Hide and Seek in the Halo of the Milky Way
Tian, Hao

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2017

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
4. Recovering Substructures in Gaia DR1

Abstract

The aim of this chapter is to identify possible substructures such as streams, globular clusters or dwarf galaxies, in the first data release of the Gaia mission. To this end, we introduce here three algorithms, named cumulative difference, cross correlation and linear method that make use of the sky position and $G$-band magnitude for the 1.1 billion sources present in the main Gaia DR1 catalog. We first apply these algorithms on a mock Milky Way and on the SDSS catalogs, where there are (known) streams and compact systems, to test their performance. We then apply the methods to the Gaia DR1. In general, we find that it is very hard to identify streams only with sky position and $G$–band magnitude, although the algorithms work reasonably well for compact substructures, such as clusters and dwarf galaxies, as we recover many of the known satellites, including the more recently discovered Gaia-1. We also identify candidates without a known counterpart, but the reality of such objects is hard to establish without additional information. We are not only limited by the lack of e.g. multiband photometry, but also by the imprints left by the scanning law of Gaia, which introduces significant spurious structures.

Authors: H. Tian, M. A. Breddels & A. Helmi, in preparation
4.1 Introduction

According to the Λ cold dark matter model, the halos of galaxies should have formed through merging of smaller systems. As a consequence of that, the stars originating in the merged systems will be found in the halo and be distributed in the form of streams and substructures. The first clear example for the Milky Way was the discovery of Sagittarius stream (Ivezić et al., 2000; Yanny et al., 2000; Ibata et al., 2001; Majewski et al., 2003) and its progenitor the Sagittarius dwarf galaxy (Ibata et al., 1994). Thanks to wide field photometric surveys, such as SDSS and Pan-STARRS, plenty of streams have been discovered during the last 23 years, including, for example, the GD-1 stream (Grillmair & Dionatos, 2006a,b), Monoceros Ring (Yanny et al., 2003; Ibata et al., 2003), Orphan stream (Belokurov et al., 2006b) and PS1 streams (Bernard et al., 2016). These streams appear as overdensities on the sky, and can be identified in configuration space and with broadband color information.

Studying those streams will thus help us understand the halo’s formation history, such as its merging tree (Johnston et al., 2008). With dynamical information on those streams, the mass distribution of the Milky Way can also be studied (see e.g. Law & Majewski, 2010; Vera-Ciro & Helmi, 2013, for recent results). On the other hand, the faint satellites are also interesting for understanding galaxy formation, as they give insights into galaxy evolution at the low mass end, and for example could help understand the missing satellite problem (Klypin et al., 1999; Moore et al., 1999; Simon & Geha, 2007; Bullock, 2010).

In order to uncover such structures, including streams and faint bound systems, it is important to have very wide and deep sky coverage, and dynamic information is particularly useful. Gaia, launched in December 2013, is measuring precisely the spatial location and kinematic properties within its 5-year mission for all stars brighter than $G \sim 20$ with small uncertainties. For example from 9 $\mu$as at 15 mag to 130 $\mu$as at 20 mag for M6V stars\(^2\) (Prusti, 2015). By the end of the mission, full phase-space information will be obtained for stars brighter than $G = 16$. But for the fainter ones, we need surveys with instruments such as WEAVE (Dalton et al., 2012) and 4MOST (de Jong et al., 2012) to obtain radial velocity and accurate chemical information.

To identify substructures, including streams and survived satellites, there are a few algorithms which use full phase-space information (e.g. ROCKSTAR, Behroozi et al. (2013) and OPTICS, Sans Fuentes et al. (2017)). However, when only position in the sky and photometric information is available, the matched filter method is often applied, which led to the discovery of e.g. the GD-1 stream (Grillmair & Dionatos, 2006a,b). Similarly, the new class of ultra faint dwarf galaxies was identified using density maps in combination with template fitting (Zucker et al., 2006; Belokurov et al., 2006a, 2007a; Irwin et al., 2007; Koposov et al., 2007; Walsh et al., 2007; Belokurov et al., 2008, 2009, 2010).

Gaia DR1 is a catalog with more than 1.1 billion stars, and includes sky positions and $G$–band magnitudes for these sources. Although it has the advantage of being a full sky survey, it has no color information, and so the traditional matched filter method cannot generally be used. This is why in this Chapter, we test three algorithms which make use simultaneously of the characteristic magnitude distribution of a single stellar population (i.e. the luminosity function) and the presence of an overdensity on the sky. This chapter is organized as follows. We introduce the datasets in Section 4.2, and describe the methods in

\(^2\)https://www.cosmos.esa.int/web/gaia/science-performance
Section 4.3. In Section 4.4 and 4.5, we present the results of applying the methods on these different catalogs. Section 4.6 includes a discussion and our conclusions.

### 4.2 Data

As stated earlier, the first data release of Gaia contains the sky position and $G$ band magnitudes of 1.1 billion sources. As Gaia surveys all the objects with $G < 20.7$, across the full sky, there may be unknown clusters or dwarf galaxies to be discovered in the Gaia dataset. Before we analyze this dataset, we aim to establish how well our methods work for the identification of substructure. To this end, we first introduce a mock Milky Way-like catalog. In this mock catalog, we combine the GUMS catalog (Robin et al., 2012) and Aquarius C (Lowing et al., 2015) simulation. GUMS, Gaia Universal Model Snapshot, is a simulated catalog which contains all the objects which Gaia is potentially able to detect. Those objects belong to four different components, namely the bulge, thin disc, thick disc and halo. The Aquarius C catalog was introduced by Lowing et al. (2015), and lists the individual stars generated from tagged dark matter particles (Cooper et al., 2010) from halo C in the Aquarius cosmological simulation (Springel et al., 2008). To make it more realistic, the mass of the stellar halo of Aquarius C (see Table 3.1 in Chapter 3) is re-scaled to match that of the Milky Way ($2.64 \times 10^8 M_\odot$, see Robin et al., 2003), and used to replace the halo in GUMS. The main reason for this combination is to embed the substructures from merging events (from Aquarius C) into a smooth background (GUMS without the halo component). This allows us to test our methods on a more realistic catalog. As only the stars with $M_{\text{abs}} < 7$ are listed in the Aquarius C catalog (Lowing et al., 2015), we also remove similarly faint stars ($M_{\text{abs}} < 7$) from GUMS. We also convert the $g-$band magnitudes in Aquarius C to the Gaia $G-$bands following Jordi et al. (2010).

