Hvorfor er t-SNE en god ide?
Hvad er forskellen på stjernespektre og kvasar spektre?
Hvad gør perplexity parameteren?
Hvorfor kan det tænkes, at grupperne blander sig?
Hvordan kan vi slippe af med flere stjerner?

t-SNE and Quasar Selection

Exploring the efficacy of using t-SNE in quasar selection

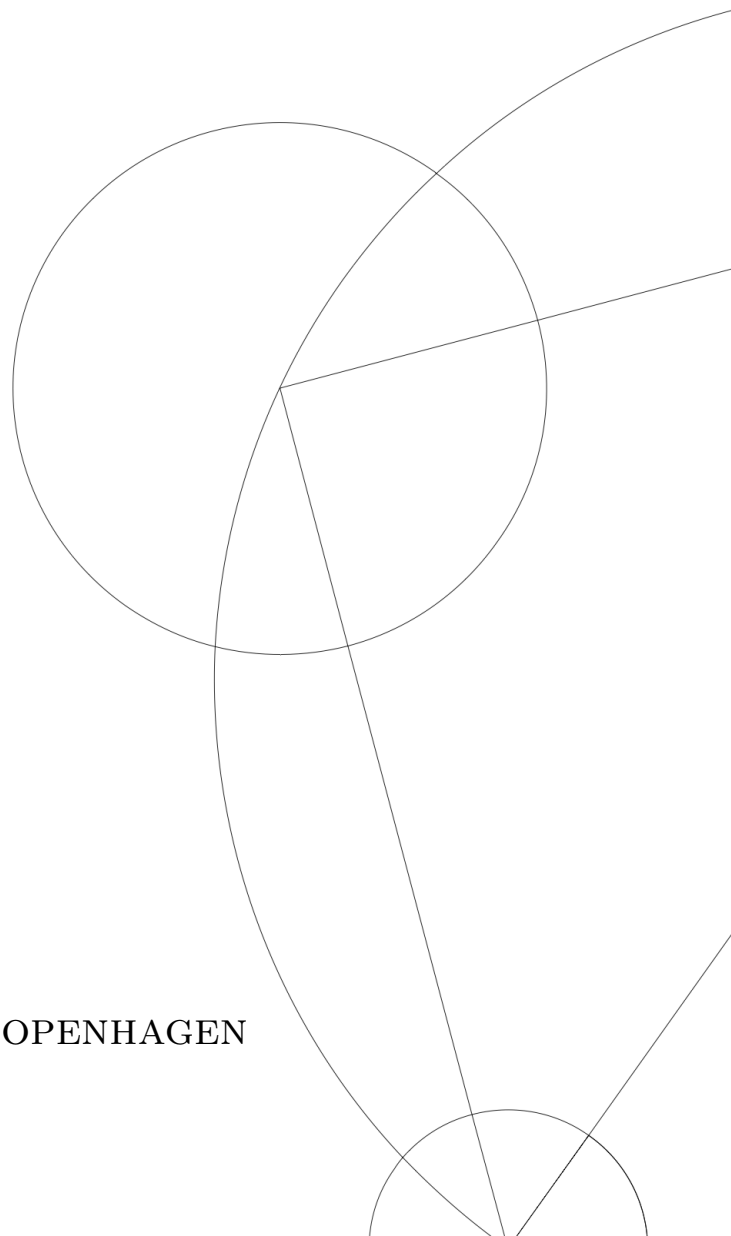
Bachelor Project

Written by *Kasper Holmberg Kamstrup*

January 20, 2021

Supervised by

Johan Fynbo



University of Copenhagen

Name of Institute: Niels Bohr Institute

Author(s): Kasper Holmberg Kamstrup

Email: dfs989@alumni.ku.dk

Title and subtitle: t-SNE and Quasar Selection
- Exploring the efficacy of using t-SNE in quasar selection

Supervisor(s): Johan Fynbo

Handed in: January 20, 2021

Defended: February 5, 2021

Abstract

This thesis aims to explore whether the t-SNE algorithm can be used in the problem of quasar selection, specifically whether the algorithm can distinguish between quasars and stars. The relevant aspect of the t-SNE algorithm are summarized. Then follows a discussion of the data and how it was selected. This thesis uses a catalog that was compiled using astrometric data from the Gaia satellite and photometric data from the Sloan Digital Sky Survey, the UKIRT Infrared Deep Sky Survey, and the Wide-field Infrared Survey Explorer. The experimental setup and results are then outlined. The thesis then discusses the results, that they suggest the application of t-SNE in quasar selection could prove useful, and how one might improve the experiment. Finally, it is concluded that while this thesis did show that the t-SNE algorithm clustered the data, the level contamination prevents any definitive conclusion, and that further research is needed.

Contents

1 Introduction

In recent years, there has been a dramatic increase in the application of machine learning in many fields of science. This year a paper by Christian K. Jespersen [8] managed to successfully classify gamma-ray bursts based solely on their prompt emissions by using the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm. The classification of astronomical objects can often be difficult. Using a simple set of criteria can lead to an incomplete classification or include too many contaminants. Jespersen's paper presents an intriguing new option with t-SNE. In this thesis I will thus attempt to apply the t-SNE methodology to the problem of quasar selection. By quasar selection, what is meant is specifically selection of quasars among point sources on the sky on the basis of photometric measurements.

