University of Groningen

Bachelor thesis

Astronomy & Physics

A statistical test for validation of Gaia data

Author:
Tomas Heldeweg

Supervisors:
Prof. Dr. Amina Helmi
Dr. Maarten Breddels
Dr. Jovan Veljanoski

25 July, 2016

Abstract

The objective of this report is to find strange correlations or trends in the 2D distributions of simulated TGAS data, so that we can later use these as a tool to validate the Gaia data. To this end, we created a program that finds symmetric areas in the galactic coordinates system with similar number of observations. By comparing the mutual information of parameters in the symmetric regions, as well as some antisymmetric areas, we were able to find several trends in the data that can be an helpful insight for validating the Gaia data. We found that the photometric parameters RP, BP and G magnitude, and the astrometric parameters Pmra, Pmdec and their respective errors had very different values for mutual information for both the symmetric and antisymmetric areas.

Contents

1 Introduction

The universe is extremely vast and full of complex structures. For millenia astronomers have been trying to get a better understanding of the cosmos. Many galaxies have been discovered and studied thoroughly. Our own galaxy, the Milky Way, is a great tool for learning more about other galaxies. Unfortunately, some parts of the Milky Way are challenging to investigate. It is literally like the old saying: You cannot see the forest for the trees; there are many stars and dust clouds that block the optical pathway, but with the increasingly advanced technology there is a lot that we can discover.

The Milky Way is the only galaxy where we are close enough, relatively speaking, to the stars so that we can determine the their motions, their chemical composition and their ages. The lower limit to the age of the Galaxy can be determined by finding the oldest star in the Galaxy. The older stars typically reside in the globular clusters of a galaxy. The oldest star known to date is 13.9 Gyr old [1] and thus this star sets the lower limit to the age of the Galaxy. The Milky Way is a barred spiral galaxy with a super massive black hole at the center. The mass of the Galaxy is $\sim 10^{11}$ [2] solar masses and the number of stars in the Galaxy has therefore often been approximated to be in the order of 10^{11}; however, the initial mass function suggests that there are more low mass stars than high mass stars, so the number of stars is probably higher. The diameter of the Galaxy's disk is ~ 30 kpc [3]. The properties of the Milky Way make it a very typical galaxy. So, not only is the Milky Way very interesting to study as we reside in it, but it is also a very useful galaxy to study. Being a very typical galaxy, the Milky Way can help us understand more about other galaxies in the universe.

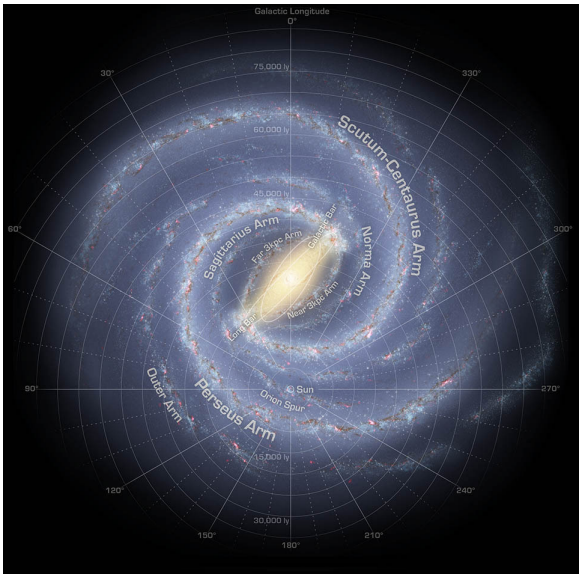

Figure 1: Artist impression of the Milky Way [4].

Although we might never be able to map the entire Galaxy, there is still a lot that can be deduced from only a fraction of the Galaxy. This is what the newest space mission from the European Space Agency (ESA) tries to accomplish. The Gaia space telescope was launched by ESA on 19 December 2013 with the primary mission to chart a three-dimensional map of the Milky Way. This is achieved by surveying around 1 billion of the brightest stars, stars with an apparent magnitude up to 20 mag, about 70 times over the course of 5 years [5]. The data obtained from the mission can then be used to address fundamental questions about the formation, structure and evolution of our galaxy. The telescope operates in a Lissajous orbit around the L_2 Lagrange Point, which has the advantage that the spacecraft will remain in the same relative position to the sun and earth so that it has a large field of view, unhindered by eclipses.

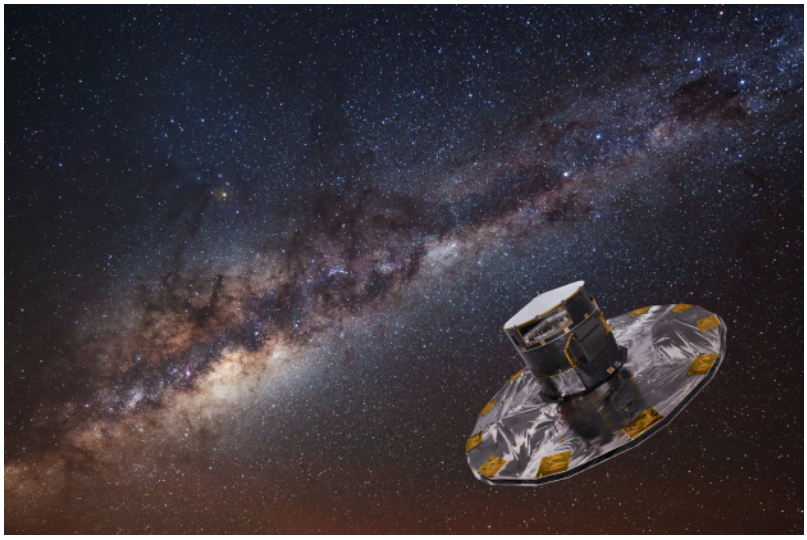

Figure 2: Artist impression of Gaia [5].

Gaia is not the first space mission that has had the primary goal of measuring the motions of stars in the Milky Way. Gaia's predecessor, Hipparcos, was launched by ESA in 1989 and its data is still used to this day. Gaia will significantly improve the quality of the observations compared to Hipparcos. This is because of three main components [6]: Firstly, the larger optics on the telescope have a larger collecting areas and a smaller diffraction pattern. Then, the improved quantum efficiency and bandwidth of the detector result in better photon statistics. Finally, the use of a CCDs provides a multiplexing advantage. Multiplexing is a way of sending multiple signals or streams of information over a communications link at the same time in the form of a single signal [6]. For those reasons, Gaia will be able to improve the accuracy of the measurements by a factor of 100, the limiting magnitude (faintest stars detectable) by a factor of 1000, and number of stars observed by a factor of 10000 with respect to Hipparcos [6].

Beside the primary objective of Gaia, the mapping of the Milky Way, there are many other scientific goals, which broadly include: The origin and history of our Galaxy, the formation and evolution of stars in the Milky Way, distance scale and reference frame, Local Group and beyond, Solar System, Extra-solar planetary systems, fundamental physics and specific objects. The census of the Milky Way will help answer some of the fundamental questions in each of these topics. Hence, the Gaia data will be used to study the kinematical, dynamical and chemical structure, and the evolution of the galaxy. The Gaia data will be useful in many fields of science like stellar physics, solar-system bodies, fundamental physics, and exo-planets [7].

Before the Gaia data is released publicly it has to be validated. There are many tests that need to be done to fully validate the Gaia data. In this work a statistical test will be developed involving the 2D distribution of several parameters of the Tycho-Gaia Astrometric Solution (hereafter TGAS) data. This will be explained in more detail in the Methodology section.
The real data cannot be used widely (i.e. beyond members of the Gaia Validation Working Group), so therefore the algorithms and calculations in this report are done on simulations. The algorithms and calculations are done in such a manner that they can also be applied on the real data.
This report is organized as follows, in the section Methodology we will first explain how the simulations are obtained and describe the dataset of the simulations. After that there is a section describing the statistics that will be used. Then we will describe the software we developed and how to apply it. In the section Results we will discuss the plots and calculations that were made and in the section Conclusion we will summarize the results that were found and conclude whether the tests can be done on the actual Gaia data.