In the top panel of Figure 4.1, we show the number distribution, $\log_{10}(N)$, of the stars in the mock Milky Way on the sky computed for bins of 0.5 degree width. There are 232 progenitors in Aquarius C (also see Chapter 3), 142 of which have more than 5 stars with magnitude brighter than $G = 20.7$. We define as progenitors all systems accreted by Aquarius C, most of which are well mixed and are not visible now with the background in our selection. For 91 objects their sizes, as measured by the space dispersion $\sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2}$, are smaller than 10 kpc, and hence they are relatively compact, although not always easily identifiable in our maps.

The second catalog that we use for testing our algorithms is the SDSS dataset, which contains stars with $g$ band magnitude brighter than 22.2. Even though SDSS is not a full sky survey, it is particularly suitable for testing because many streams and dwarf galaxies have been discovered in its footprint, for example the Sagittarius stream, Orphan stream, GD-1 stream and the Palomar 5 stream (see also references in the Introduction section). We also calculate Gaia $G-$band magnitude for stars in SDSS with the conversion law from Jordi et al. (2010).

In the middle panel of Figure 4.1, we show the number distribution on the sky, $\log_{10}(N)$, of all the stars from SDSS DR12 with magnitude $10 < G < 22.2$. In the rest of this chapter, we will mainly focus on the data in the region $120^\circ < \text{Alpha} < 246^\circ$ and $-2^\circ < \text{Delta} < 60^\circ$, which includes all the streams mentioned above.
Figure 4.1: Top: Number distribution on the sky of all the stars in the mock Milky Way catalog with $10 < G < 20.7$. Middle: Number distribution of all the stars in SDSS with $10 < G < 22.2$. Bottom: Number distribution of all the stars in Gaia DR1 with $10 < G < 20.7$. The color bar indicates the number of stars in a logarithmic scale in Alpha versus Delta space in pixels of $0.5^\circ$ width. Note that because the disc region is not included in SDSS, the color spans a smaller range up to $\log_{10}(N) = 4$ per sky bin.
4.3 Method

Finally in the bottom panel of Figure 4.1, we show the number distribution on the sky, $\log_{10}(N)$ of sources in Gaia DR1. Note the features caused by the satellite’s scanning law (also see Figure 2 in Gaia Collaboration et al., 2016a), which result in artificial structures because of density fluctuations. Since we are interested only in stars, we have removed extended sources in this catalog using the criterion from Koposov et al. (2017)

$$\log_{10}(\text{astrometric\_excess\_noise}) < 0.15(G - 15) + 0.25. \quad (4.1)$$

4.3 Method

Because Gaia DR1 only provides sky positions and $G$ magnitudes and no color information, we focus on three different algorithms that just use these observables to recover substructures in the Milky Way. We are interested in dwarf galaxies, globular clusters and streams. All our algorithms compare the number of stars in a given pixel on the sky and for a given magnitude bin with some suitable distribution. We denote the number of stars in this grid as $N_{ijk}$ where $i$, $j$, and $k$ are the indices of Alpha, Delta and magnitude pixels respectively. In all algorithms, the bin size in the magnitude dimension is 0.25. We normalize the distribution in the magnitude dimension, $f_{ijk} = N_{ijk}/N_{ij}$, where $N_{ij} = \sum_k N_{ijk}$ is the number of stars in the pixel on the sky. We also define a smooth background in the counts $N_{Sijk}$. This is generated with a Gaussian kernel on $f_{ijk}$ using the function \texttt{ndimage.gaussian\_filter} in the PYTHON package \texttt{scipy} with parameters $m_{Sb}$, $m_{Sl}$, and $m_{SG}$, which are the standard deviations of the Gaussian kernel, in unit of bins. Then $N_{Sijk} = f_{Sijk} \times N_{ij}$. This smoothing allows us to get a background where neighboring bins on the sky have a similar distribution in the magnitude dimension.

Having computed the observed number distribution $N_{ijk}$ and the background $N_{Sijk}$, we are now ready to apply our three algorithms, namely the cumulative difference (CD hereafter), the cross correlation (CC hereafter) and the linear method (LM hereafter).

4.3.1 Cumulative Difference

For each bin on the sky, the CD method takes the difference of the normalized cumulative distributions in magnitude of the observation ($F_{ijk}$) and background ($F_{Sijk}$), where the cumulative distributions are defined by

$$F_{ijk} = \sum_{m=0}^{k} (f_{ijm}) \quad \text{and} \quad F_{Sijk} = \sum_{m=0}^{k} (f_{Sijm}). \quad (4.2)$$

Then the signal of $i$th pixel is calculated as

$$S_{ij}^{CD} = N_{ij} \sum_{k} |F_{ijk} - F_{Sijk}|. \quad (4.3)$$

4.3.2 Cross-Correlation

The Cross-Correlation method uses the cross correlation algorithm to calculate the signal of the observed distribution compared to that of a known template, for example, corresponding to a dwarf galaxy or globular cluster. The key idea is that the member stars in the satellite
are distributed in magnitude according to a very characteristic distribution (the luminosity function), which depicts a bump at a characteristic magnitude (corresponding to the red clump or horizontal branch), and a sharp increase as soon as stars fainter than the bottom of the RGB are included, as shown in Figure 4.2. In the analysis of the mock Milky Way catalog, we use as template $T_k$ the magnitude distribution of stars from a survived satellite from the Aquarius C simulation. The template for the analysis of the SDSS catalog is the magnitude distribution of stars from the central part of the Palomar 5 globular cluster (see in Figure 4.10). The template for Gaia DR1 is the magnitude distribution of stars from NGC 5272 (also named M3) observed by the Hubble Space Telescope (Anderson et al., 2008), and which is shown in Figure 4.15.

The difference between the background $N^S_{ijk}$ and the observed distribution $N_{ijk}$, is a residual distribution, $R_{ijk} = N_{ijk} - N^S_{ijk}$, which we cross-correlate in the magnitude direction and for each pixel on the sky to the template using the function `correlate` in PYTHON package `numpy`, with the mode `full`. This option will return an array of values with different numbers of the overlapped elements between $R_{ijk}$ and $T_k$, and we use the maximum value as the signal of the pixel, $S_{ij}^{CC}$. Furthermore, the index $k$ of the location of the maximum value will indicate how much the template was shifted, and will result in an estimate of the distance to the structure found.

### 4.3.3 Linear Method

For the Linear Method, we assume that in each pixel the observed distribution $\vec{O} = \{O_k\}$ in magnitude is composed by two components, the background $\vec{B} = \{B_k\}$ (which is obtained in the same way as in the CC method), and the structure $\vec{T} = \{T_k\}$. That means

$$N_{ij} \vec{O} = \alpha \vec{T} + (N_{ij} - \alpha) \vec{B}$$

(4.4)

where $N_{ij}$ is the total number of stars in $ij$-bin on the sky, and $\alpha$ is the number of stars that belong to the structure. Note that $\vec{O}$, $\vec{T}$ and $\vec{B}$ are vectors that represent the distributions in magnitude, and they are all normalized, $\sum O_k = 1$, $\sum T_k = 1$ and $\sum B_k = 1$. Here $O_k$ for $ij$th bin is $f_{ijk}$ and $B_k$ is $f^S_{ijk}$.

