (3 pts)

$$\chi^2 = \sum_{i=0}^{N} \Delta_i^2 = \sum_{i=0}^{N} \left[y_i - (mx_i + b)\right]^2 .$$

   (c) Is this an over-determined or an under-determined system? (1 pt)
   There are only two unknowns, and four equations of the form $y_i = mx_i + b$ (one per point), so the system is over-determined.

   (d) Motivate qualitatively the relationship between χ^2 and the least-squares fit parameters (the slope, m, and the y-intercept, b of the fitted line) – explain how we will get from χ^2 to m and b. (3 pts)
   We seek to minimize χ^2, and so will take partial derivatives with respect to m and b and set them equal to zero.



(e) Use this relationship to derive the normal equations which define m and b in terms of various combinations of x and y. (4 pts)

Evaluating the partial derivatives, we see that

$$\frac{\partial\chi^2}{\partial b} = \frac{\partial}{\partial b}\sum_{i=0}^{N}\left[y_i - (mx_i + b)\right]^2 = -2\sum_{i=0}^{N} y_i - (mx_i + b) = 0$$

and thus that

$$\sum_{i=0}^{N} y_i = m\sum_{i=0}^{N} x_i + bN.$$

In similar fashion,

$$\frac{\partial\chi^2}{\partial m} = \frac{\partial}{\partial m}\sum_{i=0}^{N}\left[y_i - (mx_i + b)\right]^2 = -2\sum_{i=0}^{N} x_i\left[y_i - (mx_i + b)\right] = 0$$

and thus

$$\sum_{i=0}^{N} x_i y_i = m\sum_{i=0}^{N} x_i^2 + b\sum_{i=0}^{N} x_i.$$

We can combine these equations to solve for m and b.

$$b = \frac{\sum_{i=0}^{N} x_i^2 \sum_{i=0}^{N} y_i - \sum_{i=0}^{N} x_i y_i \sum_{i=0}^{N} x_i}{N\sum_{i=0}^{N} x_i^2 - \sum_{i=0}^{N} x_i \sum_{i=0}^{N} x_i},$$

and

$$m = \frac{N\sum_{i=0}^{N} x_i y_i - \sum_{i=0}^{N} x_i \sum_{i=0}^{N} y_i}{N\sum_{i=0}^{N} x_i^2 - \sum_{i=0}^{N} x_i \sum_{i=0}^{N} x_i}.$$

(f) Calculate m and b for this data set. (5 pts)

$$b = \frac{30 \times 11 - 33 \times 10}{4 \times 30 - 10 \times 10} = 0.$$

$$m = \frac{4 \times 33 - 10 \times 11}{4 \times 30 - 10 \times 10} = 1.1.$$

(g) Prove that an unweighted linear least-squares fit will always pass through the point (x_{av}, y_{av}), where x_{av} and y_{av} are average values of the input data arrays. (3 pts)

We begin with the expression derived from the partial derivative of χ^2 with respect to b,

$$\sum_{i=0}^{N} y_i = m\sum_{i=0}^{N} x_i + bN.$$

Dividing through by N,

$$\frac{\sum_{i=0}^{N} y_i}{N} = m\frac{\sum_{i=0}^{N} x_i}{N} + b,$$

and substituting in x_{av} and y_{av},

$$y_{av} = mx_{av} + b$$

which demonstrates that (x_{av}, y_{av}) clearly lies along the fitted line.

3. One often performs both direct (y-on-x) fits and indirect (x-on-y) fits for a Tully-Fisher sample, in order to examine biases in the distribution of the observables caused by sample selection effects. Consider what would happen if you swapped your x and y arrays ($y' = x$, $x' = y$), and redid your least-squares fit. **[8 pts total]**