2 Methodology

2.1 The Gaia data

The Gaia dataset is tabular, comprising absolute astrometry, broad-band photometry, spectrometry and low-resolution spectro-photometry. The data is stored in columns with each column representing a unique parameter/subspace and each row a star [7]. There are 5 astrometric parameters: two angles for star position, the two proper motions of the stars (time derivative of the positions), and the parallax of the stars, which can be used to calculate the distance to the stars. Each of these astrometric parameters also have an error parameter associated with them. The photometric parameters consist of apparent magnitudes and fluxes in 3 bands: The G-band magnitude, the RP-band magnitude and the BP-band magnitude. The G-band is the visual spectrum, which runs from 330-1050 nm. The RP-band (Red Photometer) is the bandpass for long wavelengths, which runs from 640 - 1000 nm . The BP-band (Blue Photometer) is the band for short wavelength, running from 330 - 680 nm [8]. The parameters derived with intermediate resolution spectro-photometry are temperature, surface gravity and metallicities. The radial velocities are obtained for $\sim 1.5 * 10^8$ stars with spectroscopy.

2.2 The TGAS data

The Tycho-Gaia Astrometric Solution (TGAS) dataset is also tabular and is very similar to the Gaia data. It contains the astrometric parameters positions, proper motions and parallax and their respective errors, and also the photometric parameters G-band magnitude, RP-band magnitude and BP-band magnitude, although the BP and RP band magnitudes will not be released as part of TGAS. These parameters will be the focal point for the statistical work that will be done on the data. Furthermore, the data set contains a correlation matrix between astrometric parameters for each star.

2.3 The simulated data

2.3.1 The Gaia simulator

The simulations are an integral part of the validation of the data. They therefore have to be of high quality. The Gaia simulator is a collection of three data generators: The Gaia Instrument and Basic Image Simulator (GIBIS) [9], the Gaia System Simulator [10] (GASS), and the Gaia Object Generator (GOG) [11]. The Gaia simulator uses the Universe Model (UM) as a base. The UM creates object catalogues down to a limiting magnitude, which is G = 20 for the Gaia Simulator. The model simulates stellar content of the galaxy based on the Besançon galaxy model [12]. The objects it produces belong the the main four stellar populations: thin disk, thick disk, halo, and bulge. The created objects can be stars, nebulae, clusters, diffuse light, planets, satellites, asteroids comets, resolved galaxies, unresolved extended galaxies, quasars, AGN, and supernovae [11]. The UM can create a set of these objects on a section or on the whole sky. Each object will have a position and a set of observational properties [13]. The three data generators all have their own task in the Gaia Simulator: GASS simulates a huge amount of realistic telemetry stream, GIBIS produces pixel level images, and GOG generates the final data mission [10][9][11].

2.3.2 TGAS simulations

The statistical tests done in this work will be tested on the TGAS simulations. The TGAS simulations were made using AGISLab, which is a software package that was created at Lund Observatory to develop and test Gaia astrometric data processing strategies [14]. To get high accuracy one needs to have observations that have been collected over an extended period of time. Since it has only been 2 years since the launch of the Gaia telescope, the data that has been collected so far will, on its own, not be accurate enough to reliably resolve all the astrometric parameters [14]. The TGAS simulation does achieve a high accuracy for the astrometric parameters by combining data from the Hipparcos and Tycho-2 catalogues with the Gaia data. The Hipparcos data and the Gaia data have a $\sim$ 24 year time difference in the data. Using this

difference in time between the same stars one can more accurately find the astrometric parameters. A big limitation of only combining the Gaia and Hipparcos data was that the Hipparcos data contains far fewer stars than the Gaia data. At first this was solved by adding "auxiliary stars", but these could potentially bias the solution for the parameters. This problem was overcome by including the Tycho-2 data to replace the auxiliary stars, using their positions at the time of the Hipparcos measurements to constrain the proper motion. This ensures that the solution is not biased because of the auxiliary stars and allowed accurate solution for the astrometric parameters.

2.4 Kullback-Leibler Divergence

The Kullback-Leibler Divergence (KLD) is a measure of the difference between two probability distribution functions. For continuous probability distributions the formula for the KLD is given by:

$$D_{KL}(P:Q) = \int P(x) \ln \frac{P(x)}{Q(x)} dx \tag{1}$$

where P and Q are the continuous distribution functions. In this form P often represents the real data and Q the model. The KLD is always positive and is not symmetric in P and Q. The units of the KLD are related to the base of the logarithm in the equation. Hence they are called nats here.

The mutual information (hereafter MI) or degree of clustering of a multidimensional probability distribution $P(x, y)$ can be determined using the KLD by comparing $P(x, y)$ to the product of its marginal distributions $P(x)P(y)$. In equation form it would look like this:

$$D_{KL}(P(x,y) : P(x)P(y)) = \int_y \int_x P(x,y) \ln \frac{P(x,y)}{P(x)P(y)} dxdy \tag{2}$$

As the name suggests, the MI can be described as the information that 2 parameters/variables (x and y in this case) share. If the MI is 0 then the parameters are independent from each other. This can be seen from the formula: when two parameters, x and y are independent, then $P(x, y) = P(x)P(y)$ resulting in the right hand side of the equation being 0. Contrarily, having a high MI gives a lot of information about y when knowing x [15]. The MI is not only dependent on the shape of the 1D distribution, as is demonstrated by fig. 3.