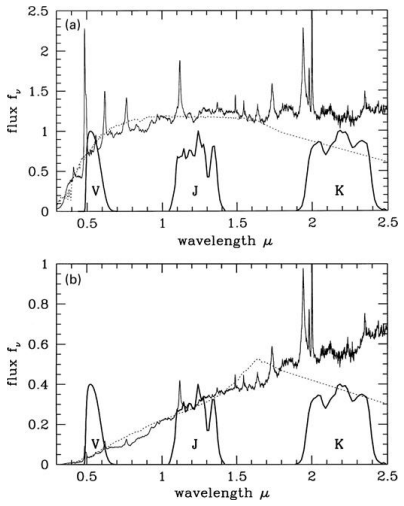


Figure 1: A figure from [14] illustrating the different SEDs of stars (solid line) and quasars (dotted line). The top panel shows how a normal quasar at redshifts around 3 looks like a K-star in the optical spectrum, but has K-band excess relative to a K-star. The bottom panel shows how a dust obscured z=3 quasar looks like an M-star in the optical spectrum, but again has excess in the K-band relative to an M-star.

There is good reason to suspect that a selection based purely on photometric information might work. Though quasars and stars can appear quite similar, they have different spectral energy distributions (SEDs). This is illustrated in figure 1 which plots the SEDs of two different stars and quasars. In this case the stars and quasars appear similar in their optical spectrum, but the quasars have significant K-band excesses which t-SNE might pick up on. Furthermore, their luminous outputs also have different physical origins. While stars are mainly powered by hydrogen fusion and roughly follow a black-body curve, quasars are believed to be black holes with an accretion disk [11][Chapter 9]. It therefore stands to reason that their SEDs are different in a way that t-SNE might pick up on.

2 t-SNE

The t-SNE algorithm is a dimensionality reduction algorithm developed by Laurens van der Maaten and Geoffrey Hinton [13], adding on the already existing SNE algorithm. In this project

I have used the scikit-learn [9] implementation of the algorithm in Python. This implementation uses the Barnes-Hut method of t-SNE as the default method. The original algorithm has $O(n^2)$ complexity whereas the Barnes-Hut method only has $O(nlog(n))$ complexity. This is achieved by approximating the attractive and repulsive terms in the t-SNE gradient using tree-based algorithms, as discussed by van der Maaten in his paper "Accelerating t-SNE using Tree-Based Algorithms" [12]. Though the dataset in this thesis has less than a million entries, it is still much larger than 10,000, this thesis therefore uses the Barnes-Hut method. A note on nomenclature, van der Maaten and Hinton refer to the input vectors as datapoints and output vectors as map points, a convention that will also be used here.

Van der Maaten and Hinton, in their original paper [13], used Principle Component Analysis (PCA) to first reduce their datasets's dimensions to 30 before running t-SNE on their examples. This was done in order to improve computational performance. Jespersen, in his paper, does not do this, though several other sources [16, 4] recommend first running PCA to reduce the number of dimensions to 50. Since the catalog used in this thesis only has 13 dimensions, it should not be an issue to run t-SNE without first running PCA. However, if one were to try and classify quasars from a catalog with significantly more dimensions, it might be prudent to run PCA or a similar dimensionality reduction algorithm first.

The t-SNE algorithm has an important hyper-parameter, perplexity. Perplexity is explicitly defined as:

$$perp(P_i) = 2^{H(P_i)}$$

Where H is the Shannon entropy of P_i, which is the conditional probability distribution in the high dimensional space, which along with the conditional probability distribution in the low dimensional space Q_i is used to calculate t-SNE's cost function [13]. Perplexity can be interpreted as the number of average neighbors each datapoint has, and is usually between 5 and 50. Perplexity is something that has to be dialed in every time the algorithm is used on a new dataset. This is an import way in which t-SNE and its likes are different from more conventional machine learning approaches where a number of parameters have to be determined in a training phase [1][Chapter 1].

As the name implies, t-SNE is a stochastic algorithm. This means that different runs of the algorithm on the same dataset are going to produce different maps unless the seed is manually