Therefore we can obtain the number of stars in the structure from the following equation

$$\alpha_{ij} = N_{ij} \frac{(\vec{O} - \vec{B}) \cdot \vec{T}}{(\vec{T} - \vec{B}) \cdot \vec{T}}$$

(4.5)

### 4.3.4 Significance

It should be noted that when we calculate the background (we use the same background $N^S_{ijk}$ for all three methods) for a pixel, we use the counts from the $ij$th pixel, implying that there is some contamination to the background, and this may be especially important for structures covering more than one pixel. Therefore we use a large standard deviation on the sky when we calculate the background $N^S_{ijk}$ namely of 10 degrees.

In order to determine the significance of a given signal $S_{ij}$ from any of our algorithms
4.4 Results

4.4.1 Testing on the mock Milky Way

As described earlier our mock Milky Way consists of the combination of GUMS (Robin et al., 2012) and the Aquarius C stellar halo (Lowing et al., 2015). We consider here stars with $G$-band magnitude from 16.7 to 20.7, which is a similar magnitude range to that covered by SDSS and Gaia DR1.

The template for the CC method is from the satellite in the simulation located at (Alpha, Delta)= $(181.0^\circ, -17.9^\circ)$ and includes all stars within 0.1 degree. In Figure 4.2 the known absolute $G$–band magnitude distribution of the selected stars has been shifted by 17.7 magnitudes to the fainter side, that means we place the satellite at a distance of 34.7 kpc. This was done to make the horizontal branch bump fall within our magnitude range $16.7 < G < 20.7$. The sharp drop in the faintest magnitude bins is caused by the conversion from SDSS $g$–band magnitude to Gaia $G$–band magnitude, and as described in Section 4.2, all stars with absolute magnitude $M_g > 7$ are removed.

For the CC method, we use the full template because for a nearer structure (with distance $D < 34.7$ kpc) fainter stars will be visible. On the other hand, for the LM method, we only consider the region of the template with $16.7 < G < 20.7$ i.e. between the green dashed lines.
Figure 4.3: The results obtained by applying the algorithms on the mock Milky Way. The top panel shows the number distribution on the sky for stars with magnitude from 16.7 to 20.7 in the $G$-band. The two black circles indicate the position of two surviving satellites. The second to the bottom panels show the distribution of the signal for the CD, CC and the LM methods, respectively. The binsize in all the left panels is 0.5° in Alpha and Delta, and the smoothing parameters are set to $(m^S_l, m^S_b, m^S_G) = (20, 20, 0.001)$, which mean that the standard deviations of the smoothing kernel for the background are 10° in the Alpha and Delta dimensions and 0.00025 in the magnitude dimension. On the other hand, the binsize in all the right panels is 0.2° in Alpha and Delta, and the smoothing parameters are set to $(m^S_l, m^S_b, m^S_G) = (50, 50, 0.001)$. The binsize in the $G$–band magnitude direction is 0.25 for all panels.
Figure 4.4: The results for a specific region selected from the left panels (binsize of 0.5°) in Figure 4.3 located at 175° < Alpha < 185° and −20° < Delta < −10°. The left most panel shows the number distribution $\log_{10}(N)$, the second panel shows the signal distribution from CD method, $\log_{10}(S_{CD})$, the third panel shows the signal distribution from CC method, $\log_{10}(S_{CC})$, and the right most panel shows the signal distribution from LM method, $\alpha$. The color coding is exactly the same as in Figure 4.3.

Figure 4.3 shows the results of applying our three algorithms. The panels on the left are for a binsize of 0.5°, while for those in the right it is 0.2°. The standard deviations for the Gaussian filter used for the smoothing to obtain the background are 10° for Alpha and Delta dimensions, and 0.00025 in the magnitude dimension (to have basically no smoothing in this direction).

In Figure 4.3, the top panels show the number distribution ($\log_{10}(N)$) of the stars on the sky. The number distribution is quite smooth at high latitudes. One satellite (in fact that used as our template) is clear above the disc with coordinates (Alpha, Delta)=(181.0°, -17.9°) for both binsizes. We can also find that it is more clearly seen in the left panel with a larger binsize, as marked with the black circle in the top panel. Figure 4.4 shows the specific region, with 175° < Alpha < 185° and −20° < Delta < −10° around the satellite.

The second panels from the top in Figure 4.3, show the signal $S_{CD}$ distribution from the CD method. We find that the survived satellite is clear in four pixels, as shown in the second panel of Figure 4.4. Besides that, a second satellite now located in the disc with (Alpha, Delta)=(99.5°, 8.0°) (and marked with a circle in the top panels) is also found with a signal which is significantly higher than the surrounding pixels, and with $SNR = 44.6$ for the 0.5° binsize.

The signal distributions $S_{CC}$ from the CC method are shown in the third panels. The distributions are similar to those from the CD method. The two satellites previously identified are also clear with the larger pixels (left panel), but weaker for the smaller pixels (right panel).

The signal $\alpha$ from the LM algorithm is shown in the bottom panels. The distribution of the $\alpha$ values seems to follow more closely the distribution of counts shown in the top panels, although it is possible to identify the same compact satellites as for the other methods.

The results from the three methods are presented more quantitatively in Figure 4.5, where we show the $SNR$ per pixel on the sky. As before the plots on the left correspond to pixels of 0.5° while those on the right to 0.2°. The top panels show the signal to noise ratio distributions
Figure 4.5: The distribution of the signal to noise ratio from the three methods CC, CD and LM for the top, middle and bottom panels, respectively, for a binsize of 0.5° on the left, and 0.2° on the right.
from the CD algorithm. We find that the signal to noise ratio $SNR$ of the two candidates (composed of a few pixels) described before are very high, with $SNR > 10$. The reason some of the neighboring pixels are "masked" is that the candidates extend over a few pixels and they have very high signal. According to Equation 4.6, the average value $< S_{i,j} >$ and the deviation $\sigma_{S_{i,j}}$ are therefore relatively large resulting in neighboring bins having a low signal to noise ratio.

In the second panel we show similar signal to noise ratio distributions from the CC method, while in the bottom panel we plot the results from the LM method. In both cases the compact high $SNR$ candidates are visible. In general for the LM, fewer pixels have $SNR > 3$. This is because the template for the LM method is the magnitude distribution of the stars in the satellite for a distance of 34.7 kpc, which means the template is not very sensitive to the structures with very different distances.