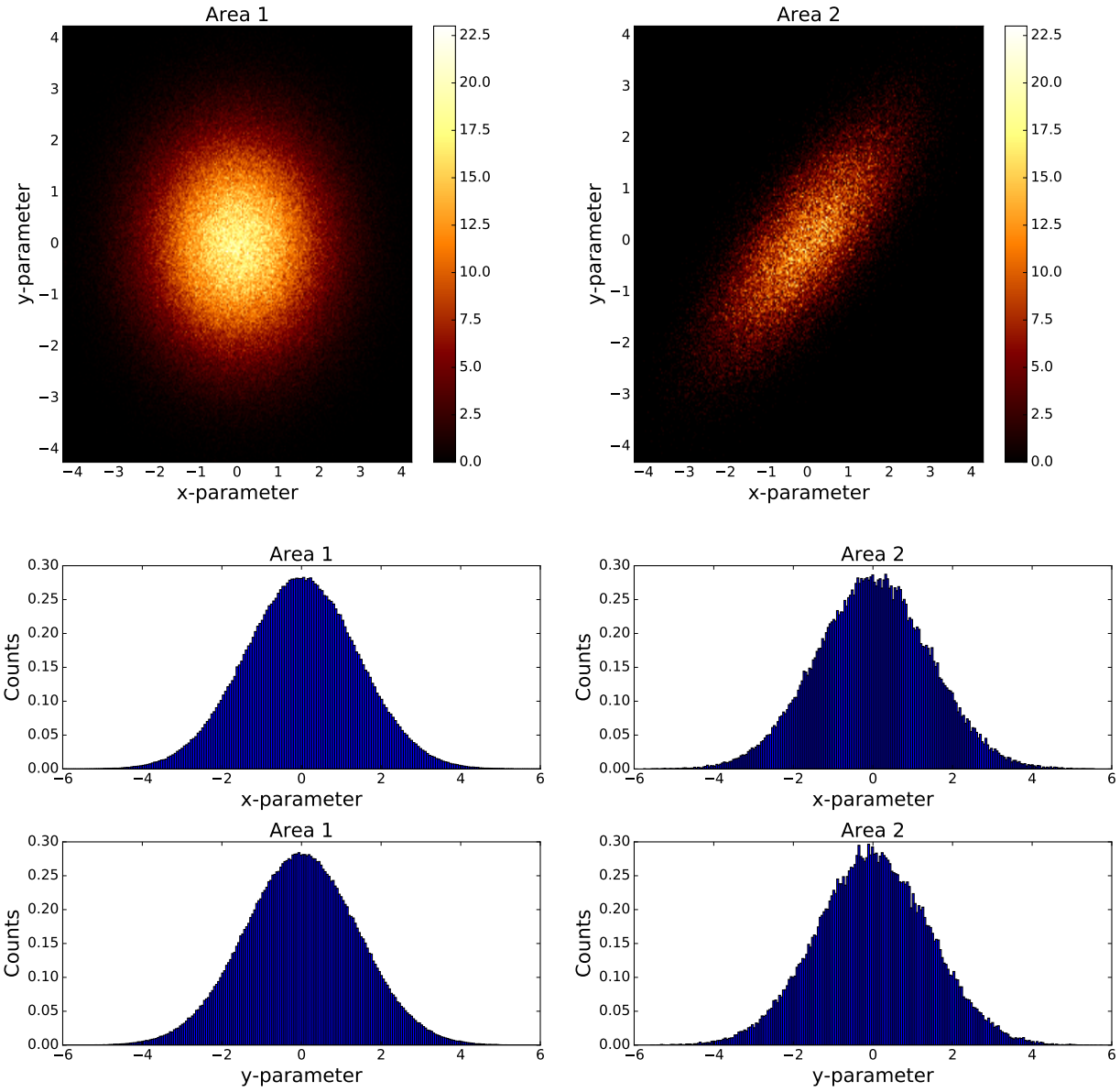

Figure 3: Top: 2D histograms and bottom: 1D histograms of areas with different mutual information.

The top two images of fig. 3 are the 2D histograms of the some parameters x and y and the bottom two are their respective 1D histograms. Although the shape of the 1D histograms is the same their degree of clustering is different; area 2 has a higher MI, which means that the parameters of area 2 are more dependent on each other than those of area 1. The MI is calculated for all combinations of the relevant subspaces.

2.5 Handling the data

To handle the large amount of data VaeX [16] is used. VaeX is a graphical tool and a python library developed by M. Breddels that is used to visualise and handle large tabular datasets. VaeX can visualise large amounts of data ($\sim 10^8$, $\sim 10^9$ rows) within seconds and that is exactly what is needed for the Gaia

data. VaeX can also do conversions to galactic and ecliptic coordinates. Furthermore, VaeX allows you to select certain regions on the sky using boolean statements and calculate the mutual information of the subspaces in the selected area.

2.6 Selecting regions

So far the validation effort in Groningen has been on global comparison between data and simulations, (such as those described in section 2.3). To validate further it is useful to compare regions on the sky between data and data. This way any systematic errors in the data can be detected that might have been missed by the global comparison.

Regions are compared on basis of similar number of observations and symmetry with respect to the galactic disk. The reason for this is that the number of observations is directly related to the error in measurement and areas having a similar error is desirable. The symmetry with respect to the disk is chosen because one would expect those areas to have (mostly) similar properties.

The regions that are compared are chosen to be circular as that is the most symmetric shape. We wrote a program to find symmetric areas that are circular and have similar number of observations. When we find the positions and radii of these areas, we can select regions on the sky and calculate the mutual information of subspaces in the regions using VaeX. By comparing the mutual informations of the region we can find correlations and trends. First we use VaeX to make a 2D histogram with the values of mean number of observations in galactic coordinates, which looks like this if plotted:

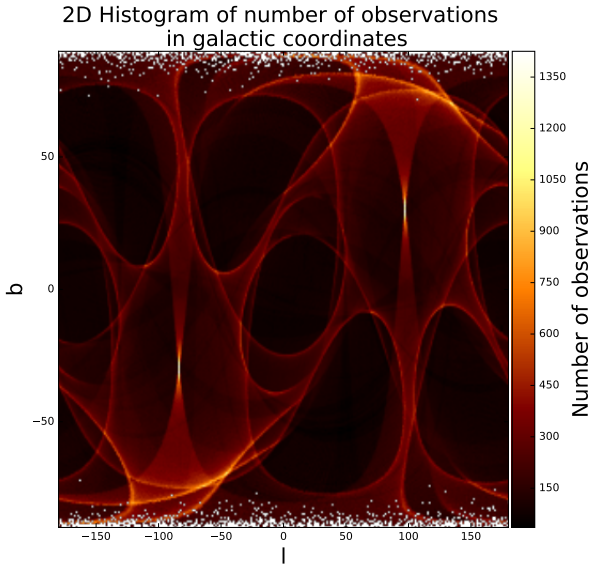

Figure 4: 2D histogram of number of observations in galactic coordinates.

The 2D histogram is a 256 by 256 grid with each gridpoint containing the number of observations on that element. To accurately display the data the y-axis will be transformed from b to $\sin b$ using VaeX. This is because the equation for a surface element on the unit sphere given by:

$$dA = \sin b db dl \tag{3}$$

where b is the latitude and l the longitude. The resulting image is shown in fig. 5.

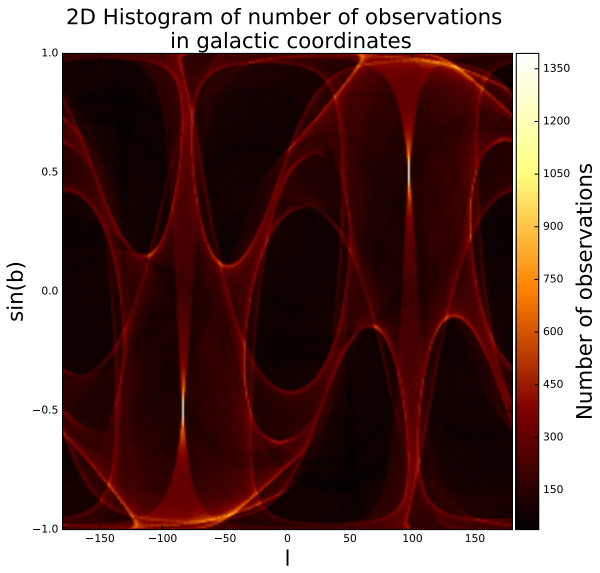

Figure 5: 2D histogram of number of observations after the transformation.

To select regions with similar NObs we have to set a range of NObs to look in. After setting this range we create a mask. The mask is a grid (with the same dimensions as the 2D histogram) that has boolean values, i.e. true and false, on each grid point. true is assigned to the gridpoints that are within the range and false where this is not the case. From now on we will refer to this mask as the NObs mask. If we take, for example, the NObs range to be between 50 and 110, the following NObs mask is created:

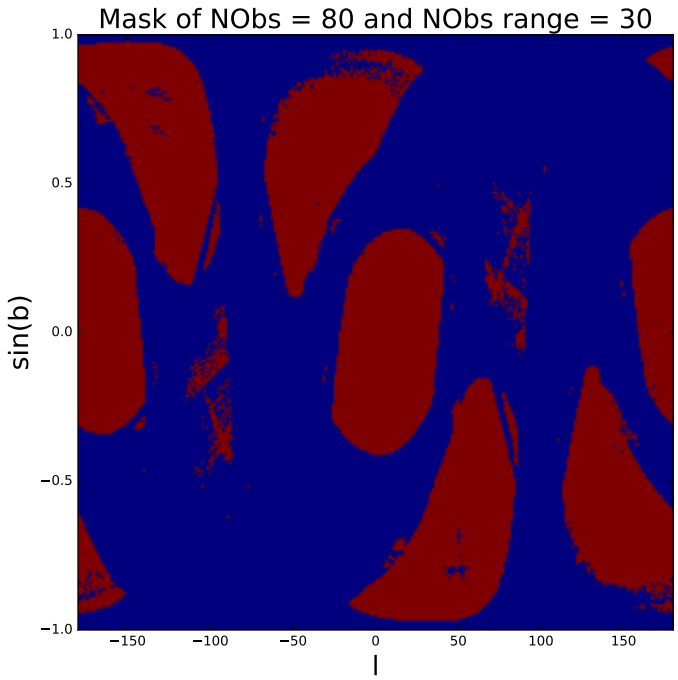

Figure 6: Mask where NObs are in the range of 50 and 110.

