Gaia DR1: Velocity and Correlation Analysis

Brent Tan$^1$, Stephanie O’Neil$^1$

$^1$ Department of Physics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

October 18, 2016

ABSTRACT
With Gaia’s first data release comes a large amount of new data on stars in the Milky Way. We were interested in a subset of these that were cross referenced with the Tycho-2 catalogue. In this project, we look at two ways of analyzing the data - spatial correlation functions and velocity profiles. To find a velocity profile, we looked at a subset of stars along the sun to galactic center line to determine their velocities and distance from center of the galaxy. We found a loose correlation between distance and velocity. For the spatial correlation, we compute the two point correlation function of the data set and found that stars cluster more at smaller scales as expected.

Key words: Gaia DR1, velocity profile, two point correlation

1 INTRODUCTION
Gaia is a satellite launched by the European Space Agency in December of 2013. After it reached its orbit location, the L2 Lagrange point, it scanned the Milky Way, collecting data on over one billion stars and other astronomical objects. By the time the satellite has finished its data collection in 2022, the project will have astrometric, photometric, and radial-velocity catalogues, variable- and non-single-star solutions, source classifications, and an exoplanet list among other information Brown & et al. (2016). This data will vastly expand our understanding of our galaxy, and has the potential to lead to significant developments in our understanding of stellar evolution, the history of our galaxy, and the structure and contents of the Milky Way. As the most precise collection of astronomical data so far, our ability to track near-Earth objects and identify other objects, such as exoplanets identified by the wobbling of stars, will be greatly increased.

The first data release, which occurred on September 14, 2016, includes the positions and G-magnitude for all the surveyed stars, as well as positions, parallaxes, and proper motions for those stars that could be cross-referenced with the Tycho-2 catalogue, about 2 million stars. With this information, many types of analysis can be done. One type of analysis is a rotation curve relating the velocity of the stars and the distance from the center of the Milky Way. This analysis is often done on other galaxies where we have the ability to resolve points along the rotation curve. Rotation curves can lead to information on the matter content of a galaxy, as is done by McGaugh, Lelli, and Schombart (S. & Lelli 2016).

This report will be divided mainly into two sections. Section 2 will deal with reflecting on the results and suggesting future directions. Section 3 will deal with the motivations behind investigating the rotation curve. Section 4 will deal with the methods involved in finding the rotation curve. Section 5 will deal with analyzing the results from attempting to find the velocity profile. Finally, Section 6 will include the containing remarks.

2 BACKGROUND (CORRELATION)

2.1 Clustering
When looking at spatially distributed point cloud data, a natural question to ask is: How are the points distributed? One way to reframe this question is to examine the clustering of points at different length scales. However, we want to have a quantitative measure of clustering that is a function of separation distance, and this is where the two point correlation function comes in.

2.2 Two Point Correlation Function
The spatial two point correlation function is amongst the simplest statistical methods for studying clustering, and has been effective for analysis of astronomical datasets together with its Fourier Transform, the power series. It is most widely used in studies involving large scale structure, where the clustering of galaxies is examined. However, it is rarely used in studies of galactic structures. The next subsection describes the motivations behind trying to apply the two point correlation at these smaller scales. Essentially, the two point correlation function is a measure of the overdensity of the number of points separated by a certain separation distance over a range of length scales, and is a hence a function of separation distance. The overdensity is with respect to the equivalent distribution
of randomly distributed points. We hence measure the likelihood of points being a certain distance apart.

2.3 Motivation

As previously mentioned, the two point correlation function is mostly used in the study of large scale structure and is not commonly used to look at structure on a galactic scale. However, the Milky way provides us a unique opportunity to study galaxy data in detail, since with the recently release Gaia dataset, we now have an overwhelming amount of data about the spatial distribution of the numerous stars in the Milky way, on a scale that we cannot get when studying any other galaxy. Previous studies have also shown that there is interesting and significant spatial structure in our galaxy, for example in the form of stellar overdensities that would be informative (?)juri). These studies will help to test constraints for models of galaxy formation and evolution and help us achieve a better understanding of galactic physics.

2.4 Previous Work

While the use of correlation functions to study clustering on the galactic scale is not widespread, there is nonetheless still a significant body of previous work on the subject. Early work found excess clustering at small scales below 25 pc, while other made use of variants such as the phase space correlation function and the angular two point correlation function. It was also used to study spatial and kinematic substructure in stellar halos.

However, the most relevant recent work involves the use of the spatial correlation function to directly probe galactic structure (Mao et al. 2015). The paper looks at a sample of 18000 stars in various pencil beam line of sights up to a distance of 1-3 kpc, and used mocks to test the sensitivity of the correlation function to various effects such as survey geometry. They found that below 50 pc, there was always excess clustering, but on lager scales, it was model dependent, and more specifically dependent on the density gradient predicted by the model. There was also an interesting change in behavior at 500 pc. This model dependence was due to their comparison to a model distribution of stars rather than a random one when computing the correlation function.