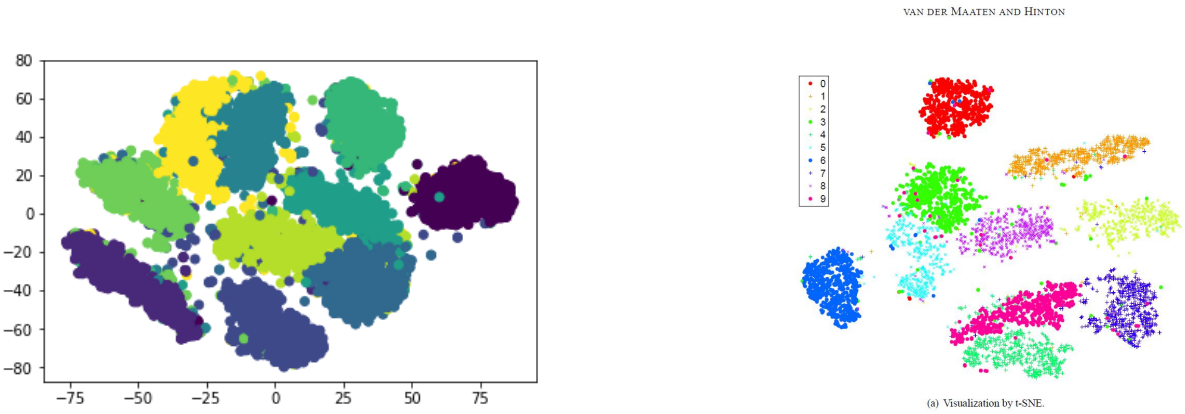


Figure 2: On the left, my attempt at applying t-SNE to the MNIST dataset. On the right, again t-SNE applied to MNIST but from van der Maaten's original t-SNE paper [13]

kept constant.

Before running any experiments on the data for this thesis project, I wanted to make sure that the code worked, and that I knew how to use it. One of the experiments van der Maaten conducted in his original paper [13] used the MNIST dataset. I therefore ran the t-SNE algorithm on the MNIST dataset, and fortunately I got a result very similar to van der Maaten, though with less separation between the clusters, which can be seen in figure 2.

3 Data

Several previous studies [5, 7, 6] have used the Gaia satellite for quasar selection, so Gaia will also be used here. The catalog of point sources was assembled from the Gaia Data Release 2 (Gaia DR2) [2] specifically and several photometric surveys discussed later. Using the catalog from Gaia DR2, the point sources had to satisfy three criteria in order to be included in this catalog.

Firstly, they had to be within 30 degrees of the North Galactic Pole (NGP). The reasoning behind this is that since the galactic poles are the furthest away from the Milky Way disk, the poles have low numbers of contaminating stars, which has been confirmed in a previous study [5]. Though all point sources within 1 degree of NGP were considered, only point sources from a sub area were ultimately added due to limited sky coverage of the photometric satellites, as depicted in figure 3.

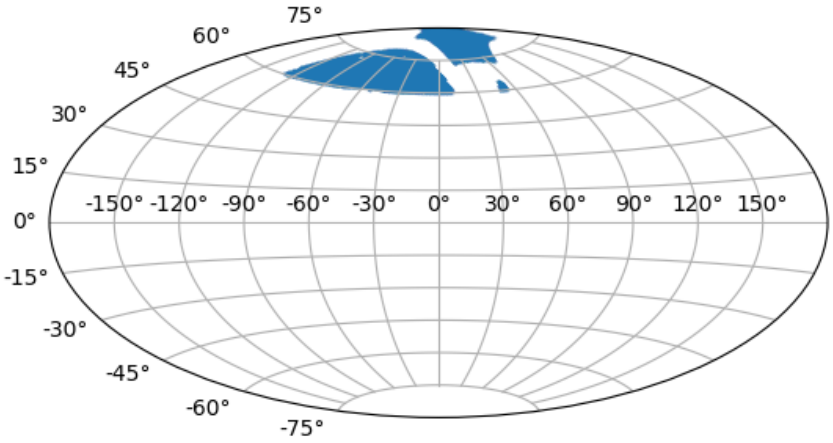


Figure 3: The blue region on this Aitoff projection of the sky in galactic coordinates shows the location of the point sources used [Picture provided by Johan Fynbo]

Secondly, the point sources had to have no proper motion. The Gaia satellite delivers precise astrometric measurements of various objects in the sky. These measurements can be used to calculate very precise parallaxes and proper motions. Since quasars are at least hundreds of megaparsecs away, they should have essentially no measurable proper motion to the precision available to Gaia (a few tenth mas/yr). The precise criterion used is that the proper motion divided by the error in the proper motion measurement had to be less than 3, i.e. that the proper motion is consistent with 0 within 3σ.

Thirdly, they had to have a magnitude, as measured by Gaia, of less than 21.

All the point sources from Gaia DR2 that satisfied the criteria, were then cross-matched with a set of photometric catalogs (done by Kasper Heintz prior to the work in this thesis). The photometric data was drawn from three sources:

- Sloane Digital Sky Survey (SDSS, [3])
- UKIRT Infrared Deep Sky Survey (UKIDSS, [15])
- Wide-field Infrared Survey Explorer (WISE, [17])

SDSS has 5 data channels from the ultraviolet into the near-infrared (u, g, r, i, z), UKIDSS has 4 near-infrared channels (Y, J, H, K), and the satellite mission WISE has 4 infrared channels

($W1$, $W2$, $W3$, $W4$) for a total of 13 dimensions. Bear in mind that the point sources in the data set had to be detected by each survey, but there was no requirement of detection in all passbands. This means that several data points contain "Not a Number" values or nans, which have to be dealt with.

At this point, a lot of the datapoints were of unknown source, which meant another filtering would have to take place such that all the datapoints were either a known star or a known quasar. The whole point of this thesis project is to test whether the t-SNE algorithm can correctly select quasars in a dataset that contains both quasars and stars. It would thus be difficult to asses whether or not the algorithm actually did that if the datapoints were unlabeled, hence this filtering. After this step, at total of 910 stars and 16,747 were included in the catalog.

