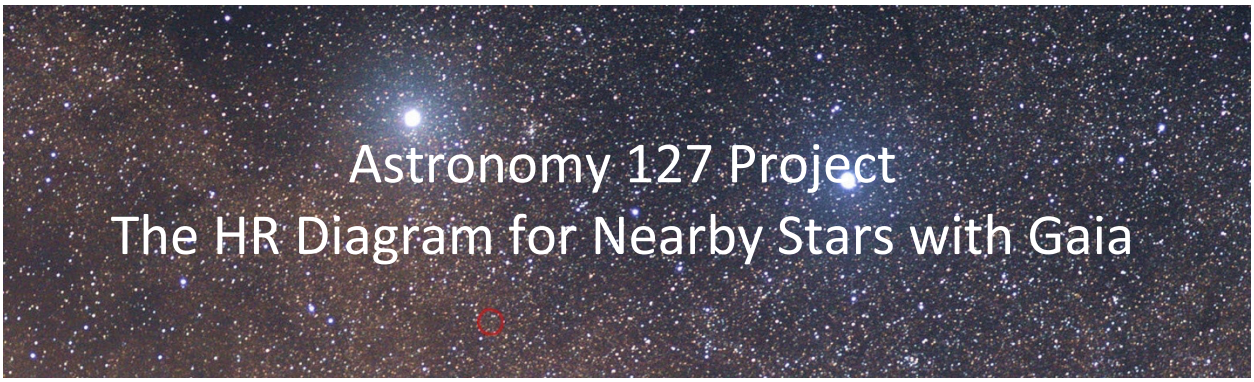

Astronomy 127 Project
The HR Diagram for Nearby Stars with Gaia

The Gaia mission does precision astrometry. It is currently observing from space, a mission of the European Space Agency. While Gaia data is designed for dynamical studies of stars & stellar systems, it also has photometric data and color information.

This project will allow you to use Gaia data as you would for a research project. You will make your own HR diagram for nearby stars, and analyze it. You will determine a mass-magnitude (mass-luminosity) relation and apply it to predict masses for the stars. You will use these masses and the Gaia counts to fit an IMF.

In the first part of this document, you will be led through the process of downloading the data from the Gaia archive, and executing the R script. You will also read in data on low mass stars from Mann+19 to fit a mass-magnitude relation. In the second part of the document, what is expected for the resulting paper will be described.

The paper is due the end of the 10th week of the quarter. There are no limits to submissions, so you can continue to edit and tweak your paper right up till the due date. It is recommended that you run the script fairly early, to give yourself time for writing and analysis.

You may consult with others but must modify and run your own script, produce your own plots and write your own paper. Identical papers will be considered a violation of the student code.

Running the R Script

It is assumed that you have already installed R and RStudio. If you ever want to clean up your space & start fresh, clicking the "X" in the tabs closes scripts and datasets in the upper left window; control-L cleans up the console in the lower left window; executing the command rm(list=ls()) will clear all datasets and variables in the environment in the upper right window; and clicking the broom will clear everything in the plots area in the lower right window.

I have included line numbers in the script for you to consult but these might be off by a little.

Step 1. Downloading data from the Gaia archive.

a) Go to https://gea.esac.esa.int/archive/
Pick the "Search" tab and under that the "Basic" tab. This brings up a window. Choose the "Position" tab at the top. Leave the first part blank. Start with: Extra conditions: Click on "Add condition" and from the list, pull down "parallax" and next to that box "≥" and insert "70" in the box (70 mas)

b) Choose display columns: You want "sourceid", "ra", "dec", "parallax", "phot_g_mean_mag", "phot_bp_mean_mag", "phto_rp_mean_mag" (these three In far right column near the bottom", and "bp_rp". These are observed magnitudes, in green, blue, and red filters and blue minus red color. (Gaia filters are not the same as the Cousins UBVRI filters: here are the Gaia passbands, https://www.cosmos.esa.int/web/gaia/iow_20180316)

At the bottom is "Max results" pull down the number "1000".

