PHSX815_Project4:
Exoplanet Populations

Yoni Brande

May 1, 2021

1 Abstract

We use clustering algorithms to identify subgroups of exoplanet populations, and try to determine whether our clusters are physical or artifacts from observational biases.

2 Introduction

Over 4000 exoplanets have been found since the first planets outside our solar system were discovered in the early 1990s [1] [2]. Of the 4383 confirmed planets, 3354 have been discovered by the transit method, observing periodic characteristic dips in the brightness of a host star as the planet moves in front of the star's disk in our line of sight. These discoveries have mostly been driven by space-based transit surveys such as the wildly successful Kepler, K2, and TESS missions [3] [4] [5], as well as ground-based surveys such as MEarth, KELT, and HAT [6] [7] [8]. Radial velocity surveys have discovered the next largest sample of exoplanets, and other methods have discovered the rest.

Now that we are in an era of statistically useful exoplanet samples, we are able to do population-level studies of exoplanet properties. For example, one of the first unexpected populations noticed were the Hot Jupiters, large gaseous exoplanets orbiting very close to their host stars. This challenged conventional wisdom on planet formation and evolution, and gave us a more complete understanding of the astrophysical processes taking place in planetary systems. Unfortunately, our observational capabilities are limited, and we can't observe every planet (or small body) that possibly exists. This leads to a fundamental conflict in our population studies: Are we observing a truly representative sample, and are the trends we exist physical? Or are we just looking at the effects of observational biases? Here we attempt to present exoplanet population data in ways that allow us to find real clusters of similar exoplanets.

3 Exoplanet Data

Different exoplanet discovery methods rely on different ways to observe planets either directly or indirectly. The transit method measures periodic dips in the brightness of a star as the planet transit's the star's disk, giving us a measurement of the relative areas of the planet's and star's disk. By knowing the size of a star, we easily find the size of the planet. The period of the signal gives us the size of the planet's orbit, and other orbital parameters can be determined by more subtle aspects of the transit signal. However, this gives us no information on the planet's mass, so we look to other detection methods, like the Radial Velocity method. Since a planet and its host star both orbit a common barycenter, we can look for the effects of the motion of the star around that

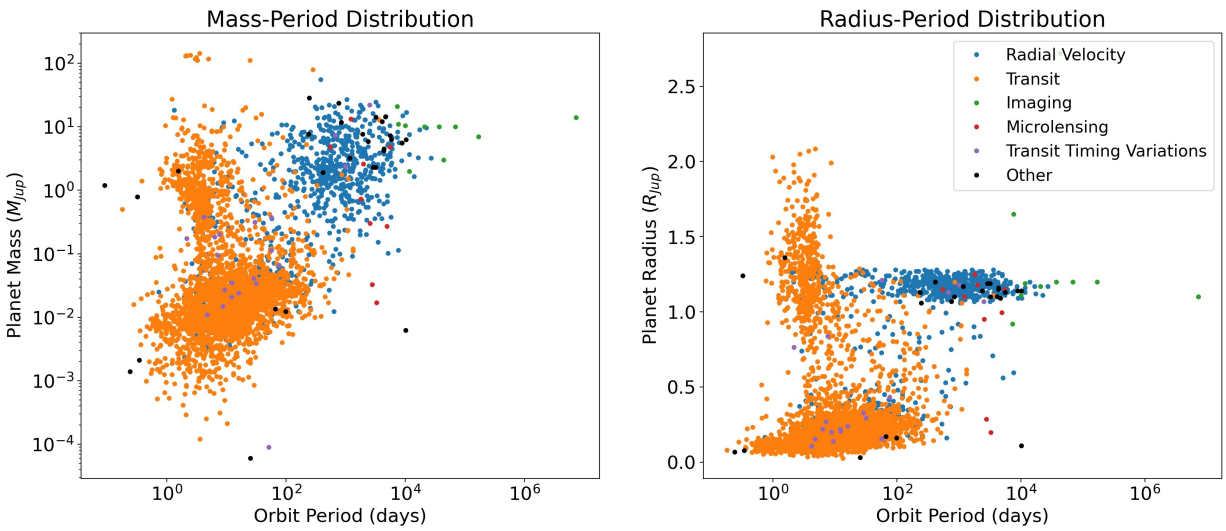

Figure 1: Confirmed Exoplanet Mass-Period/Radius-Period distributions

barycenter. As the star moves towards and away from us along our line of sight, we can observe the red- and blueshifting of the star's spectrum, and use the strength of the Doppler signal to determine the mass of the planet inducing it. Both transits and radial velocities are most sensitive to large planets on short orbits, and since transiting planets all have inclinations near 90°, many of them are also amenable to radial velocity observations, giving us both mass and radius (and thus density and bulk composition) for many observed exoplanets.