Mistake

4 Experimental Setup

The most immediate concern is how to handle the nan values. One approach is to discard any datapoint with nan values in one of the channels (nan infested datapoints), though this risks biasing the data since the fact that a detection was not made might be a crucial piece of information. One reason why a datapoint might contain a nan value in a channel might be because the signal was too weak to be measured. With this in mind it might be appropriate to replace all the nan values with zeros. In this case, however, what is being measured is the magnitude, a logarithmic scale where low numerical values correspond to a high signal and vice versa. In this system, a magnitude of zero is quite large. With this in mind, simply discarding nan infested datapoints seemed to make the most sense. Figure 4 shows that while most data channels contained at least a few nan values, the data channels from UKIDSS were by far the most nan infested.

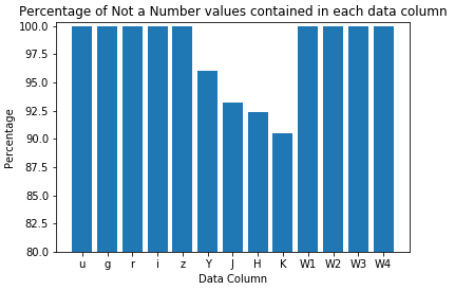


Figure 4: A bar chart starting from 80 % illustrating how many nan values each data channel in the dataset contained.

Several different experiments were run with different setups in order to explore which aspects of the data affect the algorithm. All in all, five experiments were conducted:

- The first used all 13 data channels (u, g, r, i, z, Y, J, H, K, $W1$, $W2$, $W3$, $W4$).

- The second excluded the SDSS data channels (Y, J, H, K, $W1$, $W2$, $W3$, $W4$).
- The third took a random sample of 650 quasars along with all the stars, which resulted in a more even distribution of stars and quasars (668:650 after dropping nan-infested datapoints).
- The fourth used the data channels g, r, i, z, Y, J, H, K, $W1$, $W2$.
- The fifth series of runs using all 13 data channels, but different G magnitude cut-offs, specifically $G < 21$, $G < 20$, $G < 19$, and $G < 18$.

All of these runs were also done with a series of perplexity values, specifically: 5, 30, 50, 100, 200, 600, 1000.

5 Results

The first experiment included all the data channels from the three different photometric satellites, and resulting in some clustering behavior (see appendix for full results). This experiment would seem to suggest that the ideal perplexity value is somewhere between 30 and 100. The runs with perplexity 30, 50, and 100 can be seen in figure 5. While there is a small cluster consisting only of stars in these results, the main cluster which should consist only of quasars is contaminated with stars. I therefore conducted more experiments to see if this result could be improved.

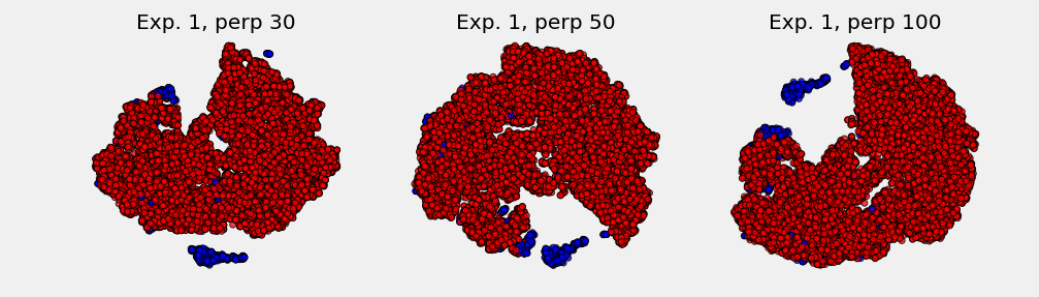


Figure 5: Selected results from the first experiment. Quasars are red, and stars are blue.

For the second experiment, I elected to exclude the data channels from the SDSS satellite. The logic is that since some quasars appear similar to some stars in the optical spectrum (see figure 1), and since SDSS mainly measures in the optical spectrum, the data from SDSS may have obscured some of the differences between the quasars and the stars. Looking at the results from

experiment 2 however, not including SDSS did not result in better clustering, and it actually resulted in the previous pure cluster of stars from the first experiment to be contaminated (see appendix).

For experiment 3, I then wondered if the ratio of stars to quasars were problematic. When people [13, 16] use t-SNE as, essentially, an n-ary classifier, the different categories usually have a roughly equal amount of members. This is certainly not always the case. In Jespersen's [8] paper, the L-type GRBs are much more numerous in his dataset, yet he still achieves some nice clustering. With that in mind I decided to see if changing the relative amount of stars and quasars to a roughly 50-50 split would yield a better result. I used the pandas library's sample method to randomly select the quasars as well as reshuffle the datapoints in order to minimize selection bias. This did unfortunately not yield better results. Since there was a roughly equal amount of stars and quasars, one would expect there to be two clusters of roughly even size; this was not case.