In fig. 6 the red areas are gridpoints where the NObs mask has the value true. Now we try to find the largest circular areas in this NObs mask that only contains trues. We do this by looking for the largest circle it can find for each gridpoint. Each gridpoint has a certain i and j index with 0 representing the first gridpoint. These indices represent respectively the x and y location in the grid. When the program looks at a gridpoint it uses the values for i and j to create a new grid. This grid has the special property that it contains values that represent the radial distance squared from the selected gridpoint. To illustrate, if we have a 9 by 9 grid and we selected the grid point $i = 4$ and $j = 4$ it will create the grid shown in fig. 7.

$$
\begin{bmatrix}
[32 & 25 & 20 & 17 & 16 & 17 & 20 & 25 & 32] \\
[25 & 18 & 13 & 10 & 9 & 10 & 13 & 18 & 25] \\
[20 & 13 & 8 & 5 & 4 & 5 & 8 & 13 & 20] \\
[17 & 10 & 5 & 2 & 1 & 2 & 5 & 10 & 17] \\
[16 & 9 & 4 & 1 & 0 & 1 & 4 & 9 & 16] \\
[17 & 10 & 5 & 2 & 1 & 2 & 5 & 10 & 17] \\
[20 & 13 & 8 & 5 & 4 & 5 & 8 & 13 & 20] \\
[25 & 18 & 13 & 10 & 9 & 10 & 13 & 18 & 25] \\
[32 & 25 & 20 & 17 & 16 & 17 & 20 & 25 & 32]]
\end{bmatrix}
$$

Figure 7: Grid of radial distances from $i = 4$, $j = 4$.

Where the radial distance squared from the center point is determined by:

$$\text{radius}^2_{\text{grid}} = \Delta i^2 + \Delta j^2 \tag{4}$$

Where Δi^2 and Δj^2 are respectively the i and j distances from the selected gridpoint. The grid can be used to approximate a circular area around the selected gridpoint. This is done by applying a boolean statement that only selects the gridpoints within a certain radius. After applying this boolean statement we get yet another mask, where true values are assigned to gridpoints that are in this circular area. This mask from now on will be referred to as the circle mask. To illustrate, if we take the radius to be 2 or less for the grid in fig. 7 then we will get the following circle mask:

$$\begin{bmatrix}
[0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0] \\
[0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0] \\
[0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0] \\
[0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0] \\
[0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0] \\
[0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0] \\
[0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0] \\
[0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0] \\
[0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0]]
\end{bmatrix}$$

Figure 8: Circle mask with radius ≤ 2 of grid from fig. 7.

As the radius gets higher, the shape will look more and more like a circle. A circle mask with a certain radius will then be applied to the NObs mask. The resulting values are the values of the NObs mask that are within the circle mask. If all the values are true then we know that all values of NObs in the circle are indeed between the range of NObs that was set at the beginning. In summary, we first select a certain range of NObs. Then we compute the accompanying NObs mask. After that we loop over every gridpoint in the NObs mask. At each gridpoint a circle mask is created with radius ≤ 1. If all values returned from applying the circle mask to the NObs mask are true then the radius is increased by 1 and the process repeated. This is done until a false value is encountered. When a false value is encountered the last radius is stored together with i and j coordinate of the gridpoint. This way we can find the largest circular area for each gridpoint for which the NObs of every element is in the set NObs range. We repeat this entire process for several ranges of NObs. We then search for areas that have a symmetric counterpart in l, $\sin(b)$ and NObs. First we convert the data to positions and radii on the sky. The axes of $\sin(b)$ are not the same in length we compute 2 radii; one for each direction. The radii can thereafter be used by VaeX to select the regions on the sky. Before using the data it has to be converted from grid points to galactic coordinates. The grid is 256 by 256 so the range of i and j are $0 - 255$. The radii and position for each direction are therefore converted using the following equations:

$$\text{radius}_l = \frac{\text{radius}_{\text{grid}} \cdot (l_{\max} - l_{\min})}{255} \tag{5}$$

$$\text{radius}_{\sin b} = \frac{\text{radius}_{\text{grid}} \cdot (\sin b_{max} - \sin b_{min})}{255} \tag{6}$$

$$l = \frac{i \cdot (l_{\max} - l_{\min})}{255} - l_{\max} \tag{7}$$

$$\sin b = \frac{j \cdot (\sin b_{\max} - \sin b_{\min})}{255} - \sin b_{\max} \tag{8}$$

After the conversion we have a data file containing the positions and radii for each gridpoint. Then we find all the areas that have a symmetric counterpart for $\sin b$ and l. The values for radii and coordinates of the symmetric areas are then used in VaeX to select the regions in the sky using the following statement:

$$\left(\frac{l - l_{\text{coordinate}}}{l_{\max} - l_{\min}}\right)^2 + \left(\frac{b - b_{\text{coordinate}}}{b_{\max} - b_{\min}}\right)^2 \leq \left(\frac{255}{\text{radius}_{\text{grid}}}\right)^2 \tag{9}$$

This formula normalises the axes so a we can select a circle in the coordinate system. After the selection we calculate the mutual information of the area for the combinations of subspaces mentioned in 2.3 and do the same for its symmetric counterpart. We plot the mutual information of the areas against each other to obtain MI plots. These plots show how much each combination of subspaces deviates from the 1:1 line. If the subspaces cluster around the 1:1 line, then the distributions of the compared regions are very similar. Therefore, deviations from the 1:1 line point to a difference in distributions. We will also compute some MI plots for antisymmetric areas and see if they will deviate more.

3 Results

We plot all the symmetric regions that were found in fig. 9:

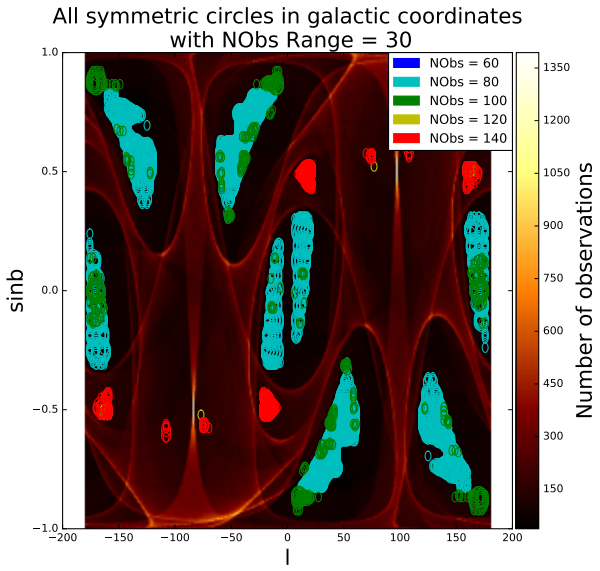

Figure 9: Symmetric Areas with Nobs range = 30.

The NObs of the circles are chosen to be around 80, since this is the median of the NObs of the entire sky as one can see from fig. 10.

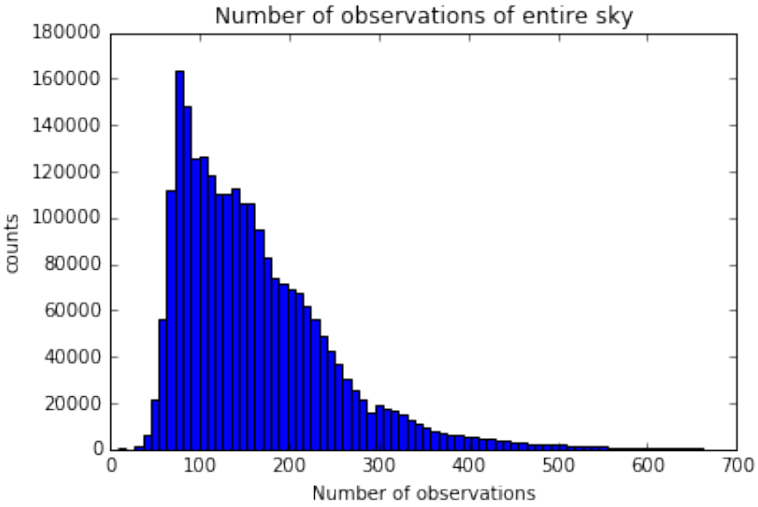

Figure 10: Histogram of the number of observation of entire sky.