Then hit the button "submit query". This will select stars with parallaxes > 70 mas.

Underneath the table, which displays the first 20 rows, you will see "Download results". Select "csv" and then click on "Download results". You should have a little over 900 sources (rows).

Step 2. Reading the Gaia data into RStudio, adding variables, and plotting an HR diagram.

Now that you have the csv file from the Gaia archive, download it into RStudio. We will add some variables to this data frame, which we will call, dat.

Either double-click on the R script or open RStudio and open the script. In RStudio you will see four windows, with the script in the upper left. You will modify the basic R script you were given in the assignment, as directed below. Commands are executed by inserting the cursor in the typing command-enter (⌘⏎) in R (for Macs). It might be control-enter on Windows machines. You can always download the original script if you mess this one up, but it is recommended that you save work in intermediate versions. The script tab shows as red if you have not saved, so it is possible to bail out and restart from your last save if you mess something up.

Comments are set off by hashtags (#) in R. There are many comments in this script. Read the comments. Things that you are to modify will have a long line of hashtags to set them off.

a) Install packages. You don't need all of these. Wait till they are installed, then execute the library(ggplot2) command in the script. ggplot2 should then be checked under "packages" in the lower right window. If you are returning and R has saved your environment & ggplot2 is already checked, you don't need to do this. If you try to do a ggplot later and it doesn't recognize ggplot, then try the library(ggplot2) command again.

b) Set the current working directory, where the R script and the csv file can be found. Edit the command

setwd("/Users/myusername/directory/subdirectory")

or you might also use, for Macs, or unix, or linux?

setwd("~/subdirectory")

(where "~/" means home directory) to put the path to your working directory in between the quotes. Quotes set off string arguments (known as "character" type) in R. Be sure the quotes are there. If you notice, in the install.packages command above that, there are also quotes. (You do not have to modify that command.)

c) Next, modify the next command to specify whatever name you gave your csv file.

dat = read.csv("filename")

Execute this command. read.csv() defines a data frame, dat, which is like a spreadsheet. The rows are different "observations" (here, stars) and the columns are variables. Execute the next commands to omit missing data (na.omit) from the file.

d) The Gaia data has parallaxes. It is convenient to also have distances in the data frame. Variables in R are denoted by the data frame name with a "$variable" appended. Modify the expression on line 45 to correctly define the variable dat$Distancepc from the Gaia variable dat$parallax in the data frame dat. Note the units of parallax that Gaia uses and make sure your distances make sense by clicking on the dat line in the environment window at the upper right, which will display the dat data frame as a tab in the command window. Do the distances look good? Then return to the script.

e) Since we have distances, that means we have absolute magnitudes. The Gaia data, variable dat$phot_g_mean_mag, is apparent magnitudes. Modify the definition of absolute G magnitude, dat$Magnitude, in the script at lines 54-55 to put in the correct expression for relating apparent to absolute magnitude. Again check these in dat.

f) You will make an initial plot of Magnitude ~ bp_rp from dat in line 65. It won't look like an HR diagram. You need to modify the x and y axis limits in the ggplot command starting on line 78. The message "Removed XX rows containing missing values" means that you need to extend your limits. Also add labels for the x and y axes. You might make note of these, because you'll make more than one plot. Now it looks like an HR diagram. A really nice one.

g) The sample of nearby stars is a mixed bag of many different kinds of stars. We can separate them into classes. Gaia colors can be related to the standard B-V colors and thus to spectral classes, as you saw in Problem Set 1. I've already calculated the Gaia colors separating the classes and on line 102, have some code to classify the stars based on absolute magnitude and color. This is good enough for our purposes. The separation isn't perfect between "LT" and main sequence types, this could be done better, but we'll mostly consider these together

anyway. Execution of the lines 102-108 will give a new variable in the data frame dat, which is dat$classif, a character variable, which identifies the spectral classification, OBAFGKM, LT, or WD. After that, starting line 112, we define four datasets containing MS+LT stars, LT, MS only, and WD; these data frames are called mainseqlt, ltdwarfs, mainseq, and whitedwarfs. There is no action item for you in this g) part, beyond executing the command.