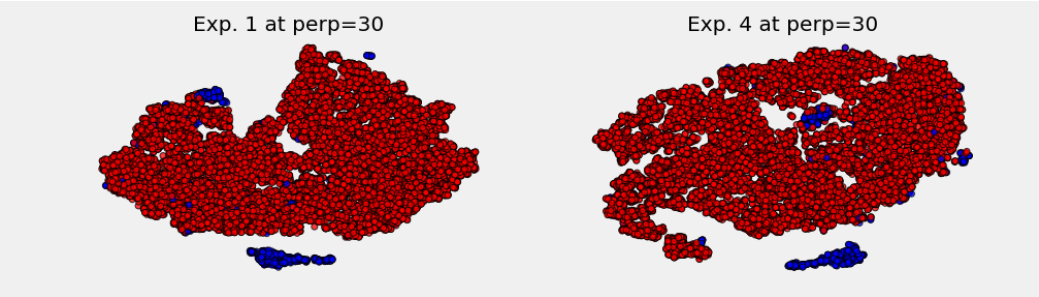


Figure 6: A comparison between the results from experiment 1 and 4 at perplexity 30. Quasars are red, and stars are blue.

I then decided to exclude the data channels u from SDSS and $W3$, $W4$ from the WISE satellite for experiment 4. These channels measure respectively in the ultra-violet and mid infrared regions, which are the extreme ends of the spectra. Additionally, these measurements also have large errors. I therefore ran an experiment without these data channels. This ended up not significantly changing the results from the first experiment though. In figure 6, you can see that both have a small cluster of pure stars, and a larger cluster consisting mainly of quasars but contaminated with stars.

Finally, for experiment 5 I decided to set different limits for the Gaia magnitude (the original catalog had used Gaia magnitude less than 21 as a selection criteria). This experiment was a little different from the others, since in addition to varying the perplexity I was also varying the

Gaia magnitude cut-off (though not at the same time). Good

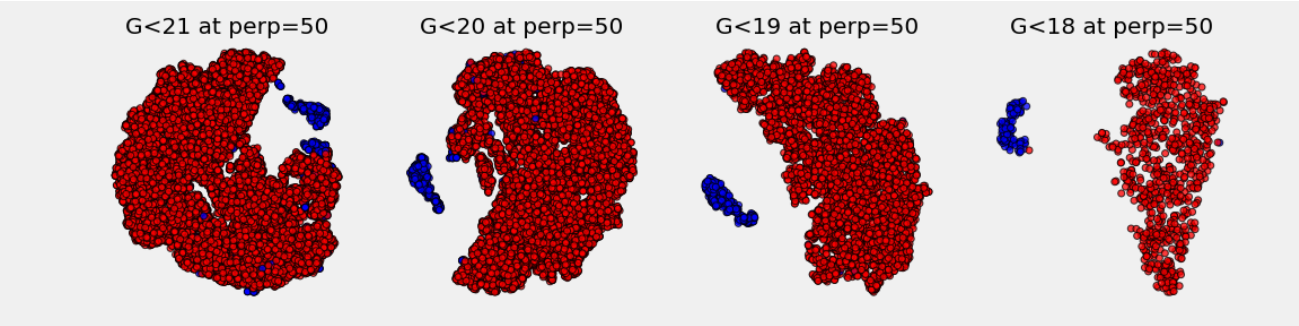


Figure 7: A comparison between the results from experiment 5 at perplexity 50. Quasars are red, and stars are blue.

As the cut-off for the Gaia magnitude decreases, the cluster appear more well separated and with lower levels of contamination, though that is somewhat difficult to discern by just looking at figure 7. One should not in haste conclude anything definitive based on the distances between the clusters, since they mean nothing [16]. These results do however suggest that the Gaia magnitude is important.

6 Discussion

Though these results are modest, several minor conclusion can be drawn, and there are several lines of inquiry moving forward.

Looking at all the results, the main issue at the moment seems to be star contamination in the large cluster. I think the results from experiment 5 provide a strong hint. Throughout this project, the data used has been the apparent magnitude of different point sources in several passbands. This is undoubtedly quick and easy, but it may be insufficient. Jespersen in his paper [8], does normalize his data in preparation for t-SNE. The Gaia magnitudes of the datapoints are mainly between 21 and 17, which means that, since the magnitude scale is logarithmic, the brightest objects are about $100^{\frac{4}{5}} \approx 40$ times brighter than the dimmest objects. I hypothesize that the reason the data appeared to be more well separated and less contaminated for lower G cut-offs is because the variance in magnitude decreases. This is of course just speculation on my part, but I think it would be interesting to either redo the experiment on photometric data that has been normalized, or on photometric data that has been subdivided into different magnitude classes.

Another similar point one could bring up would be redshift. The justification for this experiment is the fact that quasars and stars have different physical origins behind their spectra, maybe not accounting for redshift biases the data in a way such that the t-SNE algorithm doesn't work. However, if that were the case, that would immediately invalidate this whole approach. Quasars are the nuclei of galaxies far far away [11][Chapter 4.3], which means that they have significant redshift. On the other hand, the stars we observe are inside of the Milky Way galaxy, and thus have significantly smaller redshifts (or even blueshift if they're moving towards us) than quasars. Hence, if we know the datapoints redshifts, we already know whether they are stars or quasars (provided that's a valid dichotomy). Therefore, if knowing a datapoint's redshift is a prerequisite for this method to work, it is useless. It might still be interesting to explore whether correcting for redshift changes the result, but if it be necessary, the practical applications will be limited.