This is also apparent from Figure 4.6 where we show the distribution of $SNR$ for each of the methods for the two different pixel sizes, 0.5° and 0.2° for the left and right panels respectively. We see that the CD method returns more high $SNR$ bins than the other two methods, with the LM fewer high $SNR$ pixels. For the smaller pixel size, there are many more low $SNR$ pixels, and that is because the statistical noise is larger because of the smaller numbers of stars in each bin.

Figure 4.7 summarizes these results and shows the distribution of the pixels that have $SNR > 5$ in each method and on average $SNR > 6$. The dashed lines represent Galactic latitude $|b| = 30°$.

To develop some intuition about what kind of systems or overdensities are being detected by our methods, we explore some of the properties of the progenitors of Aquarius halo C in Figure 4.8. In the left panel, we show the angular size on the sky versus the number of stars in the relevant magnitude range for each progenitor. The angular size is calculated as $A = \frac{\sigma \times 180}{<D> \times \pi}$, where $\sigma$ is the 1D space dispersion, $\sigma = \sqrt{\sigma_x^2 + \sigma_y^2 + \sigma_z^2}$ and $<D>$ is the average distance of those stars. The red dots are the progenitors dominating the pixels with $|b| > 30°$ in Figure 4.7. In the right panel, we have plotted the number of pixels with at least 5 stars
Figure 4.7: The pixels with the SNR for each of the methods larger than 5 and average SNR larger than 6 are shown for a binsize of 0.5°. The two dashed lines represent Galactic latitude |b| = 30°.

Figure 4.8: Left: angular size versus number of stars with magnitude 16.7 < G < 20.7 for all progenitors. The red dots correspond to the progenitors dominating the candidates with |b| > 30° in Figure 4.7. Right: number of bins each progenitor has contributed to with at least 5 stars versus the number of stars from the progenitors in those bins. The dots represent the median star number found for those bins, while the error bars denote the minimum and maximum values. The colors for the dots represent the number of selected stars in each progenitor. The red dashed line shows the average number of stars with 16.7 < G < 20.7 from the background (GUMS without halo) for |b| > 30°. The black dots mark the progenitors which are dominating the candidates with |b| > 30° in Figure 4.7.
for each progenitor, against the median number of stars from the progenitor over all those bins. The error bars represent the minimum and the maximum star numbers of those bins. The colors indicate the number of the stars from each progenitor in the Aquarius C halo. The average number of stars from the background (GUMS without halo) in the bins with $|b| > 30^\circ$ is 108.6 and is represented by the red horizontal dashed line. The black dots denote the progenitors dominating the pixels with $|b| > 30^\circ$ in Figure 4.7.

From Figure 4.7, we see there are 2 pixels with $|b| > 30^\circ$ with $SNR > 5$ from each method and average $SNR > 6$. Two progenitors (the red dots in the left panel and black dots in the right panel of Figure 4.8) contribute to the two high $SNR$ candidates in Figure 4.7. The pixel with $(\text{Alpha}, \text{Delta})=(167.5^\circ, 3.5^\circ)$ is dominated by a progenitor contributing 1165 out of 1171 stars present in the pixel. The other high $SNR$ pixel is dominated by the progenitor which is used to generate the template for both the CC and LM methods, with 9962 of all the 17510 stars in the selected pixel.

Thus far we ignored high $SNR$ pixels in the disc, because the background is very variable, and a low latitude pixel will require a typically higher contribution from a progenitor for this to be detected. However there is one such pixel with $(\text{Alpha}, \text{Delta})=(99.5^\circ, 8^\circ)$ which is dominated by a single progenitor contributing with 2664 of all the 3039 stars in this pixel. The other candidates at low $b$ in the disc appear to be caused by fluctuations of the background.

In the left panel of Figure 4.8, there are a few compact progenitors with $A < 10^\circ$. Most of these progenitors are located around the center of the mock Milky Way, and are rather well mixed. For example, of the 8 progenitors shown in the right panel with a maximum contribution larger than the average value from the background (the red dashed line), six have most of the stars near the center, and hence a much higher local background. This is the case, for example for the largest progenitor with more than $3 \times 10^6$ stars, which has a scale 7.2 degree, and is not recovered by any of our methods.

If we compare the differences caused by the pixel size, we find that the compact satellites are recovered by 1 bin for the larger pixel size, but by many pixels when these are smaller. This means the angular size of the structure is larger than 0.2$^\circ$ and comparable to 0.5$^\circ$. This means that the $SNR$ will depend on the pixel size, e.g. if the structure’s angular size is comparable to 0.2$^\circ$, then it will have high $SNR$ when a smaller binsize is used, but a lower $SNR$ for a larger binsize.

In summary, the algorithms are able to recover some of the compact structures with the ideal catalog, where a bigger binsize will typically yield lower $SNR$ for smaller structures. We still need additional information to confirm or rule out those candidates caused by fluctuations of the background. As there is no thin streams in this catalog, we are not sure if the algorithms can recover them or not.

### 4.4.2 Testing on SDSS

In this subsection, we show the results of applying the algorithms on the SDSS catalog and focus on whether the known streams and satellites can be recovered. In Figure 4.9, we have marked the location of 31 known objects, 11 globular clusters (yellow crosses) which are listed in Table 4.1 and 20 dwarf galaxies (red pluses) which are listed in Table 4.2. Four streams are also marked here, the GD-1 stream with a red dashed line, the Orphan stream
Figure 4.9: The known objects in SDSS northern footprint are shown, with yellow pluses indicating the globular clusters, red pluses indicating the dwarf galaxies. The red dashed line shows the approximate path of the GD-1 stream, the yellow for the Orphan stream and cyan for the Palomar 5 stream. The two magenta lines show the bright (the lower one) and faint (upper one) arms of the Sagittarius stream. The background is the number distribution $\log_{10}(N)$ of stars with $18 < G < 22$ in SDSS.

with a yellow dashed line, the Palomar 5 stream with a cyan dashed line and the Sagittarius stream with magenta lines, where the upper one is for the faint branch and the lower one for the bright branch (all those lines are extracted from the Figure 1 in Belokurov et al. (2006b). We select the stars with $G$-band magnitude from 18 to 22 and use similar parameters for the algorithms as in the previous section, i.e. we consider binsizes in Alpha and Delta of 0.5° and 0.2°, and in the magnitude dimension of 0.25. We focus on the region $120^\circ < \text{Alpha} < 246^\circ$ and $-2^\circ < \text{Delta} < 60^\circ$, because a lot of known streams are located here (see also the figures and tables in Grillmair & Carlin, 2016). As a template for both the CC and LM methods we use the $G-$band magnitude distribution of stars within 0.07° of the globular cluster Palomar 5 at its true distance as shown in Figure 4.10.