h) At line 127 you need to modify the x and y limits and labels for the plot. Use the values from part f) if you want. Save the plot if you wish.

i) At line 138 you will plot the main sequence subset, which you can save or not as you please. Needs the same x, y cutoffs and labels.

j) At line 152 you will plot the stars, with color coded to classification, and put a title on it. Put the usual x,y cutoffs, labels, and modify the chart title, and replace anything with "XXX". Also, you will need to pick a color palette for color_brewer. Pick whatever you want. Use G**gle to do look up color_brewer. Take a minute to admire your work. SAVE THIS PLOT#1 (152-161)

Step 3. Star counts.

a) Now that you've compiled the Gaia data set into the data frame, dat, start the analysis part of the script by doing plots of star count vs. distance. This is a histogram, using the hist() function, on line 174. This plots the stars in each bin distance, in each shell bin going out. You can control the number of bins with an optional addition of ", breaks=XXX", where XXX=number of bins+1, but hist does an optimal binning given the data ("Sturges" count).

b) This histogram might not be the best thing for analysis; if we want to determine stellar density, we want the total number of stars within a volume defined by R, $N_*(R)$, rather than the counts in a shell, R to R+dR, which is the plot in part 3-a. Instead, you want the cumulative number of star counts at distance R. To do this cumulative sum we save the histogram into the object disthist, which is a list. Then do a cumulative sum of the variable disthist$counts within disthist with the command in lines 184-187. This redefines the variable disthist$counts to reflect the cumulative sum. In line 188 you plot this cumulative sum, which is $N_*(R)$.

c) One final thing to do here is to add a curve() to this describing how the number of stars grows with R. Think about this: how should $N_*(R)$ vary with R? Substitute a number in the expression in line 193 where there is XXX and execute it. How well does this curve fit? SAVE THIS PLOT#2 (200)

d) Now separate the sample into spectral types. We'll do the counts of the different spectral types. ggplot will plot the categorical variable dat$classif along the x-axis. All you have to do is choose a fill color for the plot. You can go with something obvious and/or garish or you can look for plot colors online with G**gle. Fill in the XXX at line 226 with something that works (make

sure it is within quotes) and plot numStars. Note that the object numStars saves counts for the different categories. SAVE THIS PLOT#3 (229)

Step 4. Mass-luminosity relation.

If we can assign masses to the stars, we can do a mass function. This is huge. However, we only have the mass of a star if it is in a binary. There are definitely binaries within the sample although they may not be obvious. (Think: what are binaries doing to the HR diagram). However, what we need is a mass for every star. We can assign masses because on the main sequence, mass is related to magnitude, and we have a Gaia G magnitude for each star. What we will do next is determine a mass-magnitude relation.

To determine a mass-magnitude relation, we'll use the masses determined by Mann et al. in their 2019 paper. (I may have used this table in lecture to demonstrate the cut-and-paste-from-pdf method of acquiring data). I downloaded the Gaia magnitude & color information for this sample of stars, so we have mass-Gaia G magnitude for the sample. We'll plot those stars in the Mann sample and see if there is a correlation we can use to predict masses for the stars in our Gaia sample.

In this step, at line 254, you will read in the csv file redlist.csv. redlist.csv is the Mann sample with my additions. Check the working directory. The next line omits any rows with missing data. A new data frame, redstars, should show up in the environment in the upper right window.

a) Now plot the variable redstars$Mtot vs. redstars$M_G in line 260. This is a pretty good correlation between M_{tot} and M_G. Warning: M_{tot} is total mass of the binary, since that is what you get from Kepler's 3rd law. But that should be within less than a factor 2 of the mass of the dominant star. It will introduce a little scatter. We'll just go with this, close enough.

b) Next we will do a linear regression on the redstars$Mtot vs. redstars$M_G relation, using the R function lm(). (See documentation on this important function at https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm) We will save the results of lm() into the object massmodel in line 267. This is a simple linear fit, for which the syntax is

```
lm(y ~ x)
```

To add the fit to our previous plot of the data, which is still in the window, use the abline() function, which adds to the previous plot, and execute line 268:

```
abline(lm(redstars$Mtot ~ redstars$M_G))
```

which will plot the fit. (We could also have typed abline(lm(massmodel)) since we saved the fit into object massmodel.) The fit is pretty good; there is some scatter; why might there be scatter?