Generally, when presenting data for most exoplanets, we prefer to rely on radius, mass, and orbital period as the basic parameters. Mass-period and radius-period distributions are common presentations, and mass-radius distributions allow for curves of constant density to be plotted. By coloring each planet with its discovery method, we see some natural clusters appear.

In the mass-period plot (left of Fig. 1), we see that there are two distinct clusters of transit and radial velocity planets, plus possibly a third cluster of mixed detection methods. In the radius-period plot (right of Fig. 1), we start to see the effects of some other planetary models. While most transiting planets are amenable to RV study, giving us their exact masses, many RV discovered planets are not known to transit, and as such we can only estimate their radii from planetary mass-radius relations. This leads to a very tight, visible cluster.

4 Clustering Algorithms

Given a multidimensional set of data, we can use clustering algorithms to find subsets of data points that are all more similar to each other than they are to data outside their clusters. Here, we adopt the use of k-means clustering, a centroid-based method that assigns points to one of K possible clusters given K starting points, and iteratively reassigns points as the cluster centers move to some steady state. This minimizes the variance within each cluster, and when the cluster centers no longer move, the solution has converged to the minimum-variance solution. Of course, not every possible number of clusters will be useful, as for a dataset with N points, you might expect to be able to assign the dataset to any number of clusters in the range $[1, N]$. Some care will be needed to not overfit the data with too many possible clusters, but we must also be careful not to underfit the data, as too few clusters will also be uninformative.

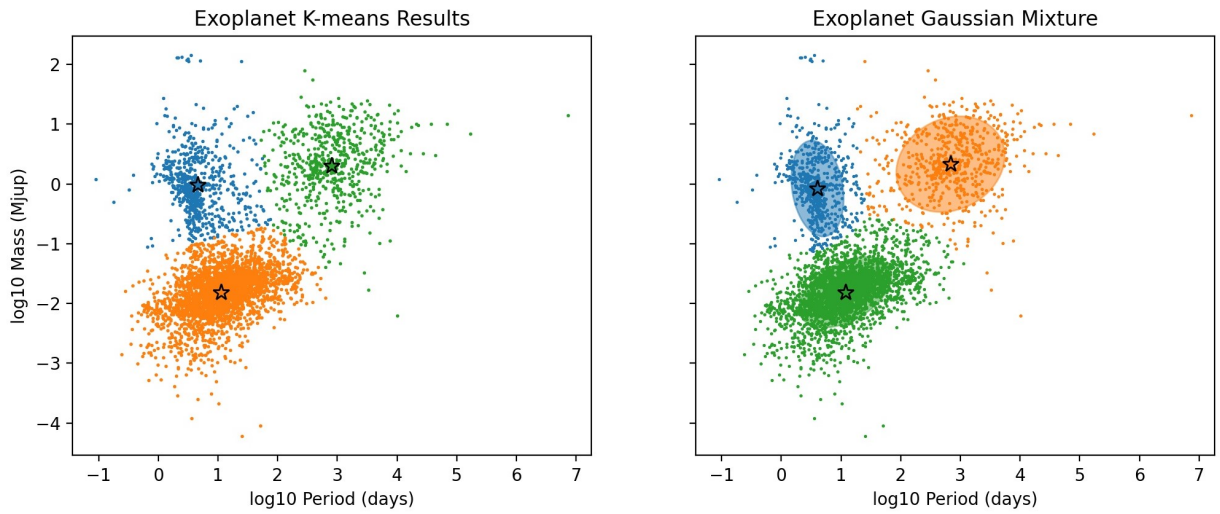

Figure 2: k-means vs GMM clustering for the exoplanet mass-period data

For data where it may be expected that the populations will have some overlap, we may also use expectation maximization to assign points to one or several overlapping Gaussian distributions. This Gaussian mixture method may be more versatile than K-means as it can account for points that somewhat fit both of two separate distributions, like a trough in a bimodal distribution, rather than placing a hard border and assigning them to one or the other cluster. Each cluster then becomes a Gaussian function with its centers and variances, and as the algorithm iterates, those centers and variances should converge to some maximum expectation value. We will compare the use of these two methods to sort exoplanets into varying populations.