3 CORRELATION : METHODOLOGY

3.1 Data

The data points we used were extracted from the Gaia DR1 data. Specifically, we used a subset cross matched with the Tycho2 catalog. This gave us approximately 2 million stars, with information on their ra,dec, and parallax. Next, we trimmed the data by first ignoring stars with negative parallaxes and then transforming to Cartesian coordinates. Upon examining the data, we found that most of the data was concentrated in a smaller volume, with some outliers much further out. We hence decided to use data in a smaller 4 kpc box. The number of stars at this point was still close to 2 million. We then generated 2 million data points randomly within the box to comprise our random data set. Once we had these two data sets, we then simply ran a correlation code.

3.2 Estimators

While the correlation function is a continuous function, the data we have is discrete. Hence, when computing the two point correlation from data, we have to use binning and an estimator in order to compute the correlation, and reduce errors resulting from factors such as survey geometry. In galaxy surveys, we thus commonly separate our data into distance separation bins [r, r + δr], and then count the number of pairs bin by bin.

In this project, we use the following estimator from Davis and Peebles (Davis & Peebles[1983]).

\[
\hat{\xi}(r) = \frac{N_{r} DD - N_{r} DR}{N_{d} DR} - 1
\]

where \(N_r\) and \(N_d\) are the number of randoms and stars respectively, \(DD\) represents the star-star pairs separated by distance \(r\), and \(DR\) represents the star-random pairs separated by distance \(r\).

3.3 Dual KD Trees

The pair counting problem that presents itself in the correlation problem can easily be solved a brute force approach by comparing each particle with every other, resulting in a algorithm with a complexity \(O(n^2)\). However, this naïve implementation scales poorly. Instead, a more efficient algorithm utilizes constructing a KD-Tree on the dataset to be able to more efficiently perform range queries and greatly reducing the amount of time required to perform the pair counting operation, making use of the fact that we are only interested in pair separated up to some cutoff distance. Further optimizations are also utilized, such as performing dual tree traversals and parallelizing operations over 16 cores. This allows us to run the algorithm in reasonable times at the scales we are interested in.

4 CORRELATION : RESULTS

The results for running the correlation function can be seen in Figures[1 and 2]. In Figure[1] we plot the value of the correlation function found for each distance bin against the value at the middle of the distance bins. We have 200 equally spaced bins over a distance range of 1 kpc. We observe a decreasing correlation function as we would expect, since we observe that stars cluster more on smaller scales. However, there is not much more information from Figure...
save for the changing gradient. Also, while it is difficult to tell without error bars, the function obtains seems to have surprisingly little noise.

In Figure 2 we plot the same function, but this time scaling the y axis by a factor $r^2$. Interestingly, we see to see a peak at 500 pc, which might be related to the feature found in Mao’s paper (Mao & et al. 2015).

5  CORRELATION : FURTHER WORK

Non-static situation: moving stars means there would be noise -¿ how do we quantify this noise? Is it important? What kind of clustering/substructure do we expect in a galaxy based on theory? Does the planar aspect matter? Non-isotropic nature Impact of different Estimators

As for things that could have been improved or followed up on, the following are a few. Firstly, we could have used a denser random sample. Typically, the random sample should be far denser than the data sample. However, in this case, we used a random sample of the same size to run the code faster. It should be noted that we tested with 10% of the sample with the randoms to see if there was any visible difference, and we did not find any.

Another thing we could have looked at was to investigate earlier results about excess clustering at low distances, which previous works mentioned. This could have been done by using log binning of the distances.

We also tried to get errors bars by using the jackknife method. However, there were issues with normalization, likely due to the non isotropic nature of the sample even at large scales (relative to sample size).

Next, we could have attempt to fit galaxy models instead of using a random sample, as done in Mao’s paper (Mao & et al. 2015), where they showed how using the correlation function could place constraints on model parameters via a Monte Carlo Markov Chain method.

Lastly, the code still takes a long time to run even on a data set of this size. In order to scale to the full Gaia dataset, there would be a need to develop massively parallel multi tree algorithms that perhaps utilize million scale cores to do the computation.

6  VELOCITY PROFILES : MOTIVATION

The rotation curve of a galaxy can give information on the structure and matter content of the galaxy. With Gaia’s higher precision and larger data set, more information is now available on the Milky Way, which can lead to new insights of our home galaxy, especially in terms of the mass distribution of the galaxy. If the mass of the galaxy is concentrated at the center of the galaxy, the velocity of stars as they orbit would peak just outside the concentration, and the magnitude of the velocity would then fall as the distance from the center increased. If there is uniform mass distribution throughout the galaxy, the velocity would increase with $r^2$ as the distance from the center increased. If, however, the mass distribution were more complicated, as is likely the case, the rotation curve would take on different properties, though it should remain smooth.