To look at the parameters of massmodel, on line 270 do

summary(massmodel)

This gives the intercept, slope, all the information on the regression. It assigns a significance to the correlation based on a 95%? confidence interval, and rates the significance with stars. Here, it gets three stars for a highly significant correlation. Well, surprise, we already knew this.

c) We can maybe do a little better than linear regression. The problem is that if you run this you may end up with negative values for the faintest stars, since there is an indication that the relation flattens out for the faint stars. (You can try it and see). Let's try something else: a loess fit, "locally weighted scatterplot smoothing". This is the equivalent of introducing extra parameters, generally not the best thing to do, but may be warranted here since we have no a priori reason to expect it should be a strict linear function. You want the closest thing to a straight line, the smoothest curve that goes through the data. So let's use the neighboring data points to estimate masses.

Execute lines 294-295. You can see that this loess fit captures the flattening at the low mass end and is probably better than the linear. If this were a real research project we would fiddle with the kernel in loess, and pick a goodness-of-fit criterion to apply and do statistical tests. But this basic eyeball fit serves our purposes. SAVE THIS PLOT#4 (294)

Let's save the loess model into a new object, massmodel2, in line 305.

d) Now we can predict masses for our Gaia sample using massmodel2, the relation between magnitude and mass using the loess model. The predict() function in R can use a model to do this. (We could use either massmodel, the linear fit, or massmodel2, the loess fit, to do the prediction but we prefer the loess fit, which fits the data better).

However, we will not put these masses into dat because dat also contains white dwarfs and this relation does not apply to them. So we'll put it into the data frame mainseqlt, which excludes the white dwarfs. Execute the command in line 310,

mainseqlt$mass = predict(massmodel2, mainseqlt$Magnitude)

The predict function puts the predictions of the model as the new variable, based on the data in mainseqlt$Magnitude. The new variable, mainseqlt$mass, will appear in the data frame mainseqlt.

Step 5. The luminosity function and the present-day-mass function (PDMF).

Now that we have a data set of stars with masses and magnitudes, and spectral types (a valuable dataset!) we can analyze the local stellar luminosity function and mass function.

a) Redo the histogram of magnitude by executing line 319. The histogram is stored in the list object magCounts. You can play with the bins if you want but 30 is good. SAVE THIS PLOT#5

b) Do a similar histogram for mass. The mass histogram is stored in massCounts, line 330. The PDMF will come from the massCounts histogram data. Don't save the plot just yet.

c) Next we will determine an IMF. For this, you should read the first two pages of the paper by Bochanski+20. We'll use the definition based on number counts, N, per mass, $\psi(M) = dN/dM \sim M^{-\alpha}$. The tilde means "proportional to". We want to determine the power law index, α, this specifies the "IMF". Rearrange this equation for $\psi(M)$ to get $dN \sim M^{-\alpha} dM$. Since we are using equal mass bins, the bins are constant, and just go into the constant of proportionality, so this simplifies to $dN \sim M^{-\alpha}$. Take the log of this equation. $\log(dN) \sim \alpha \log M$.

The histogram in massCounts gave us dN vs M in equal mass bins, so we just need to take the logs of the two variables and find the slope of their relation. Histograms are saved as R list objects, which are difficult to work with. What we need from the histogram object are dN and the value of M it corresponds to, and those are the variables massCounts$counts, and massCounts$bins, respectively. Note that they are the same length by examining massCounts. We pick those two variables out and cbind ("column bind") those into a data frame called countsMassbins in line 344. (If you don't specify "as.data.frame" it saves it as a matrix instead; data frames are easier). Then for convenience we relabel the columns as "massbin" and "counts" using the command colnames().