5 Results

We ran both k-means clustering and expectation maximization for the mass-period and radius-period planet data. For the mass-period runs, both k-means and GMM/EM converged nicely. These results are shown in Fig. 2, with stars indicating the cluster centers, and the ellipses indicating the $\pm 1\sigma$ bounds for the Gaussians. Three major clusters were found, and the k-means centers identified were very close to the GMM/EM centers, easily within the 1σ ellipses.

However, we start to see the limitations of k-means clustering with the radius-period data. Although by eye we might convince ourselves that there are three main clusters, the k-means algorithm struggles to identify them. There are a lot of data points between the main cluster cores, and these serve to drag the k-means centers away from where they might ought to be. By increasing the number of desired clusters to 4, we now get good identification of the top two, but the bottom cluster gets split in half. Luckily, GMM steps in to help. By using this method, we can fit for all three major clusters and a fourth can start to incorporate the interstitial data points.

So, after successfully identifying some clusters of exoplanets, we move to interpreting what these might mean. In the mass-period data, we found three distinct clusters. By overplotting these with the discovery methods of these planets (see Fig. 4), we see that two clusters correspond to the majority of the transiting planets, and one corresponds to the majority of the wide orbit radial velocity planets. Each transit cluster seems to coincide with some of the well-known planet populations. The top cluster looks to be hot gaseous planets, similar in mass to Jupiter and orbiting

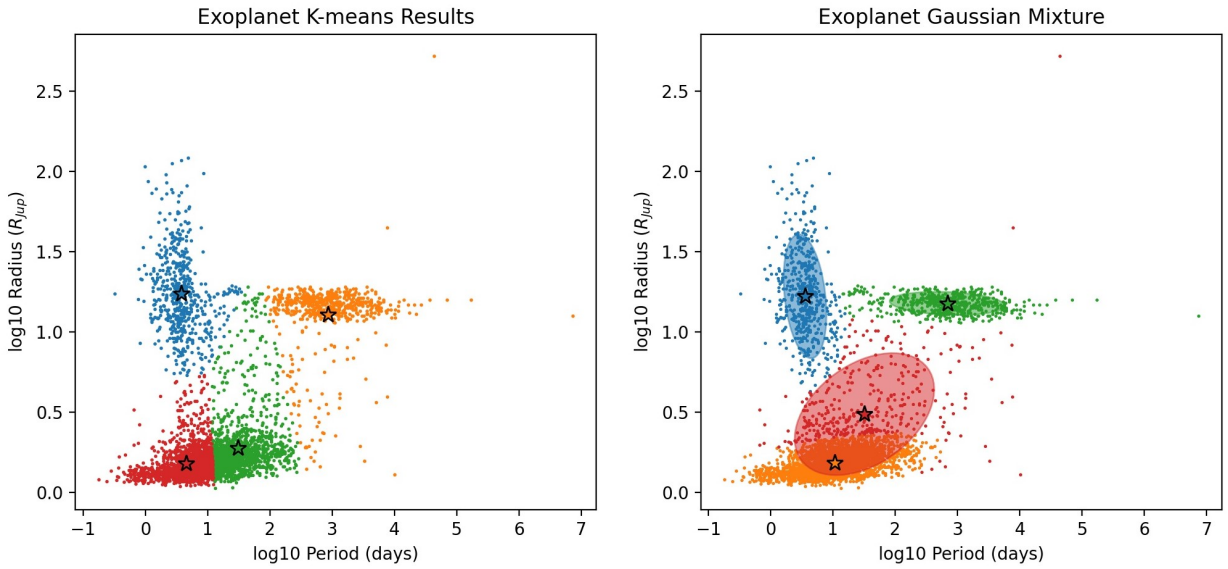

Figure 3: k-means vs GMM clustering for the exoplanet radius-period data

with periods of a few days. The lower cluster is broader, but includes the known exo-Neptunes and terrestrial-mass planets. We can also see that this cluster contains planets on both very short and medium orbits. The radial velocity cluster contains many long-period and high-mass planets. In general this all seems to be expected. Transit surveys are most sensitive to short-period planets at all masses, and radial velocity surveys are most sensitive to high-mass planets, which are less likely to form on or migrate to short orbits.