It would be great if the contamination could be entirely eliminated, though t-SNE could still prove useful even if that cannot be achieved. One intriguing line of inquiry would be to check whether it is the same datapoints that become contaminants, or if it be new every time. t-SNE is a stochastic algorithm, so it would be out of the question that any given datapoint which was a contaminant in one run would not be a contaminant in another. If this be the case, it would be possible to run t-SNE multiple times and then ascribe a confidence number to every datapoint based on how many times it was mapped into either category.

Another approach could be trying different passband filters. I have already done this to an extend. In experiment 2 I exclude the SDSS data channels which seemed to yield slightly worse results. In experiment 4 I excluded the "extreme" data channels, which seemingly did nothing. It is of course possible that more data could yield a better result, though based on my experiments, I am not optimistic that this would be the case.

Perplexity gets a lot of attention because it always has to be manually set when using the t-SNE algorithm, but it isn't the only hyperparameter. In addition to perplexity there are three other hyperparameters [13, 16, 10]:

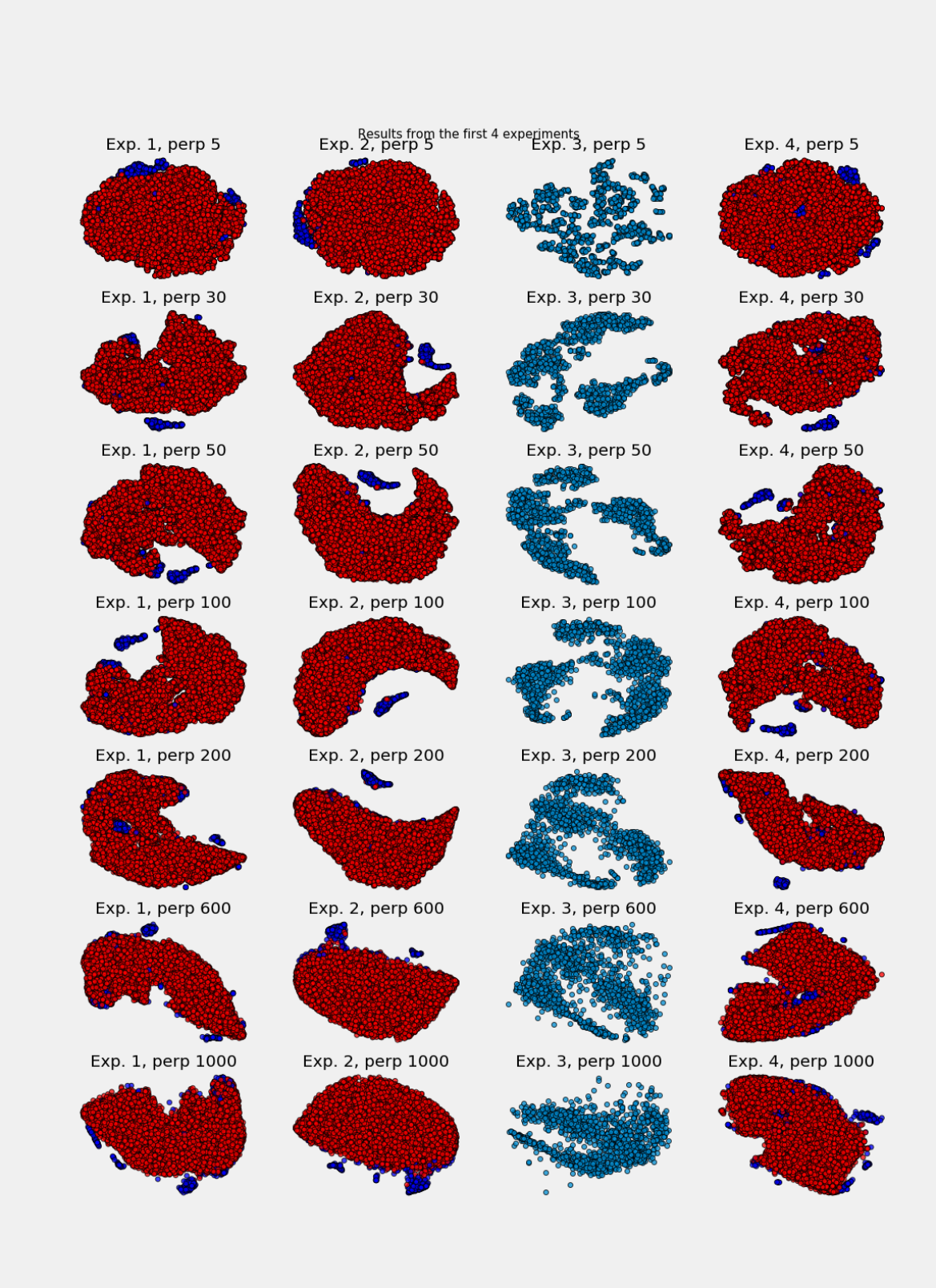
- Number of iterations.
- Learning rate.
- Early exaggeration.