So now we have our dN (countsMassbins$counts) and M (countsMassbins$massbin). We need the logs of these variables. So we add two variables (columns) to our data frame, countsMassbins$logcounts and countsMassbins$logbin, which are log dN and log M, respectively, in lines 352-353. We can plot these two columns to see how they look. Above a mass of $0.5M_{sun}$, it is a linear function, so linear regression can be done.

Now you have to do some work; you need to specify the linear regression to fit a slope to log(dN) vs. log(M).

At line 364 at the "salpeter =" (this saves your fit into object "salpeter"), write an expression in R that will do a linear regression model for logcounts ~ logbin, with data=countsMassbins. Bear in mind that you only want to include data with mass > $0.5M_{sun}$; it is not the same power law below that mass; try adding ", subset=(massbin > 0.45)" to your command. The subset means that the linear fit is only for masses higher than 0.45 M_{sun}.

summary(salpeter) in line 365 will give you the intercept and slope of your linear regression line. Plot this line over your data with the abline() command, line 369, as modified in your script. Admire your fit. Save this plot if you wish, it is optional. (line 369 over line 355 plot)

d) A final plot to make: overlay your plot on the histogram of masses directly, rather than their logs. Put your alpha into the "points" command on line 371 and SAVE THIS PLOT#6 (385-386)

e) There are commands starting line 397 to make a data frame containing the mass per bin, rather than the counts per bin, which was done above. This shows us how much mass is in each mass bin (M*dN). We check that the total masses summed over all bins, totalmass, agrees with the total mass summed up from our table of stars, totalmass2. It agrees pretty well (there are errors associated with binning, but the result is within a fraction of a percent, so we have a good mass function.) Make a note of total stellar mass, totalmass2, line 422.

You are now done with the script. SAVE THE WORKING SCRIPT! (Make it read-only!) Also you may want to save dat and mainseqlt into csv files, they have useful data that you've created.

Writing the paper

Please follow the traditional style for a journal article on astrophysical data: Introduction, Observations and Analysis, Results, Discussion, Summary. Data papers are usually centered on the results, plots, tables, and graphs. This document will guide you through the presentation of your results in this form, and things to include. However, be creative. This is your paper.

While you may consult with others, you need to demonstrate that you yourself have done the script and the plots and understand the findings and their implications. You cannot share writing with anyone else. The text and plots must be your original work based on the script.

I have given you some references to consult.

Mann+19: You should reference this because you will use this sample to obtain a mass-magnitude (mass-luminosity) relation.
Bochanski+10: Has a nice summary of what the IMF is on the first two pages.
Kroupa+13: Is a scary big review. Do NOT try to read the whole thing. I point you to a few relevant pages later, but pages 58-70 are most relevant. Ignore all the high mass and cluster sections.

You will construct your paper around the minimum six plots that you will have accumulated from your analysis (include others if you want).

1. A beautiful HR diagram for nearby stars.
2. A plot of cumulative number of stars vs. distance, R
3. A bar plot of number of stars vs. spectral classification
4. The plot of loess fit of Gaia G magnitude vs. mass
5. A histogram of absolute G magnitudes
6. A histogram of mass counts with the fit to the mass function, your IMF

7. You will have numbers of different spectral types within the volume (plot 3)
8. You will also have a total stellar mass within the volume

Your paper consists of the background and discussion of your findings.

A guide to your paper follows. The parts labeled with ➤ are required for the median grade (~B+).

Introduction Section:

- Always have a clear first paragraph stating the purpose of your paper. It can be short, but should be very clear.

- You should have a paragraph describing the Gaia mission and archive.

Observations and analysis section:

- You must describe how you got your data from the Gaia archive.
  You must list the software you use (R and RStudio).

- You must describe how you got masses for your sample. This was done when you read in the file redstars, reference the Mann+19 paper here. Describe how you fit this relation of Magnitude and mass, and did the prediction. You could also describe the linear fit that you did not use, because loess seemed better (this would be done differently for a journal article).