(a) Begin with the expression $y = mx + b$, and massage it to express y' as a function of x', with a new slope and y-intercept m' and b'. Express m' and b' algebraically in terms of some subset of x, y, N, m, b, and χ^2. (2 pts)

$$x' = my' + b$$

and so

$$y' = \frac{1}{m}x' + \frac{-b}{m},$$

so that

$$m' = \frac{1}{m}, \quad \textit{and} \quad b' = -\frac{b}{m}.$$

(b) Though we did not calculate them in question (2), there are errors σ_m and σ_b associated with each fit component. Express σ'_m and σ'_b algebraically in terms of some subset of x, y, N, m, σ_m, b, σ_b, and χ^2, by propagation of errors. (2 pts)

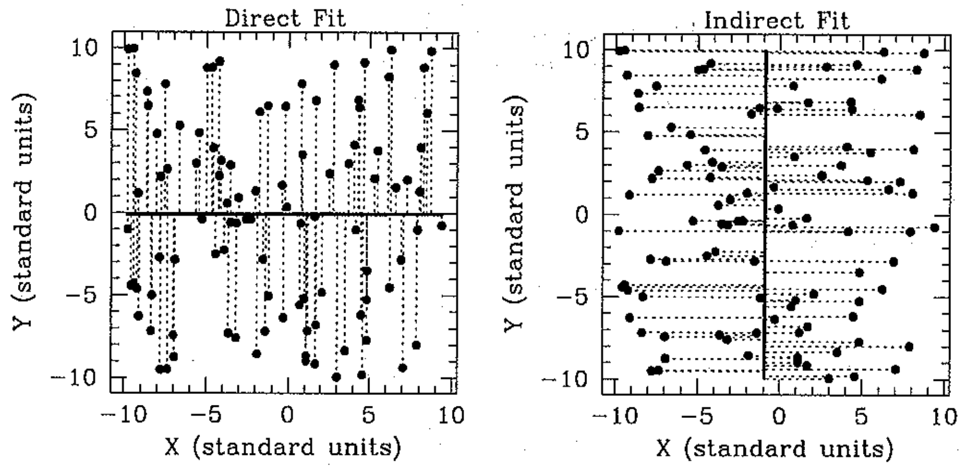
$$\sigma'_m = \frac{\sigma_m}{m^2}, \quad \textit{and} \quad \sigma'_b = \sqrt{\left(\frac{\sigma_b}{m}\right)^2 + \left(\frac{b}{m^2}\sigma_m\right)^2}.$$

(c) If you now performed a least-squares fit on your primed (swapped) arrays, would you expect to recover the values m' and b' derived in question 3(a)? Describe the relationship, or lack thereof, between m' and b' and the values derived from your inverted fit. Under what conditions might they agree, and under what conditions might they differ? (4 pts)

For a tightly constrained set of data with a linear relationship between x and y and little scatter, the direct and inverse fits will agree very well. The more scattered the data are, the more outliers will act to drive each fit toward a horizontal line (direct fit) or a vertical line (inverse fit).

The figure shown below is drawn with 100 randomly selected (x, y) data points varying uniformly between -10 and 10 (the same points lie within each window), and illustrates this point.

4. Let's examine 240 galaxies imaged by the SDSS, and divide them into six groups based on morphological properties. The distribution of galaxies within the six groups is 20, 46, 35, 45, 42, and 52. (You may visualize this as rolling a six-sided die 240 times, if that makes it easier for you.) **[5 pts total]**



(a) If the galaxies are distributed randomly between groups, how many do you expect to find within each group (what is the expected number, E)? (1 pt)

We expect each group to contain an equal fraction of the 240 galaxies, or 40 galaxies per group.

(b) Calculate χ^2 for the distribution. (3 pts)

$$\chi^2 = \sum_{i=0}^{N} \left(\frac{E - O}{\sigma}\right)^2 = \left(\frac{20^2 + 6^2 + 5^2 + 5^2 + 2^2 + 12^2}{40}\right) = 15.85, \quad as \quad \sigma = \sqrt{E}.$$

(c) Could the galaxies in fact be distributed randomly? (1 pt)

As $\chi^2 >> N$, it seems unlikely. For an even, random distribution, $\chi^2 \approx N$ (counts should deviate from the expected value by a bit less than 1σ on average, not by almost 2σ).