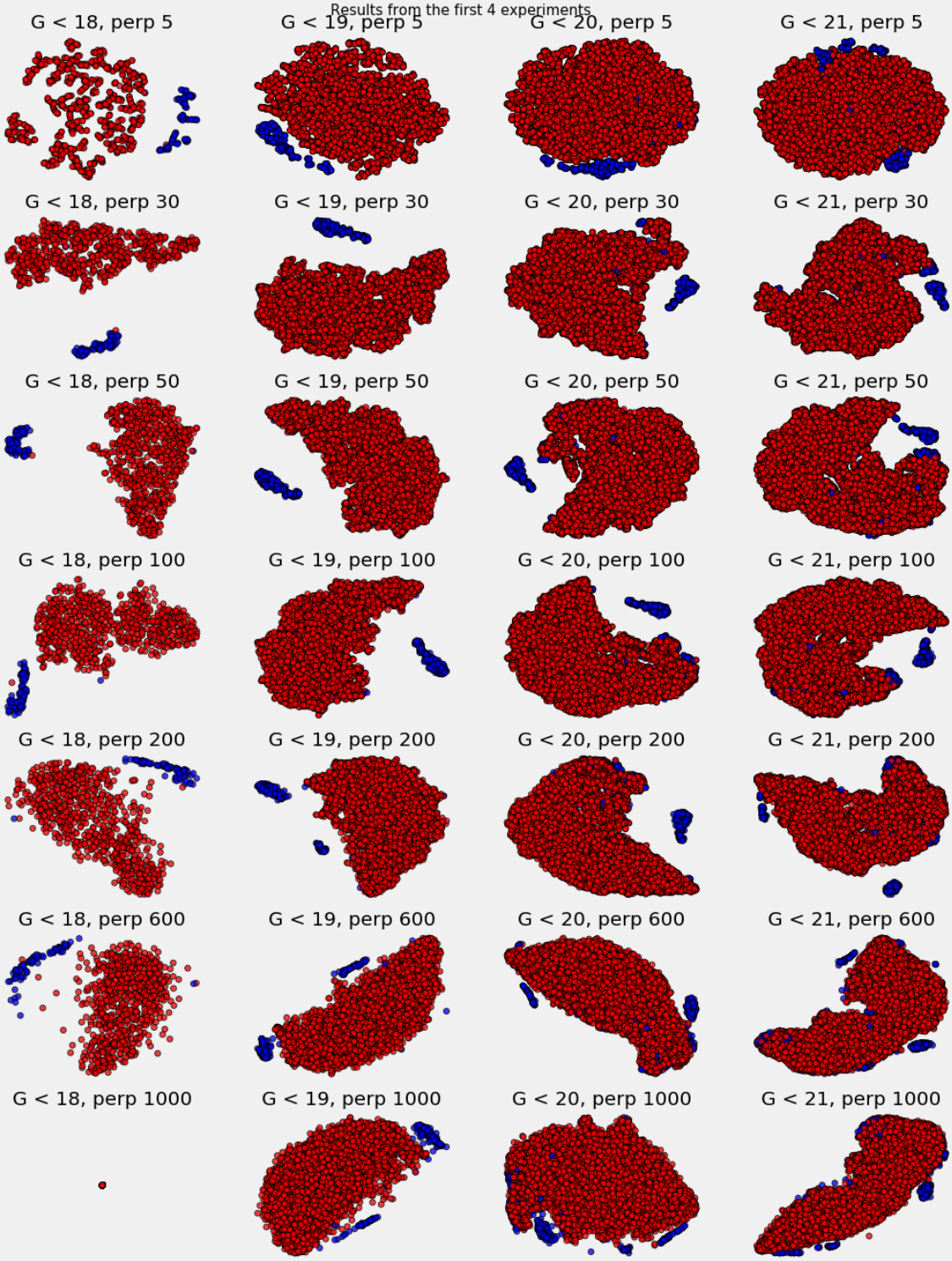
These hyperparameters have default values respectively of 1000, 200, and 12 in the scikit-learn implementation [10], which are the values that have been used in this thesis. I bring these up for the sake of being exhaustive with the possible improvements one could make, though I don't believe fiddling with these would change much. This data did separate, however, if one used a significantly larger dataset by either looking at a larger section of the sky, or at a more populated section, changing the maximum number of iteration might be prudent.

In conclusion, these results do suggest that t-SNE might be useful in quasar selection, though a more definitive conclusion demands further research. I suggest that the next step taken is controlling for variance in the Gaia magnitude by either normalizing the data using G, or subdividing the data into distinct magnitude classes.

Very short, but good.