7  VELOCITY PROFILES : METHOD

To calculate the rotation curve for the Milky Way, it was necessary to determine the speeds of stars and their distances from the center of the galaxy. Restricting the stars we used to those that were cross-referenced with the Tycho-2 catalogue gave us a set of stars with known proper motions and parallaxes as well as celestial coordinates. We further restricted the dataset to stars with a small right ascension angle and made the assumption that most of a star’s velocity would be due to its orbital motion around the center of the galaxy. Stars with a small right ascension would then have most of their velocity in the plane perpendicular to our line-of-sight view, corresponding to the proper motion of the star, which is information available for the chosen set of stars.

The allowed maximum right ascension varied with distance from the sun. For a given star, the distance from the sun to the star was known, as was the distance from the sun to the center of the galaxy (8000 pc). The angle between the radial vector between the sun and the center of the galaxy and between the sun and the center of the star is found through trigonometry by

$$\theta_{\text{star}} = \cos^{-1}((\cos(ra)\cos(dec)))$$

where $\theta_{\text{star}}$ is the angle between the distance of the sun to the center of the galaxy and the distance between the sun and the star, $ra$ is the right ascension of the star, and $dec$ is the declination of the star. This defines a triangle, and the vector from the star to the center of the galaxy is the third side of this triangle, which, using the law of cosines is

$$R_{\text{star}} = \sqrt{D_{\text{star}}^2 + D_{\text{center}}^2 - 2D_{\text{star}}D_{\text{center}}\cos \theta_{\text{star}}}$$

where $R_{\text{star}}$ is the orbital radius of the star, $D_{\text{center}}$ is the distance from the sun to the star, $D_{\text{center}}$ is the distance from the sun to the center of the galaxy, and $\theta_{\text{star}}$ is the angle between the sun’s radius and the distance to the star as defined in Equation 2. The proper motion of the star is the velocity component perpendicular to the line-of-sight, and our assumption of circular orbital motion means that a star’s velocity is perpendicular to its radius, so as the angle between the line-of-sight and the star’s radius approaches 180°, the proper motion approaches the total velocity. Using these assumptions, we solved for the angle that would allow the proper motion to be near .9 of the total velocity and limited the stars used in the rotation curve to those that fell within the calculated right ascension. We then took the velocity to be the vector sum of the right ascension proper motion and the declination proper motion.

This, however, gave the velocity of the star with respect to the
(a) Rotation curve for stars with small right ascension. The line corresponds to the radius of the sun.

Figure 3.

There is likely calculation error when finding the velocity of the stars. There are distinct breaks in the points on the plot that correspond to the radius of the sun, indicating some discrepancy in the calculations for stars nearer to the center of the galaxy and stars that are farther away. These breaks, while indicative of an error, also make some tilt clear as velocity changes with radius. In order to mitigate the effect of the breaks, we plotted just the stars on the far side of the galaxy, determined by stars with a small right ascension and a distance from the sun of more than 8000pc, the distance from the sun to the center of the galaxy. This is shown in figure 4. This cut removed the breaks, though it is still difficult to discern a distinct pattern in the rotation curve. There is a hint of an increasing velocity as the radius increases, though it is not curved enough to correspond to the $r^2$ proportionality of uniform density.

We did not expect a clean curve due to the assumptions we made. The velocities will have variation due to movements not around the center of the Milky Way (i.e. the orbits of binary stars around each other), and the small right ascension angle restriction we imposed still included stars where the proper velocity could not account for all of the orbital velocity.

9 CONCLUSION

For the spatial correlation, we found that the two point correlation does seem to be a viable way of studying spatial structure in galaxies. However, getting interesting information from correlation function of stars probably requires fitting models to be able to study excess/under clustering with respect to such models and their parameters, but does seem very doable.

We did not find a strong dependence of velocity on radius for the stars orbiting in the galaxy, though Figure 3 indicates that there is some dependence. Figures 3 and 4 show some increase in velocity as radius increases, which is inconsistent with other calculations that generally find an increase and then a flattening of the curve (Sofue 2016).

Further work on this curve could include using calculations using all possible stars instead of just those with a small right ascension. Once a satisfactory rotation curve is achieved, it would be informative to find a function that matches the curve and use this to produce a mass profile of the Milky Way.

REFERENCES

Brown et al. 2016, AA
S. M. S., Lelli F., 2016, Physical Review Letters
Sofue Y., 2016, Publications of the Astronomical Society of Japan, 1