Figure 4.11, shows the results obtained after applying the algorithms on the SDSS for binsize of 0.5° and 0.2° on the left and right panels respectively. From the number distribution (top panel), we immediately identify 5 globular clusters and 2 dwarf galaxies, marked with black circles, including the Palomar 5 globular cluster at (Alpha,Delta)=($229.02^\circ$, $-0.11^\circ$). The Palomar 5 stream is not visible in either panel (see also Figure 4.12). Notice the SDSS scanning pattern which is more pronounced in the left panel, but still visible in the right panel.

The signal $S_{CD}$ from the CD algorithm are shown in the second panels of Figure 4.11. The dense objects that are clear in the top panels are associated to pixels with high signal. Furthermore the right and left corners of both left and right panels, depict high signals which are due to the large number of stars in these regions that are quite close to the Galactic disc.
### Table 4.1: The globular clusters in the selected area shown in Figure 4.9 (Harris, 1996).

<table>
<thead>
<tr>
<th>Alpha</th>
<th>Delta</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>151.38</td>
<td>0.07</td>
<td>Palomar 3</td>
</tr>
<tr>
<td>172.32</td>
<td>28.97</td>
<td>Palomar 4</td>
</tr>
<tr>
<td>179.83</td>
<td>12.26</td>
<td>Koposov 1</td>
</tr>
<tr>
<td>182.53</td>
<td>18.54</td>
<td>NGC 4147</td>
</tr>
<tr>
<td>198.23</td>
<td>18.17</td>
<td>NGC 5024</td>
</tr>
<tr>
<td>199.11</td>
<td>17.70</td>
<td>NGC 5053</td>
</tr>
<tr>
<td>205.55</td>
<td>28.38</td>
<td>NGC 5272</td>
</tr>
<tr>
<td>211.36</td>
<td>28.53</td>
<td>NGC 5466</td>
</tr>
<tr>
<td>229.02</td>
<td>-0.11</td>
<td>Palomar 5</td>
</tr>
<tr>
<td>229.64</td>
<td>2.08</td>
<td>NGC 5904</td>
</tr>
<tr>
<td>242.75</td>
<td>14.96</td>
<td>Palomar 14</td>
</tr>
</tbody>
</table>

### Table 4.2: The dwarf galaxies in the selected area shown in Figure 4.9 (McConnachie, 2012).

<table>
<thead>
<tr>
<th>Alpha</th>
<th>Delta</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>151.77</td>
<td>16.08</td>
<td>Segue I</td>
</tr>
<tr>
<td>209.50</td>
<td>12.85</td>
<td>Bootes II</td>
</tr>
<tr>
<td>162.34</td>
<td>51.05</td>
<td>Willman I</td>
</tr>
<tr>
<td>186.75</td>
<td>23.90</td>
<td>Coma dwarf galaxy</td>
</tr>
<tr>
<td>209.30</td>
<td>26.80</td>
<td>Bootes III</td>
</tr>
<tr>
<td>210.03</td>
<td>14.50</td>
<td>Bootes I</td>
</tr>
<tr>
<td>153.26</td>
<td>-1.61</td>
<td>Sextans I</td>
</tr>
<tr>
<td>158.72</td>
<td>51.92</td>
<td>Ursa Major I</td>
</tr>
<tr>
<td>173.24</td>
<td>0.53</td>
<td>Leo IV</td>
</tr>
<tr>
<td>194.29</td>
<td>34.32</td>
<td>Canes Venatici II</td>
</tr>
<tr>
<td>172.79</td>
<td>2.22</td>
<td>Leo V</td>
</tr>
<tr>
<td>202.01</td>
<td>33.56</td>
<td>Canes Venatici I</td>
</tr>
<tr>
<td>168.37</td>
<td>22.15</td>
<td>Leo II</td>
</tr>
<tr>
<td>152.12</td>
<td>12.31</td>
<td>Leo I</td>
</tr>
<tr>
<td>143.72</td>
<td>17.05</td>
<td>Leo T</td>
</tr>
<tr>
<td>149.86</td>
<td>30.75</td>
<td>Leo A</td>
</tr>
<tr>
<td>139.01</td>
<td>52.84</td>
<td>UGC 4879</td>
</tr>
<tr>
<td>150.00</td>
<td>5.33</td>
<td>Sextans B</td>
</tr>
<tr>
<td>155.43</td>
<td>18.09</td>
<td>Leo P</td>
</tr>
<tr>
<td>243.45</td>
<td>54.37</td>
<td>KKR 25</td>
</tr>
</tbody>
</table>

Table 4.1: The globular clusters in the selected area shown in Figure 4.9 (Harris, 1996).

Table 4.2: The dwarf galaxies in the selected area shown in Figure 4.9 (McConnachie, 2012).
The Palomar 5 stream is not visible with either bin size. The smaller bin size returns a "sharper" signal for the Palomar 5 globular cluster in the right panel.

The signals $S_{CC}$ from the CC algorithm are shown in the third panels, and the results are quite similar to those from the CD method.

The bottom panels show the results from the LM method. Comparing with Figure 4.9, we can find that the Sagittarius stream is a bit more apparent, while the Palomar 5 stream is not recovered. On the other hand the globular cluster NGC 5904 located above Palomar 5 (see Figure 4.12 for a zoom-in of this region) is not recovered with either bin size, although it is recovered by the CD and CC methods. This is because the template from the Palomar 5 globular cluster with its distance 23 kpc (larger than the distance 7.5 kpc of NGC 5904), suffers from incompleteness when matched to a nearby object. That is also the reason that some other known objects are not found with the LM method. Note as well that the scanning path appears more strongly in this method than for the other two for both pixel sizes.

In none of the three methods the stream from Palomar 5 is recovered, the main reason being that each pixel along the stream has a comparable star number to the background. Secondly, the background for all the methods is contaminated by Palomar 5 itself, which does decrease the signal from all the methods slightly.

In order to pick out those candidates with high significance we calculate the signal to noise ratio with Equation 4.6. Figure 4.13, from top to the bottom shows the pixels with $SNR > 5$ from the CD, CC and LM methods respectively. As before the LM method finds fewer high $SNR$ candidates. This is caused by the limited apparent $G$ magnitude distribution of the template. For objects at different distances, the LM will find it difficult to return a high $SNR$. For those candidates found by all the methods and with average signal to noise ratio larger than 6, we find in all cases that they correspond to a known object.