7 Appendix


Results from the first 4 experiments
Exp. 1, perp 5
Exp. 2, perp 5
Exp. 3, perp 5
Exp. 4, perp 5
Exp. 1, perp 30
Exp. 2, perp 30
Exp. 3, perp 30
Exp. 4, perp 30
Exp. 1, perp 50
Exp. 2, perp 50
Exp. 3, perp 50
Exp. 4, perp 50
Exp. 1, perp 100
Exp. 2, perp 100
Exp. 3, perp 100
Exp. 4, perp 100
Exp. 1, perp 200
Exp. 2, perp 200
Exp. 3, perp 200
Exp. 4, perp 200
Exp. 1, perp 600
Exp. 2, perp 600
Exp. 3, perp 600
Exp. 4, perp 600
Exp. 1, perp 1000
Exp. 2, perp 1000
Exp. 3, perp 1000
Exp. 4, perp 1000


Results from the first 4 experiments
G < 18, perp 5
G < 19, perp 5
G < 20, perp 5
G < 21, perp 5
G < 18, perp 30
G < 19, perp 30
G < 20, perp 30
G < 21, perp 30
G < 18, perp 50
G < 19, perp 50
G < 20, perp 50
G < 21, perp 50
G < 18, perp 100
G < 19, perp 100
G < 20, perp 100
G < 21, perp 100
G < 18, perp 200
G < 19, perp 200
G < 20, perp 200
G < 21, perp 200
G < 18, perp 600
G < 19, perp 600
G < 20, perp 600
G < 21, perp 600
G < 18, perp 1000
G < 19, perp 1000
G < 20, perp 1000
G < 21, perp 1000

References

[1] Giuseppe Bonaccorso. *Mastering Machine Learning Algorithms*. Packt Publishing, 2018. ISBN: 978-1-78862-111-3.

[2] A. G. A. Brown et al. "Gaia Data Release 2. Summary of the contents and survey properties". In: *Astronomy & Astrophysics* 616, A1 (2018), A1. DOI: 10.1051/0004-6361/201833051. arXiv: 1804.09365 [astro-ph.GA].

[3] Daniel J. Eisenstein et al. "SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way, and Extra-Solar Planetary Systems". In: *The Astronomical Journal* 142.3, 72 (2011), p. 72. DOI: 10.1088/0004-6256/142/3/72. arXiv: 1101.1529 [astro-ph.IM].

[4] Tyler Folkman. *Why You Are Using t-SNE Wrong*. 2019. URL: https://towardsdatascience.com/why-you-are-using-t-sne-wrong-502412aab0c0.

[5] K. E. Heintz, J. P. U. Fynbo, and E. Høg. "A study of purely astrometric selection of extragalactic point sources with Gaia". In: *Astronomy & Astrophysics* 578, A91 (2015), A91. DOI: 10.1051/0004-6361/201526038. arXiv: 1503.02874 [astro-ph.HE].

[6] K. E. Heintz et al. "Spectroscopic classification of a complete sample of astrometrically-selected quasar candidates using Gaia DR2". In: *Astronomy & Astrophysics* 644, A17 (2020), A17. DOI: 10.1051/0004-6361/202039262. arXiv: 2010.05934 [astro-ph.GA].

[7] K. E. Heintz et al. "Unidentified quasars among stationary objects from Gaia DR2". In: *Astronomy & Astrophysics* 615, L8 (2018), p. L8. DOI: 10.1051/0004-6361/201833396. arXiv: 1805.03394 [astro-ph.GA].

[8] Christian K. Jespersen et al. "An Unambiguous Separation of Gamma-Ray Bursts into Two Classes from Prompt Emission Alone". In: *The Astrophysical Journal Letters* (2020). DOI: 10.3847/2041-8213/ab964d.

[9] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[10] F. Pedregosa et al. *sklearn.manifold.TSNE Documentation*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html (visited on 01/18/2021).

[11] L. S. Sparke and J. S. III Gallagher. *Galaxies in the Universe: An Introduction*. Cambridge University Press, 2007. ISBN: 978-0-521-67186-6.

[12] Laurens van der Maaten. "Accelerating t-SNE using Tree-Based Algorithms". In: *Journal of Machine Learning Research* (2014).

[13] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* (2008).

[14] S. J. Warren, P. C. Hewett, and C. B. Foltz. "The KX method for producing K-band flux-limited samples of quasars". In: *Monthly Notices of the Royal Astronomical Society* 312.4 (2000), pp. 827–832. DOI: 10.1046/j.1365-8711.2000.03206.x. arXiv: astro-ph/9911064 [astro-ph].

[15] S. J. Warren et al. "The United Kingdom Infrared Telescope Infrared Deep Sky Survey First Data Release". In: *Monthly Notices of the Royal Astronomical Society* 375.1 (2007), pp. 213–226. DOI: 10.1111/j.1365-2966.2006.11284.x. arXiv: astro-ph/0610191 [astro-ph].

[16] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. "How to Use t-SNE Effectively". In: *Distill* (2016). DOI: 10.23915/distill.00002. URL: http://distill.pub/2016/misread-tsne.

[17] Edward L. Wright et al. "The Wide-field Infrared Survey Explorer (WISE): Mission Description and Initial On-orbit Performance". In: *The Astronomical Journal* 140.6 (2010), pp. 1868–1881. DOI: 10.1088/0004-6256/140/6/1868. arXiv: 1008.0031 [astro-ph.IM].