First we computed a few MI plots of symmetric areas to see what their general form is.

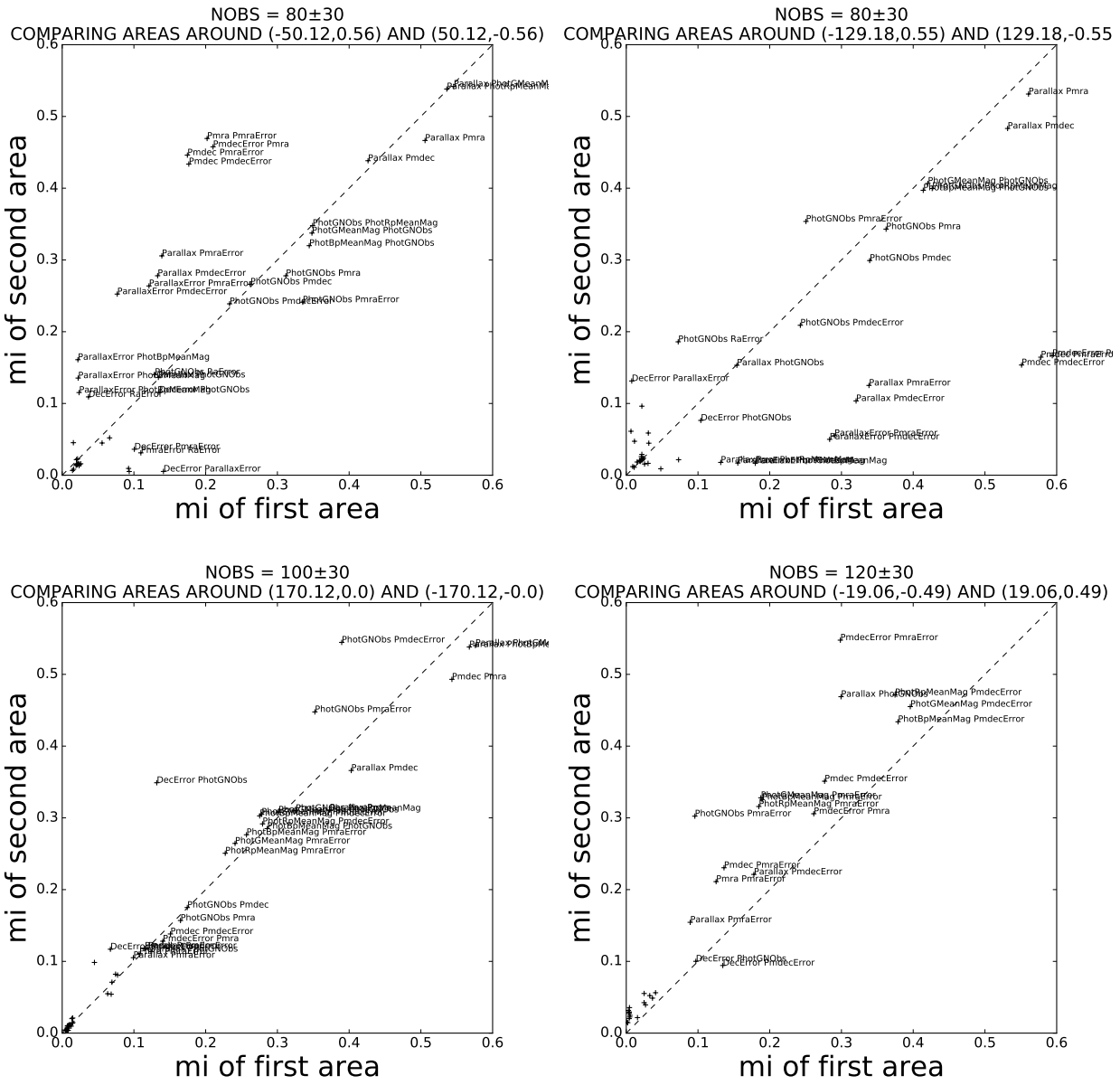

Figure 11: Examples of MI plots for symmetric areas.

As one can see from the plot a lot of subspaces cluster in the bottom left corner. For these subspaces the mutual information is very low, which means that the subspaces are mostly independent of each other. These combination of subspaces are therefore not very interesting; they do not contain a lot of information. The points we are interested in are the more deviating ones. We also computed some MI plots for antisymmetric areas to see if they looked different.

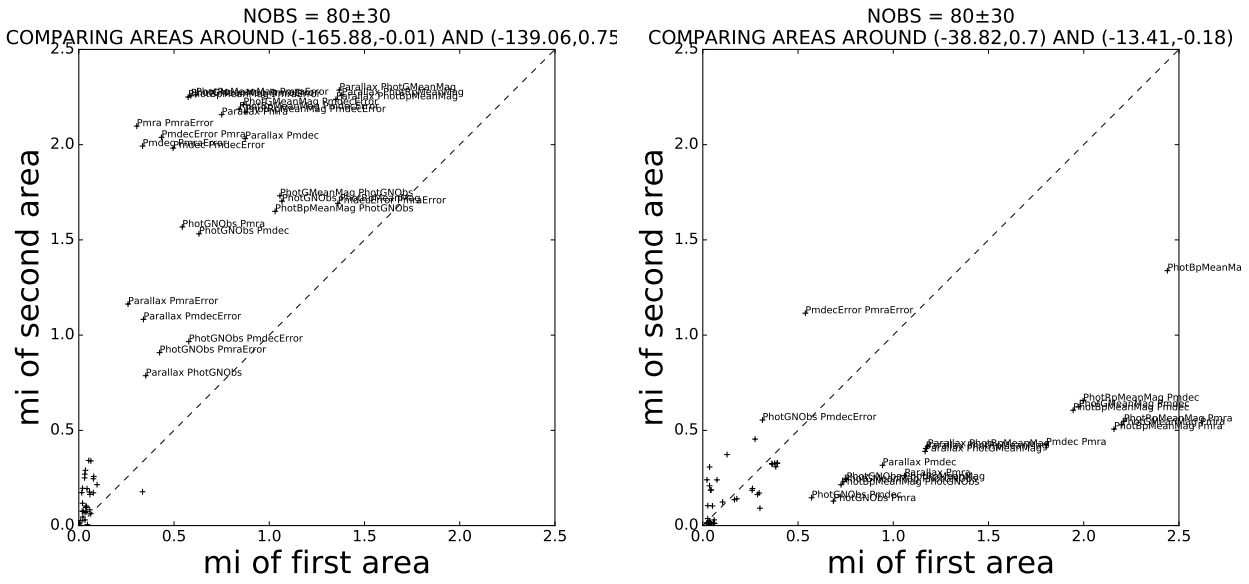

Figure 12: Examples of MI plots for antisymmetric areas.

From the plots one can see that the antisymmetric areas deviate more from the 1:1 than the symmetric areas (mind the fact that the range of the axes are different). This is what is expected for regions that are not symmetric to the disk; different areas in the sky will have different properties and thus different distributions for quantities

After creating the MI plots we calculated the average deviation, median deviation and normalised average of a large sample of symmetric and anti symmetric areas.