Looking at the clusters for the radius-period data, we see similar results, easily identifying the Hot Jupiters, the medium-sized transiting planets, and the radial velocity sample. The intersitial cluster, containing some of the data between the main clusters, is a bit dubious. It may be that there are two clusters, one terrestrial sized, and one gas dwarf sized, with the extra planets just being artifacts of the diversity of observed exoplanetary systems. It is important to note here that the radial velocity cluster does not get much larger than $\sim 1.3 M_{Jup}$, even though these planets may have masses near $10 M_{Jup}$ due to the effects of gravitational self compression at super-Jupiter masses (see: e.g. [9] for a common mass-radius relation). Objects that get much larger than these are likely to be Brown Dwarfs, not exoplanets, and as such are not included in these data. However, we see many super-Jupiter radius planets at shorter periods, due to heating and atmospheric inflation. It seems likely that the true radii of these planets more gradually blends into the ∼Jupiter radius sample, but we do not currently have the transit data to prove this.

We do, however, see some notable voids. In the mass-period and radius-period data, there is a hint of a gap, just smaller than Jupiter and closer orbiting than the main body of transiting planets, just to the left of where the transit clusters meet. This is the known Neptunian Desert, an observed paucity of Neptune sized planets on short orbits, theorized to be an effect of either photoevaporative atmospheric loss [10], or a result of the formation and migration of gaseous exoplanets [11]. We also note that there appears a lack of intermediate mass and radius planets on longer orbits. Although we have good coverage of the large mass, long orbit RV sample, and good coverage of transit data out to a few hundred day orbits, we only have a sparse sampling of the low-mass, long-orbit parameter space. This makes sense from formation theory, as past the snow line we expect more giant planets to form as they are able to accrete lighter gases and condensates. However, we also would not expect

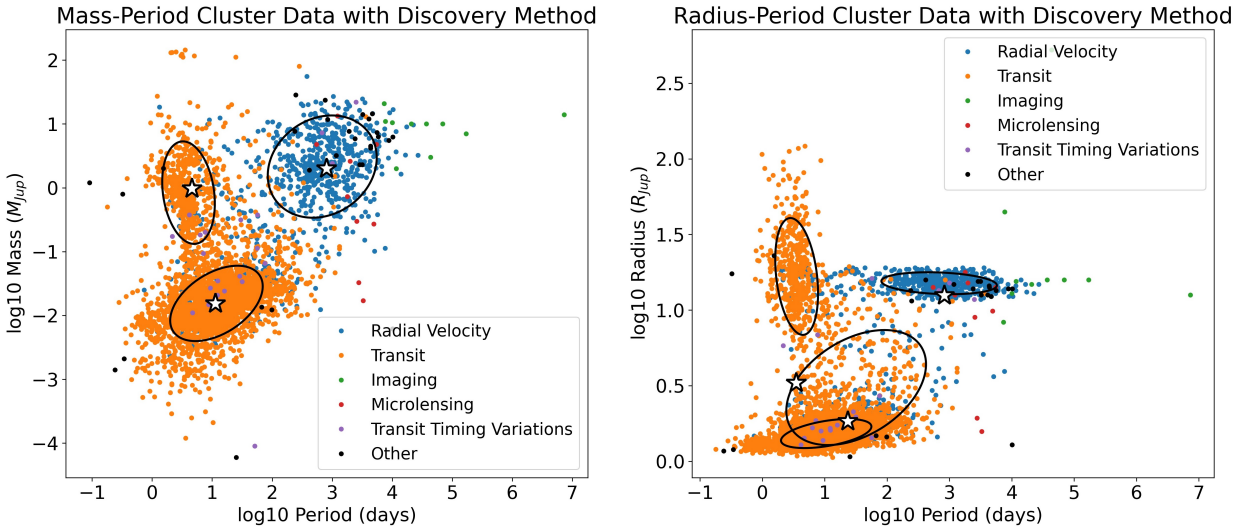

Figure 4: Clusters shown with exoplanet discovery methods

to be able to observe very low mass, long period planets as these would be less likely to transit, and require much higher sensitivity and longer baseline RV surveys. Surveys of this sort are only recently coming online, and as such have not had the time to search for planets such as these yet.

6 Conclusions

We find that clustering algorithms can be useful tools to help classify exoplanetary data. k-means clustering can be productive when appropriately applied, but will break down with particularly complex mixed datasets. Gaussian mixture modeling can overcome some of the limitations of k-means, and works much better in those circumstances.