According to the algorithm, those objects with a scale comparable to the bin size should have a higher $SNR$ than in the case of a much larger bin size. Take Palomar 5 as an example, with a scale about 0.12 degree on the sky, the $SNRs$ from the three methods for bin size of 0.2° are 25.28 (CD), 17.01 (CC) and 14.75 (LM), while the $SNRs$ for bin size 0.5° are 6.62 (CD), 3.99 (CC) and 5.43 (LM). This means that a smaller bin size is generally better for the
4.4 Results

Figure 4.11: The top panel shows the number distribution $\log_{10}(N)$ of the stars with $G$-band magnitude from 18 to 22. The black circles represent the locations of 7 very significant known objects, including 5 globular clusters and 2 dwarf galaxies. The second panel is the signal $\log_{10}(S_{CD})$ from the CD method. The third panel is the signal $\log_{10}(S_{CC})$ from the CC method. The bottom panel shows the number ($\alpha$) distribution with the LM method. The binsizes and kernel smoothing parameters are the same as in Figure 4.3.
Figure 4.12: The results of a zoom in of Figure 4.11 around the region where the Palomar 5 globular cluster, (Alpha, Delta)=(229.0°, −0.11°), and NGC 5904, (Alpha, Delta)=(229.64°, 2.08°), are located.

Figure 4.13: Pixels with SNR > 5 from the three methods, CD, CC and LM in the top, middle and bottom panels, respectively.
recovery of smaller structures. The reason that the Sagittarius stream does not have a strong signal and $SNR$ is because its contrast compared to the background is smaller, and further the background has been computed over similar scales as the stream implying it has similar distribution as the stream itself.

On the other hand, the distribution of the $SNR$ of all the bins are shown in Figure 4.14, with binsize $0.5^\circ$ in the left panel and $0.2^\circ$ in the right panel. The last bins show the numbers of pixels with $SNR > 10$. As with the mock Milky Way catalog, the CD methods are returning relatively higher $SNR$. By crossmatching to a list of known globular clusters, of the 11 located in this region, 5 of them are recovered by all the three methods. 4 out of 20 known dwarf galaxies in this region are recovered by all the methods with binsize $0.5^\circ$. In the results with $0.2^\circ$, two more globular clusters and two more dwarf galaxies are found. For those that are not identified, their $SNR$ is lower than 3, such as for Palomar 3 with $(\alpha, \delta) = (151.4^\circ, 0.1^\circ)$, whose $SNR$ from the three methods are 1.7 (CD), 2.7 (CC) and 1.0 (LM). Some other objects are recovered by two of the three methods with high $SNR$, like NGC 5904, for which the $SNR$ of the CD, CC and LM methods is 6.0, 6.9 and 1.5, respectively.

We conclude that the methods are working relatively well for compact objects, especially given that no color information has been used.

4.5 Results from Gaia DR1

4.5.1 Analysis in the SDSS footprint

In this section we apply our algorithms to Gaia DR1. This dataset suffers from incompleteness at the faint end (which also varies across the sky), and which is especially severe in crowded regions such as in the central parts of globular clusters. This is illustrated in Figure 4.15. In this figure the green histogram shows the distribution of stars in Gaia DR1 within $0.1$ degree from the center of NGC 5272 (also known as M3), while the black histogram shows the distribution observed by the Hubble Space Telescope (Anderson et al., 2008), where the Gaia $G$–band magnitude of stars is calculated from the $V$ and $I$ magnitudes following the conversion law from Jordi et al. (2010). We find that from $G = 18$, the difference becomes larger and larger, because of the incompleteness of Gaia DR1. This is why in this section, we
Figure 4.15: The template comparison with data from Gaia DR1 (green line) and Hubble observation (black line) (Anderson et al., 2008).

use the distribution of stars in NGC 5272 from the Hubble Space Telescope as the template. The reason for choosing NGC 5272 is that it is quite near (approximately at 10.2 kpc), which implies that it will not bring magnitude incompleteness when we shift it in distance.

Figure 4.16 shows the results obtained with the three algorithms when applied on the SDSS-like footprint of Gaia DR1 for $G$-band magnitudes from 14 to 20.25. The binsizes are 0.5° for the left panels and 0.2° for the right panels. The smoothing parameters for the computation of the background are $(m^L, m^S, m^G) = (20, 20, 0.001)$ for the left panels and $(50, 50, 0.001)$ for the right panels.

The top panel shows the number distribution of stars and may be compared to a similar selection applied on SDSS and shown in Figure 4.17. Note that some of the satellites are better seen in Gaia than in SDSS. In fact the same satellites are recovered both in SDSS and Gaia DR1 as we discuss below. The exception is Palomar 5 which is not identified in Gaia DR1 by any method. The reason is that here we use a brighter magnitude range, which leads to a large fraction of member stars being lost (see Figure 4.10), and hence resulting in a low SNR for Palomar 5.

Figure 4.18 shows pixels with $SNR > 5$ for all three methods. Cross-matching with the known objects from Tables 4.1 and 4.2, we find that NGC 4147, NGC 5053, NGC 5272, NGC 5466, NGC 5904, Leo I and Leo II are identified for a pixel size of 0.5°. NGC 4214 (an irregular galaxy at (Alpha, Delta)=$($183.9°, 36.3°$)$) is also identified. All those pixels with $SNR > 5$ found with bin size 0.5° are also identified for bin size 0.2°. In this case two more objects are recovered, Sextans I and NGC 4449.

4.5.2 Analysis of the region around Gaia-1 and Gaia-2

In the previous sections, we have shown that the algorithms are working relatively well for the identification of compact systems such as globular clusters, and the smaller bin size is more sensitive for the smaller size structures. Inspired by the recently discovered Gaia-1 and Gaia-2 (Koposov et al., 2017) with an overdensity searching algorithm, in this section we first test our algorithms on two regions around them to see if we can also recover these systems. Now we set the bin size to 0.05°. We do this because we aim to discover structures with a similar or smaller size than Gaia-1, whose half-light radius is about 0.1° (Koposov et al.,
Figure 4.16: The results of applying the algorithms on the SDSS footprint in Gaia DR1 are shown in the various panels. The top panel shows the density distribution of the stars with magnitude from 14 to 20.25 in the $G$-band. The circles are the same known objects as in Figure 4.11. The second, third and fourth panels depict the signals obtained with the CD, CC and the LM algorithms respectively. We use the same binsizes and parameters for the background as in Figure 4.11.
Figure 4.17: The number distributions from SDSS for the same magnitude selections as that used in Figure 4.16. Binsize is $0.5^\circ$ in the left panels and $0.2^\circ$ in the right panel.