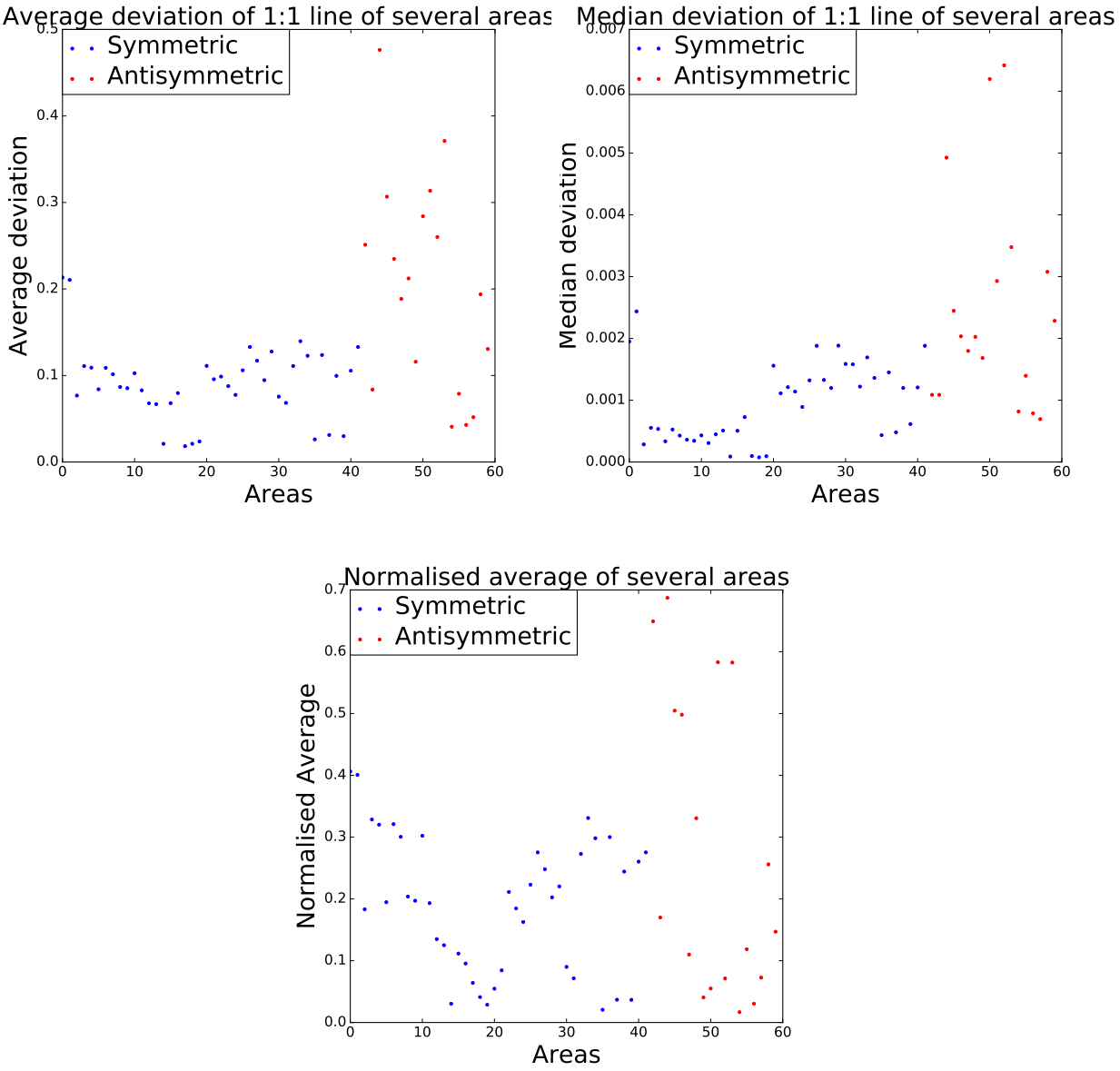

Figure 13: Top left: Average deviation, top right: Median deviation, bottom: Normalised average.

As one can see from fig. 13 the antisymmetric areas tend to have a higher average deviation, which confirms our prior belief. A key of the coordinates, size and number of stars for each area can be found in the appendix.

After computing some MI plots and the averages we looked at which subspaces drove the deviation from the one to one line. We did this by calculating how many times each subspace occurred in the top 11 subspaces. The top 11 was chosen as there are in total 12 subspaces, so each subspace makes 11 combinations with the other subspaces. This way we have some sort of normalisation in the occurrence. This information is given in the following tables:

subspaces	percentual occurence in top 11 deviating subspaces
PhotRpMeanMag	29.44
PmdecError	25.54
Pmra	24.68
PhotBpMeanMag	23.59
PmraError	22.73
PhotGMeanMag	22.08
Pmdec	18.61
Parallax	16.67
PhotGNObs	11.69
DecError	3.68
ParallaxError	1.3
RaError	0.0

Table 1: Occurence in top 11 deviating subspaces of symmetric areas.

subspaces	percentual occurence in top 11 deviating subspaces
Pmra	35.86
PhotBpMeanMag	24.24
Pmdec	23.74
PmraError	23.23
PhotRpMeanMag	22.73
PhotGMeanMag	22.22
PmdecError	22.22
Parallax	12.63
PhotGNObs	12.12
DecError	0.51
RaError	0.51
ParallaxError	0.0

Table 2: Occurence in top 11 deviating subspaces of antisymmetric areas.

The objective of this report is to characterise the 2D distributions of simulated Tycho-Gaia Astrometric Solution (TGAS) data, so that we can later use these as a tool to validate the Gaia data itself. To this end, we developed a program that finds symmetric areas in the galactic coordinates system with similar number ofobservations. By comparing the mutual information of parameters in the symmetric regions, as well as some antisymmetric areas, we were able to identify several trends in the data that can be helpful for validating the Gaia data. We found that the photometric parameters RP, BP and G magnitude, and the astrometric parameters Pmra, Pmdec and their respective errors had very different values for mutual information for both the symmetric and antisymmetric areas.

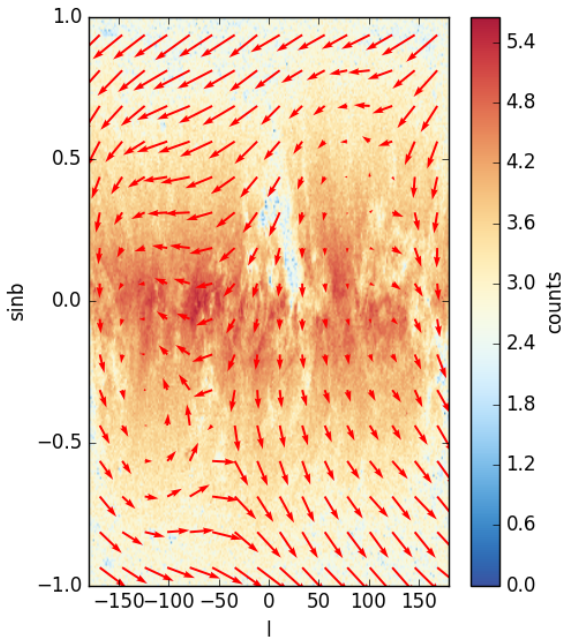

Figure 14: Vector field of Pmdec and Pmra.

From fig. 14 we can see that symmetric areas have somewhat similar directions of Pmra and Pmdec. Contrarily when comparing two antisymmetric regions on the sky the chance is a lot higher that the vectors are pointed in widely varying directions. This might be the reason that Pmdec and Pmra have a higher occurence rate in the antisymmetric areas.