Results section:

- Present your HR diagram, Plot#1. Comment on it.

- Present your plot of number of stars vs spectral class, Plot#3. What are the relative fractions of the different classes of stars?

  Are you surprised at how many WD there are? Do you usually see this many on HR diagrams? (Be aware that Gaia only goes to a brightness of magnitude 3, so Sirius and Alpha Cen are not on this plot. They were observed by Hipparcos.)

- Present Plot#2, the cumulative number of stars with distance, R. What is the density of stars, per cubic parsec, in the solar neighborhood? What is the average separation of stars?

- Make a table with the following numbers. What is the density of M stars per cubic parsec? K stars? G stars? LT stars? White dwarfs?

Leading up to the mass function….

- Show Plot# 5, the histogram of G magnitudes. Compare to Figures 9 and 15 in Kroupa+. You can see from their Figure 9 that Hipparcos did not go nearly deep enough to see these features, but Gaia easily can map out this function. (Don't worry too much about the exact function plotted; it is the form that is important. We're working on luminosity within 4π/3 (14)^3 pc^3) The features, the dips and peaks, we see in this plot are real, have names, and have a basis in stellar structure. These are discussed on p. 60 of that reference on the steepening near M_V~10 (V is close to G, so take M_V ~ M_G for this comparison).

- Explain how you predicted masses for the Gaia sample, based on a mass-magnitude model based on the data from the Mann paper that I provided. Show the loess fit that you used to do this fit, Plot#4, and explain that you used the prediction feature in R to compute masses from this model.

(You could also show the linear fit and explain why you didn't use it). You could compute a mass-to-luminosity ratio also, but often what is presented in the literature as "mass-luminosity" relation is really "mass-M_v". This is directly comparable to the function shown in Kroupa+13, Figure 11. Compare and state if it is consistent or not.

Finally the mass function…

- Plot the histogram of masses, Plot#6, and the curve plotted over it, with the exponent that you used for the mass function. Is the mass function consistent with what is found in Bochanski+19 & in Kroupa+? I suggest reading "Main results" in Kroupa+13, p. 70, and also p. 101, as well as the Table in Bochanski. State if your plotted curve is consistent or not.

You are now in a position to compute the mass density, ρ, of stars of the solar neighborhood. Compare to what Kroupa+13 says on p 102. (Although if you can't parse the Kroupa, I understand. It is very dense.)

This analysis includes LT dwarfs; does that make a difference? See "Main Results" in Kroupa+13, p. 88, but recognize that this reference is relatively early in the very low mass star history.

Discussion:

The discussion is an analysis of your results, which are your data and fits. Generally this is where you compare your results to theory. Here I would suggest that you could discuss "what we could do better" in analysis (to make this publishable! You can already see that the results look very good).

- Mention at least two ways that this study could be improved and why they are

important. Some possibilities: extinction? completeness? binaries? other selection effects (see the discussion on "Biases" starting page 24 in Kroupa+13).

An interesting test for stellar models is the number of white dwarf stars. Are the numbers of WD observed in the solar neighborhood consistent with what we expect based on our IMF? We will deal with this in a problem set later on, so you can defer this paragraph till the end.

Be sure to include in your discussion why this is timely and relevant. Remember, this is Gaia data, it is new. You can do the IMF in the local volume much better than when the Kroupa+13 review was written. Then, there is the consideration of exosolar planets. How important are faint stars to the search for planets & their characterization? This is where you would put the discussion of why this is important and how it could influence current & future work. The introduction to Mann+19 has some things along this line to consider. If you are really into exoplanets or alien worlds, or Star Trek, and the possibilities for life in the local neighborhood, go crazy.

Summary:

The summary should be a brief review of what you presented and your findings, 1-2 paragraphs or maybe a bullet list. Many readers will read only the Introduction & Conclusions so this part of the paper is extremely important, along with the figures.

References:

You should have at least the three references you were given, and be sure to cite them in the text where relevant. Follow the format that you see in the Mann+19 reference.