Figure 4.18: Pixels with $SNR > 5$ for the CD (top), CC (middle) and LM (bottom) algorithms from Figure 4.16. The parameters for $SNR$ calculation are the same as those used for the mock Milky Way and SDSS catalogs.
4.5 Results from Gaia DR1

2017), or Gaia-2, which is even smaller, about 0.03°.

The template in this section is the same as that used in the previous section and shown in Figure 4.15, with magnitude from 14 to 20.25 in the $G$–band. The parameters for the Gaussian kernel $(m_S^f, m_S^b, m_S^G)$ are set to (40, 40, 0.001), while the $SNR$ is computed following Equation 4.6 using a region of 40 bins around the pixel in question.

As we mentioned in the previous sections, it is difficult for the LM method to recover objects that have very different distance from that of the template. In this section, we shift the template with magnitude offset from -1.5 (brighter side) to 5 (fainter side), which means we relocate the template at an artificial distance from 5.2 kpc to 102 kpc. In this case, the LM method should be able to discover substructures with different distances. For each magnitude offset (artificial distance), we keep a record of those sky pixels with $SNR > 3$. For those pixels recorded many times for different magnitude offsets, we keep the information ($SNR$ and magnitude offset) which yields the largest $SNR$.

We first select two regions with $100° < \text{Alpha} < 110°$ and $-20° < \text{Delta} < -10°$, which includes Gaia-1 and some open clusters, and $25° < \text{Alpha} < 35°$ and $50° < \text{Delta} < 60°$, which includes Gaia-2. As before we show the results from the three methods in Figure 4.19 and Figure 4.20, for Gaia-1 in the left panels and for Gaia-2 in the right panels.

The top panels of Figure 4.19 show the number distributions on the sky, $\log_{10}(N)$, where Gaia-1 and Gaia-2 are marked with red circles. We see that Gaia-1 has a relatively high density centered around $(\text{Alpha,Delta}) = (101.45°, -16.75°)$. The extension to smaller Alpha is caused by Sirius (Koposov et al., 2017). The open cluster Berkeley 25 with $(\text{Alpha, Delta}) = (100.3°, -16.5°)$ is also visible in the density distribution, but does not stand out clearly in the signal distributions as can be seen from the bottom panels. Meanwhile Gaia-2, which is smaller, is still visible although not as clearly as Gaia-1.

The top and second panels of Figure 4.20 show the $SNR$ distributions of the pixels from the CD and CC methods. The third panels show the pixels with $SNR > 3$ from the LM method when the template is shifted in magnitude. As before the red circles indicate the approximate positions of Gaia-1 and Gaia-2.

From the left panels, we can see that the CD and CC methods recover Gaia-1 with a few pixels with $SNR > 4$, while the LM method identifies a neighboring pixel with a $SNR = 3.5$ for a magnitude offset -1.5, corresponding to a distance 5.1 kpc, which should be compared to its estimated distance of 4.6 kpc by Koposov et al. (2017). Gaia-2 is recovered with the CD and LM methods with $SNR$ 3.3 (CD) and 3.6 (LM, for a magnitude offset of -1.5).

4.5.3 Analysis of the full sky

Encouraged by the previous results, we now apply the algorithms with same parameter values on the whole sky. For processing purposes, we divide the full sky into 18 × 9 pieces in Alpha and Delta, of $20° \times 20°$ size without overlap.

All candidates from the CD and CC algorithms with $SNR > 3$ are recorded. As we mentioned in the last section, the template generated from NGC 5272 with distance 10.2 kpc, will be shifted in magnitudes with offset from -1.5 to 5, which corresponds to a distance range
Figure 4.19: The results of applying the algorithms on selected regions around Gaia-1 and Gaia-2 in the left and right panels respectively. The top panel shows the number distribution on the sky $\log_{10}(N)$ of the stars with magnitude from 14 to 20.25 in $G$-band. The second panel is the signal distribution $\log_{10}(S_{CD})$ with the CD algorithm. The third panel is the signal distribution $\log_{10}(S_{CC})$ with the CC algorithm. The fourth panel shows the number distribution ($\alpha$) with the LM method. The binsize in all the panels are $0.05^\circ$ with smoothing parameters for the background of $(m_{SI}, m_{SB}, m_{SG})=(40,40,0.001)$. The red circles represent the Gaia-1 and Gaia-2 satellites in the left and right respectively.
Figure 4.20: The pixels with $SNR > 3$ for each method are shown, color coded by the average $SNR$. The red circles indicate the locations of the Gaia-1 and Gaia-2 satellites in the left and right panels respectively.
Chapter 4. Recovering Substructures in Gaia DR1

Figure 4.21: The pixels recovered by all the three methods with average signal to noise ratio larger than 5 are shown with colored dots, where the color coding is indicated by the bar. We have removed from here all pixels that have a cross match to a known object as well as those with $|b| > 85^\circ$. The background is the number distribution of stars in Gaia DR1 with $14 < G < 20.25$.

from 5.1 kpc to 102 kpc. Thus in the case of the LM, of all the bins with $SNR > 3$, we select the magnitude offsets which provide the maximum $SNR$ for each pixel.

Finally we do a cross match among the results from the three methods. If one candidate pixel is listed by all the three methods with $SNR > 3$, then we say it is a plausible candidate.

We remove the high latitude candidates with $|b| > 85^\circ$ because of edge effects. We then merge those neighboring high $SNR$ pixels (up to a maximum of 2 pixels). In this way we obtain 368 candidates with average $SNR > 5$, where 110 can be cross matched to known objects, from a list of globular clusters, open clusters, dwarf galaxies, NGC objects and star associations provided by SIMBAD\(^3\). The remaining 258 candidates are shown in Figure 4.21, color coded by their average $SNR$.

Figure 4.21 shows that most of the candidates are not in the disc, and that there are many candidates in the poorly scanned area centered around Alpha of 150$^\circ$. If we remove all candidates in this region, we are left with 203 objects. The average $SNR$ distribution is shown in Figure 4.22 with the green line. The distribution of the candidates recovering known objects is represented by the red line.

In order to confirm that the tentative candidates without a known counterpart are true structures, it would be helpful to inspect their color magnitude diagram (CMD).