References

[1] A. Wolszczan and D. A. Frail, *A planetary system around the millisecond pulsar PSR1257 + 12*, **355** no. 6356, (Jan., 1992) 145–147.

[2] M. Mayor and D. Queloz, *A Jupiter-mass companion to a solar-type star*, **378** no. 6555, (Nov., 1995) 355–359.

[3] W. J. Borucki, D. Koch, G. Basri, T. Brown, D. Caldwell, E. Devore, E. Dunham, T. Gautier, J. Geary, R. Gilliland, A. Gould, S. Howell, and J. Jenkins, *The Kepler Mission: A Transit-Photometry Mission to Discover Terrestrial Planets*, *ISSI Scientific Reports Series* **6** (Jan., 2006) 207–220.

[4] S. B. Howell, C. Sobeck, M. Haas, M. Still, T. Barclay, F. Mullally, J. Troeltzsch, S. Aigrain, S. T. Bryson, D. Caldwell, W. J. Chaplin, W. D. Cochran, D. Huber, G. W. Marcy, A. Miglio, J. R. Najita, M. Smith, J. D. Twicken, and J. J. Fortney, *The K2 Mission: Characterization and Early Results*, **126** no. 938, (Apr., 2014) 398, `arXiv:1402.5163 [astro-ph.IM]`.

[5] G. R. Ricker, J. N. Winn, R. Vanderspek, D. W. Latham, G. Á. Bakos, J. L. Bean, Z. K. Berta-Thompson, T. M. Brown, L. Buchhave, N. R. Butler, R. P. Butler, W. J. Chaplin,

D. Charbonneau, J. Christensen-Dalsgaard, M. Clampin, D. Deming, J. Doty, N. De Lee, C. Dressing, E. W. Dunham, M. Endl, F. Fressin, J. Ge, T. Henning, M. J. Holman, A. W. Howard, S. Ida, J. M. Jenkins, G. Jernigan, J. A. Johnson, L. Kaltenegger, N. Kawai, H. Kjeldsen, G. Laughlin, A. M. Levine, D. Lin, J. J. Lissauer, P. MacQueen, G. Marcy, P. R. McCullough, T. D. Morton, N. Narita, M. Paegert, E. Palle, F. Pepe, J. Pepper, A. Quirrenbach, S. A. Rinehart, D. Sasselov, B. Sato, S. Seager, A. Sozzetti, K. G. Stassun, P. Sullivan, A. Szentgyorgyi, G. Torres, S. Udry, and J. Villasenor, *Transiting Exoplanet Survey Satellite (TESS)*, *Journal of Astronomical Telescopes, Instruments, and Systems* **1** (Jan., 2015) 014003.

[6] P. Nutzman and D. Charbonneau, *Design Considerations for a Ground-Based Transit Search for Habitable Planets Orbiting M Dwarfs*, **120** no. 865, (Mar., 2008) 317, arXiv:0709.2879 [astro-ph].

[7] J. Pepper, R. W. Pogge, D. L. DePoy, J. L. Marshall, K. Z. Stanek, A. M. Stutz, S. Poindexter, R. Siverd, T. P. O'Brien, M. Trueblood, and P. Trueblood, *The Kilodegree Extremely Little Telescope (KELT): A Small Robotic Telescope for Large-Area Synoptic Surveys*, **119** no. 858, (Aug., 2007) 923–935, arXiv:0704.0460 [astro-ph].

[8] G. Bakos, R. W. Noyes, G. Kovács, K. Z. Stanek, D. D. Sasselov, and I. Domsa, *Wide-Field Millimagnitude Photometry with the HAT: A Tool for Extrasolar Planet Detection*, **116** no. 817, (Mar., 2004) 266–277, arXiv:astro-ph/0401219 [astro-ph].

[9] J. Chen and D. Kipping, *Probabilistic Forecasting of the Masses and Radii of Other Worlds*, **834** no. 1, (Jan., 2017) 17, arXiv:1603.08614 [astro-ph.EP].

[10] T. Mazeh, T. Holczer, and S. Faigler, *Dearth of short-period Neptunian exoplanets: A desert in period-mass and period-radius planes*, **589** (May, 2016) A75, arXiv:1602.07843 [astro-ph.EP].

[11] J. E. Owen and D. Lai, *Photoevaporation and high-eccentricity migration created the sub-Jovian desert*, **479** no. 4, (Oct., 2018) 5012–5021, arXiv:1807.00012 [astro-ph.EP].