The PmraError and PmdecError subspaces also have a high percentual occurrence in the top 11 deviating subspace. Further investigating some of the symmetric areas did indeed show that the 2D distribution of these subspaces were very different. In fig. 15 we plotted the histograms of PmraError and PmdecError of a symmetric area that had highly deviating proper motion error subspaces:

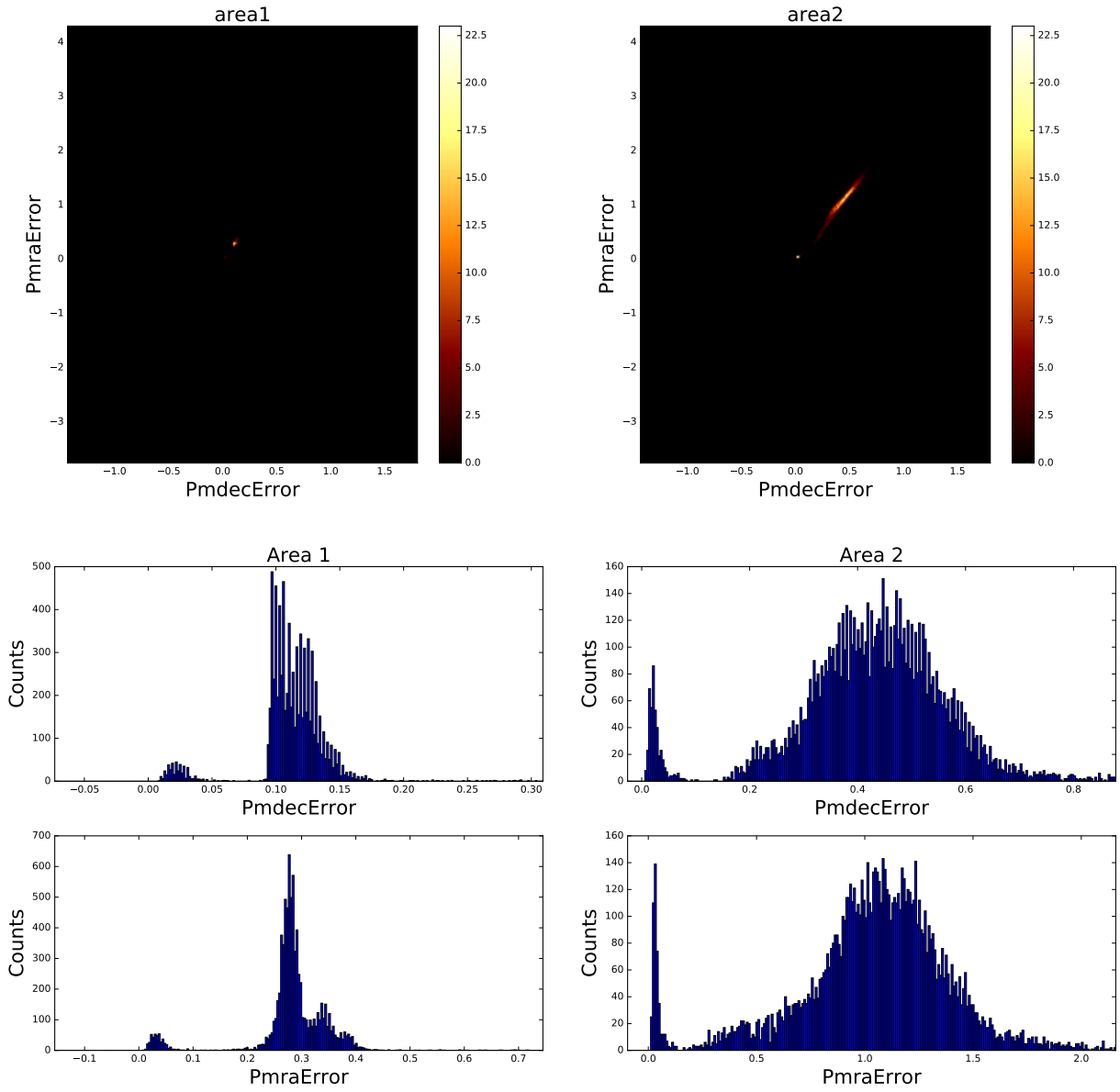

Figure 15: Histograms of an symmetric area with highly deviating error subspaces.

As one can see from fig. 15 the 2D histograms are very different and as a result their MI is very different. Subsequently, any combination with these will likely deviate a lot from the 1:1 line.

4 Conclusion

By applying the program to the simulations we have found several interesting facts. Antisymmetric areas generally have a higher deviation from the 1 to 1 line. This is what was expected because that are not symmetric with respect to the disk will generally have different properties and thus different distributions for the observables.

Furthermore, there are trends that can be found by looking at what drives the deviation: the top deviating subspaces are the proper motions and the photometric magnitude subspaces for both symmetric and antisymmetric areas. Lastly, the PmraError and PmdecError subspaces are highly deviating.

The program that was written has been successfully applied to the simulated data. The trends that were found will be a useful insight when the program is applied to the actual Gaia data. If the subspaces that did not occur often in the top 11 most deviating subspaces deviate a lot in the real data, than we know that there is some mechanic in the data that has to be understood more in depth.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Amina Helmi for the useful comments, remarks and engagement through the learning process of this bachelor thesis. Her expertise and keen eye have been of great importance to this report. Furthermore, I would like to thank Jovan Veljanosky and Maarten Breddels for their comments and readiness throughout the course of this project. They have been very helpful and were always very kind and quick to respond whenever I needed feedback.

References

[1] H. E. Bond, E. P. Nelan, D. A. VandenBerg, G. H. Schaefer, and D. Harmer. HD 140283: A Star in the Solar Neighborhood that Formed Shortly after the Big Bang. *The Astrophysical Journal, Letters*, 765:L12, March 2013.

[2] O. Gerhard. Mass distribution in our Galaxy. *Space Science Reviews*, 100:129–138, January 2002.

[3] Nasa - the cosmic distance scale. `http://imagine.gsfc.nasa.gov/features/cosmic/milkyway_info.html`. Accessed: 2016-07-17.

[4] Spitzer. `http://www.spitzer.caltech.edu/images/1925-ssc2008-10b-A-Roadmap-to-the-Milky-Way-Annotated-` Accessed: 2016-07-20.

[5] European Space Agency gaia mission. `http://www.esa.int/Our_Activities/Space_Science/Gaia/The_legend_of_Gaia`. Accessed: 2016-06-30.

[6] M. A. C. Perryman, K. S. de Boer, G. Gilmore, E. Høg, M. G. Lattanzi, L. Lindegren, X. Luri, F. Mignard, O. Pace, and P. T. de Zeeuw. GAIA: Composition, formation and evolution of the Galaxy. *Astronomy and Astrophysics*, 369:339–363, April 2001.

[7] J. H. J. de Bruijne. Science performance of Gaia, ESA's space-astrometry mission. *Astrophysics and Space Science*, 341:31–41, September 2012.

[8] C. Jordi, M. Gebran, J. M. Carrasco, J. de Bruijne, H. Voss, C. Fabricius, J. Knude, A. Vallenari, R. Kohley, and A. Mora. Gaia broad band photometry. *Astronomy and Astrophysics*, 523:A48, November 2010.

[9] C. Babusiaux. The Gaia Instrument and Basic Image Simulator. In C. Turon, K. S. O'Flaherty, and M. A. C. Perryman, editors, *The Three-Dimensional Universe with Gaia*, volume 576 of *ESA Special Publication*, page 417, January 2005.

[10] E. Masana, Y. Isasi, X. Luri, and J. Peralta. The Gaia Simulator: Design and Results. *Astrophysics and Space Science Proceedings*, 14:515, 2010.

[11] X. Luri, M. Palmer, F. Arenou, E. Masana, J. de Bruijne, E. Antiche, C. Babusiaux, R. Borrachero, P. Sartoretti, F. Julbe, Y. Isasi, O. Martinez, A. C. Robin, C. Reylé, C. Jordi, and J. M. Carrasco. Overview and stellar statistics of the expected Gaia Catalogue using the Gaia Object Generator. *Astronomy and Astrophysics*, 566:A119, June 2014.

[12] A. C. Robin, C. Reylé, S. Derrière, and S. Picaud. A synthetic view on structure and evolution of the Milky Way. *Astronomy and Astrophysics*, 409:523–540, October 2003.

[13] A. C. Robin, X. Luri, C. Reylé, Y. Isasi, E. Grux, S. Blanco-Cuaresma, F. Arenou, C. Babusiaux, M. Belcheva, R. Drimmel, C. Jordi, A. Krone-Martins, E. Masana, J. C. Mauduit, F. Mignard, N. Mowlavi, B. Rocca-Volmerange, P. Sartoretti, E. Slezak, and A. Sozzetti. Gaia Universe model snapshot. A statistical analysis of the expected contents of the Gaia catalogue. *Astronomy and Astrophysics*, 543:A100, July 2012.

[14] D. Michalik, L. Lindegren, and D. Hobbs. The Tycho-Gaia astrometric solution . How to get 2.5 million parallaxes with less than one year of Gaia data. *Astronomy and Astrophysics*, 574:A115, February 2015.

[15] R. E. Sanderson, A. Helmi, and D. W. Hogg. Action-space Clustering of Tidal Streams to Infer the Galactic Potential. *The Astrophysical Journal*, 801:98, March 2015.

[16] VaeX visualization and exploration. `https://www.astro.rug.nl/~breddels/vaex/`, note = Accessed: 2016-06-30.

Appendix

Key of areas

Area	NObs	Coordinates:(Area1),(Area2)	l-radius	sinb-radius	Nr of stars:(Area1, Area2)
0	60	(7.76,0.13),(-7.76,-0.13)	3.6250	0.0201	(744.0, 2376.0)
1	60	(-7.76,-0.18),(7.76,0.18)	3.6250	0.0201	(2286.0, 867.0)
2	80	(-50.12,0.56),(50.12,-0.56)	14.8750	0.0826	(9053.0, 9750.0)
3	80	(129.18,-0.57),(-129.18,0.57)	13.4688	0.0748	(7437.0, 6397.0)
4	80	(-129.18,0.55),(129.18,-0.55)	13.4688	0.0748	(6677.0, 7959.0)
5	80	(-50.12,0.57),(50.12,-0.57)	13.4688	0.0748	(7221.0, 7820.0)
6	80	(-129.18,0.56),(129.18,-0.56)	13.4688	0.0748	(6488.0, 7541.0)
7	80	(132.0,-0.56),(-132.0,0.56)	13.4688	0.0748	(7539.0, 6526.0)
8	80	(51.53,-0.56),(-51.53,0.56)	13.4688	0.0748	(7877.0, 7387.0)
9	80	(51.53,-0.55),(-51.53,0.55)	13.4688	0.0748	(8075.0, 7586.0)
10	80	(132.0,-0.57),(-132.0,0.57)	13.4688	0.0748	(7371.0, 6416.0)
11	80	(-48.71,0.58),(48.71,-0.58)	13.4688	0.0748	(7139.0, 7715.0)
12	100	(170.12,-0.89),(-170.12,0.89)	9.2500	0.0514	(2266.0, 2127.0)
13	100	(170.12,-0.88),(-170.12,0.88)	9.2500	0.0514	(2254.0, 2149.0)
14	100	(170.12,0.0),(-170.12,-0.0)	9.2500	0.0514	(9017.0, 9602.0)
15	100	(-170.12,0.87),(170.12,-0.87)	9.2500	0.0514	(2145.0, 2241.0)
16	100	(170.12,-0.87),(-170.12,0.87)	7.8438	0.0436	(1593.0, 1587.0)
17	100	(-170.12,-0.01),(170.12,0.01)	7.8438	0.0436	(6826.0, 6872.0)
18	100	(-170.12,0.0),(170.12,-0.0)	7.8438	0.0436	(7106.0, 6807.0)
19	100	(-170.12,0.01),(170.12,-0.01)	7.8438	0.0436	(7122.0, 6686.0)
20	100	(10.59,-0.87),(-10.59,0.87)	7.8438	0.0436	(1920.0, 1479.0)
21	100	(170.12,-0.86),(-170.12,0.86)	6.4375	0.0358	(1049.0, 1099.0)
22	120	(-19.06,-0.49),(19.06,0.49)	6.4375	0.0358	(2609.0, 1793.0)
23	120	(19.06,0.51),(-19.06,-0.51)	6.4375	0.0358	(1789.0, 2548.0)
24	120	(-19.06,-0.51),(19.06,0.51)	6.4375	0.0358	(2502.0, 1795.0)
25	120	(-17.65,-0.51),(17.65,0.51)	6.4375	0.0358	(2641.0, 1729.0)
26	120	(-16.24,-0.51),(16.24,0.51)	6.4375	0.0358	(2700.0, 1593.0)
27	120	(-17.65,-0.5),(17.65,0.5)	6.4375	0.0358	(2675.0, 1729.0)
28	120	(19.06,0.5),(-19.06,-0.5)	6.4375	0.0358	(1819.0, 2606.0)
29	120	(-16.24,-0.51),(16.24,0.51)	5.0312	0.0280	(1601.0, 970.0)
30	120	(20.47,0.5),(-20.47,-0.5)	5.0312	0.0280	(1191.0, 1498.0)
31	120	(20.47,0.51),(-20.47,-0.51)	5.0312	0.0280	(1205.0, 1484.0)
32	140	(17.65,0.49),(-17.65,-0.49)	7.8438	0.0436	(2437.0, 3956.0)
33	140	(-14.82,-0.49),(14.82,0.49)	7.8438	0.0436	(3976.0, 2199.0)
34	140	(-16.24,-0.5),(16.24,0.5)	7.8438	0.0436	(4014.0, 2344.0)
35	140	(-163.06,-0.49),(163.06,0.49)	7.8438	0.0436	(2526.0, 2747.0)
36	140	(-16.24,-0.49),(16.24,0.49)	7.8438	0.0436	(3976.0, 2329.0)
37	140	(165.88,0.49),(-165.88,-0.49)	7.8438	0.0436	(2467.0, 2244.0)
38	140	(17.65,0.51),(-17.65,-0.51)	7.8438	0.0436	(2485.0, 3855.0)
39	140	(163.06,0.5),(-163.06,-0.5)	7.8438	0.0436	(2700.0, 2404.0)
40	140	(17.65,0.5),(-17.65,-0.5)	7.8438	0.0436	(2447.0, 3937.0)
41	140	(-16.24,-0.51),(16.24,0.51)	6.4375	0.0358	(2700.0, 1593.0)

Table 3: Key of Areas for symmetric regions.

Area	NObs	Coordinates:(Area1),(Area2)	l-radius	sinb-radius	Nr of stars:(Area1, Area2)
42	80	(-38.82,0.7),(-13.41,-0.18)	9.2500	0.0514	(2449.0, 13302.0)
43	80	(52.94,-0.49),(51.53,-0.62)	6.4375	0.0592	(3339.0, 2514.0)
44	80	(-165.88,-0.01),(-139.06,0.75)	3.6250	0.0201	(1315.0, 353.0)
45	80	(132.0,-0.44),(31.76,-0.71)	5.0312	0.0280	(1581.0, 953.0)
46	80	(58.59,-0.51),(172.94,0.14)	6.4375	0.0358	(1983.0, 3075.0)
47	80	(139.06,-0.74),(44.47,-0.58)	7.8438	0.0436	(1694.0, 2578.0)
48	80	(13.41,0.13),(44.47,-0.63)	5.0312	0.0436	(2023.0, 1367.0)
49	100	(-165.88,0.89),(14.82,-0.74)	3.6250	0.0280	(479.0, 714.0)
50	100	(172.94,-0.86),(38.82,-0.52)	3.6250	0.0280	(493.0, 965.0)
51	100	(13.41,-0.87),(165.88,0.07)	5.0312	0.0280	(779.0, 1990.0)
52	100	(34.59,-0.65),(-9.18,0.88)	3.6250	0.0201	(516.0, 297.0)
53	100	(-13.41,0.87),(170.12,0.05)	5.0312	0.0280	(608.0, 2214.0)
54	120	(-158.82,-0.49),(-20.47,-0.5)	3.6250	0.0201	(602.0, 772.0)
55	120	(-17.65,-0.5),(160.24,0.51)	5.0312	0.0358	(2065.0, 1521.0)
56	120	(19.06,0.51),(160.24,0.51)	5.0312	0.0358	(1470.0, 1488.0)
57	120	(-160.24,-0.5),(-19.06,-0.51)	5.0312	0.0280	(1118.0, 1577.0)
58	140	(-16.24,-0.51),(-165.88,-0.51)	3.6250	0.0358	(1516.0, 718.0)
59	140	(17.65,0.44),(158.82,0.5)	3.6250	0.0201	(495.0, 659.0)

Table 4: Key of Areas for antisymmetric regions.