In Figure 4.23 we show the CMDs of one candidate that has been cross-matched to the open cluster Berkeley 39 with average $SNR = 6.1$. In both panels, the red dots are the stars in the candidate pixel, and the black dots are from the $3 \times 3$ neighboring pixels with $|\text{Alpha} - \text{Alpha}_C| < 0.075^\circ$ and $|\text{Delta} - \text{Delta}_C| < 0.075^\circ$, where $(\text{Alpha}_C, \text{Delta}_C) = (116.675^\circ, -4.625^\circ)$. The contours show the number distribution of the stars from the background box with

\(^3\)http://simbad.u-strasbg.fr/simbad/
4.5 Results from Gaia DR1

Figure 4.22: The average SNR histogram distributions of the candidates. The red line shows the distribution of the candidates which are recovering known objects and the green line shows the distribution of those which are not recovering known objects. Those associated with the scanning path (130° < Alpha < 170° and −40° < Delta < 50°) are not taken into account.

|Alpha − AlphaC| < 2° and |Delta − DeltaC| < 2°. The top subplots show the distribution of \( G − K \) (left) and of \( J − K \) (right), where the black line corresponds to the distribution of stars in the background box and the green of the candidate stars, and the error bars represent the Poisson noise. The cyan line in the sub panel on the right shows the magnitude distribution of the candidate stars which have colors from 2MASS. The red line shows the \( G \) magnitude distribution of all the stars in the candidate box, both histograms are normalised to the total number of the stars in the pixel. The pink and brown filled histograms represent the \( G \) magnitude distributions of all the stars and of those stars with colors in the background box, and are also normalised. The blue straight line indicates the position of the biggest difference found by the CD method. The black and yellow solid curves represent the shifted template with offset from the CC and LM methods, respectively. In the right panel, only the \( K \)−band magnitude distribution of the stars in the candidate box and background box are shown with green lines and pink bars, respectively.

Berkeley 39 is located at about 4.78 kpc. The distance derived from the magnitude offset in LM method is 5.1 kpc, and consistent with that derived from the CC method (as can be seen from the fact that the black and yellow curves overlap perfectly).

Figure 4.24 shows the CMDs for a candidate that does not have a known counterpart, located at \((\text{Alpha}_C, \text{Delta}_C) = (82.325°, 47.625°)\), which has average \( \text{SNR} = 5.8 \). The CC and LM methods obtain a similar distance for this candidate, about 9.09 kpc (magnitude offset -0.25) and 8.10 kpc (magnitude offset -0.5)\(^4\), respectively. The magnitude where the maximum difference is found in the CD method is also indicating a similar distance assuming this corresponds to the location of MSTO stars with absolute magnitude \( M_G = 4 \) (Gaia Collaboration et al., 2016a).

Figure 4.25 shows the CMDs of another three candidates with Galactic latitude \( |b| > 15 \)

\(^4\)The magnitude offset from CC method is relative to the binsize in magnitude dimension which is 0.25, while the offset binsize for the LM method is 0.5.
Figure 4.23: The CMD of one candidate is shown, corresponding to the open cluster Berkeley 39. Here we do not include any extinction correction. The left panel shows the CMD of $G$ versus $G - K$. The red dots in the main plot are the stars from the candidate pixel, while the black dots in the main plot are the stars within the box $|\text{Alpha} - \text{Alpha}_C| < 0.075^\circ$ and $|\text{Delta} - \text{Delta}_C| < 0.075$ from the candidate coordinate $(\text{Alpha}_C, \text{Delta}_C) = (116.675^\circ, -4.625^\circ)$. The contours show the distribution of stars selected in the background, with $|\text{Alpha} - \text{Alpha}_C| < 2^\circ$ and $|\text{Delta} - \text{Delta}_C| < 2$. The top subpanels show the normalised color distributions of the candidate stars and background stars in green and black respectively. The error bars indicate Poisson errors. In the subpanel of the $G - K$ CMD, the pink bars show the normalised $G$ magnitude distribution of all stars in the background box, the brown bars show the same distribution of those stars with color information in the background box. The red line shows the same distribution for all stars in the candidate pixel, while the cyan line shows the same normalised distribution for those with color information. The blue line shows the magnitude where the largest difference comes from the CD method, the black solid line shows the template distribution with a shift according to the CC method, the yellow line also shows the shifted template with offset from the LM method. In the right panel, the main plot is the CMD of $K$ versus $J - K$, while the top subplot shows the distribution of color for the candidate and background stars. The right subpanel show the magnitude distributions, the red bars for stars in the background box, and the green line for stars in the candidate box. The uncertainty in color should be smaller than 0.1 with stars brighter than $K \sim 15$ and up to 0.4 around $K \sim 16$ (See http://www.ipac.caltech.edu/2mass/releases/allsky/doc/sec2_2b.html).
Figure 4.24: Similar as in Figure 4.23, but for a possible new candidate. The coordinate of this candidate is \((\text{Alpha}_C, \text{Delta}_C) = (82.325°, 47.625°)\)

which have no known counterparts. The average \(SNR\) for all the three candidates are higher than 5.4. From these CMDs it is not possible to discern whether they are truly distant objects or whether they are simply caused by noise due to the small number of associated stars.

A clear limitation that we are facing is the lack of or incompleteness in the color information that would help confirm the nature of a candidate. Particularly for candidate pixels with a few stars, this is an important issue, since the presence of the horizontal branch, the red clump or the MSTO can indicate a clear structure in CMD, as shown in Koposov et al. (2017). For our candidates, such as the one in in Figure 4.24, the red clump or horizontal branch stars are not very clear, even when considering the 9 neighboring bins (including the candidate pixel itself).

### 4.6 Discussion and summary

Aimed at Gaia DR1, we have developed and tested three algorithms to identify substructures on different catalogs, including one mock Milky Way and SDSS. The results obtained indicate that the methods perform reasonably well in recovering compact substructures, such as dwarf galaxies and globular clusters. This is especially encouraging given that we do not use color information. On the other hand, we are not able to recover extended structures such as the Sagittarius stream present in the SDSS catalog.

We then proceed to analyze Gaia DR1 and focus on the compact structures. Of all the candidates identified with average \(SNR > 5\), 110 have a known counterpart according to SIMBAD. These counterparts include dwarf galaxies, globular and open clusters. After further cleaning, 203 tentative candidates remain, but the determination of their true nature will require follow up. Unfortunately, we are severely limited by small numbers and by the lack of color information for the fainter magnitudes. The situation will however improve significantly with the next Gaia data release planned for April 2018.
Figure 4.25: As in Figure 4.23, but we now show the CMDs of three candidates without a known counterpart. The coordinates of the candidates (Alpha$_C$, Delta$_C$) are (334.025°, -37.525°), (186.525°, -27.925°) and (58.575°, 30.675°).
Acknowledgements

HT acknowledges support from the China Scholarship Council with project code 201206040040. AH acknowledges financial support from a VICI grant from the Netherlands Organisation for Scientific Research, NWO. MB and AH are grateful to NOVA for financial support. This work has made use of data from the European Space Agency (ESA) mission Gaia (http://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC, http://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement.