CLASSIFYING EXOPLANETS WITH GAUSSIAN MIXTURE MODEL

SOHAM KULKARNI AND SHANTANU DESAI
Dept. of Physics, IIT Hyderabad, Kandi, Telangana 502285, India

ABSTRACT

Recently, Odrzywolek and Rafelski [1] have found three distinct categories of exoplanets, when they are classified based on density. We first carry out a similar classification of exoplanets according to their density using the Gaussian Mixture Model, followed by information theoretic criterion (AIC and BIC) to determine the optimum number of components. Such a one-dimensional classification favors two components using AIC and three using BIC, but the statistical significance from both the tests is not significant enough to decisively pick the best model between two and three components. We then extend this GMM-based classification to two dimensions by using both the density and the Earth similarity index [2], which is a measure of how similar each planet is compared to the Earth. For this two-dimensional classification, both AIC and BIC provide decisive evidence in favor of three components.

1. INTRODUCTION

Over the past two decades, there has been a revolution in the field of exoplanet astronomy following the confirmation of more than 3000 planets orbiting stars other than the Sun (see Ref. [3] for a recent review). Lot of work has been done to characterize the properties of the exoplanets discovered using the myriad techniques [4]. Recently, Odrzywolek and Rafelski [1] (hereafter OR16) have carried out the classification of exoplanets according to their density, following a suggestion long time back [5]. OR16 fitted the exoplanet density data to lognormal distributions to determine the optimum number of components. They found three lognormal components with peak densities at 0.71 gm/cm^3, 6.9 gm/cm^3, and 29.1 gm/cm^3 [1]. These three components correspond to ice/gas giants, iron/rock super-Earths, and brown dwarfs respectively. The optimum number of components was determined by maximizing the log-likelihood and then checking the goodness of fit for different number of components by calculating the p-value from three distinct non-parametric tests.

We would like to do a variant of the above analysis by carrying out a similar classification according to density using Gaussian mixture models, followed by information theoretic criterion to determine the optimum number of classes. We have previously used this procedure, to perform a unified classification of all GRB datasets using three different model comparison techniques [6]. We then extend the analysis of OR16 by considering two-dimensional classification using both the density and Earth similarity index.

This paper is organized as follows. In Section 2, we describe the dataset and the physical quantities used for the classification. The mathematical basis for the classification is discussed in Section 3. Our results are shown in Section 4 and we conclude in Section 5.

2. DATA

2.1. *Exoplanet Catalog*

In the one-dimensional classification, we shall study the trends obtained from the confirmed detections, by classifying the exoplanet database according to their densities, for which we need the mass and radius of the planets. We obtain the mass and radius information from the NASA Exoplanet archive[1] and the Extrasolar planet encyclopedia[2] and consider only those planets with measured values of mass and density, and which exist in both the datasets with the same observed values to avoid any irregularities and to maintain consistency in the dataset. The NASA Exoplanet archive is a NASA funded public data service, which is hosted by the Infrared Processing and Analysis Center. This catalog lists only those objects, for which their detection and planetary status is sacrosanct. It currently contains a total of 3440 planets out of which 531 have measured mass and radius values detected. The Extrasolar planet encyclopedia is maintained by the Meudon Observatory in Paris and has a total of 3567 planets, of which 615 have measured values for all the parameters. The data provided by the two catalogs is similar except for some differences in their selection criteria. The Extrasolar planet encyclopedia allows planets weighing from 60 Jupiter Mass onwards, whereas the NASA Exoplanet archive uses 30 Jupiter mass as the lower limit, which is also the reason for the smaller number of confirmed exoplanets in the latter. We have selected 450 observations, which are common to both the datasets for our study in this paper.

In addition to the one-dimensional classification using only density, we also carry out a two-dimensional classification, wherein we use both the density and the Earth Similarity Index (or ESI) [2] for the classification. For this, we need

ep13b1003@iith.ac.in
shntn05@gmail.com
[1] https://exoplanetarchive.ipac.caltech.edu/
[2] http://exoplanet.eu

some additional parameters for the calculation of ESI. The additional parameters that we need apart from the radius and density are the surface temperature and time period of revolution, as other parameters can be derived from the mass and radius. The escape velocity and surface gravity are calculated by positing that the shape of the planet is a perfect sphere, wherein the total mass is distributed uniformly throughout the volume. We only consider planets for which we have the observed values for all four of these parameters.

2.2. *Calculations for the data:*

Assuming the planet is a perfect sphere with a uniform mass distribution, the expression for density is:

$$\bar{\rho} = \frac{M}{\frac{4}{3}\pi R^3} \tag{1}$$

The escape velocity is given by:

$$v_{esc} = \sqrt{\frac{2GM}{R}}. \tag{2}$$

The surface gravity is obtained from:

$$g_{surf} = \frac{GM}{R^2}. \tag{3}$$

where G is the Gravitational Constant, M is the mass of the planet and R is the radius.

ESI is a figure of merit used to ascertain how habitable is the planet for life to develop compared to the Earth. More details on the theory behind ESI can be found in the work by Kashyap [2], which in turn follows the prescription from Schulze-Makuch et al. [7] (See also [8] for alternate indexes proposed similar in spirit to ESI). The ESI is based on six different parameters, viz. density, radius, temperature, surface gravity, escape velocity, and the time period of revolution around their Sun. All these parameters are normalized to Earth units, as it is convenient for the index calculation. The ESI is calculated based on the Bray-Curtis Similarity index [9] and is given by:

$$ESI_x = \left(1 - \left|\frac{x - x_0}{x + x_0}\right|\right)^w \tag{4}$$

where x is the parameter for which the index has to be calculated, x_0 is the reference values which in our case is one, as we have expressed all parameters in Earth units and w is the weight exponent.

The total ESI is given by:

$$ESI = (ESI_g \times ESI_{temp} \times ESI_{vesc} \times ESI_p \times ESI_r \times ESI_d)^{1/6} \tag{5}$$

The values of ESI range from 0 (completely different from Earth) to 1 (resembling a clone of Earth).

3. ANALYSIS METHODS:

We outline the method used for both the one-dimensional classification using density and the two-dimensional classification using density and ESI. For finding the best-fit parameters, we use the Gaussian-mixture Model (GMM) [10], which is part of the `Scikit-learn` package, used for a variety of machine learning applications in python. The GMM fits the data to a mixture of multiple (k) lognormal Gaussian distributions, which are characterized by their mean, covariance and their respective weights in the fit data. The GMM method uses the Expectation Maximization (EM) algorithm [11] to maximize the likelihood function over the given parameter space. The GMM method can also be generalized to include error bars and this generalized GMM algorithm is referred to in the astrophysics literature as Extreme Deconvolution [12]. However, since we are using a planet catalog measured in two separate datasets having negligible error bars, we stick to the ordinary GMM method. Given the probability distribution function $f(x, \theta)$, where x are the observed datapoints, θ are the parameters used to define the function, N being the total number of exoplanets in our study, and w_k denotes the weights associated with each of the k log normal distributions, the likelihood can be defined as:

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{k} w_j f_j(x_i, \theta), \tag{6}$$

and the probability distribution function for a univariate Gaussian as:

$$f(x, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \log \rho_{planet})^2}{2\sigma^2}\right). \tag{7}$$

A generalized bivariate Gaussian distribution can be defined as:

$$f(x,\theta) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho'^2}}\exp\left[-\frac{1}{2(1-\rho'^2)}\left(\frac{(x-\mu_x)^2}{\sigma_1^2} + \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_1\sigma_2} + \frac{(y-\mu_y)^2}{\sigma_2^2}\right)\right] \tag{8}$$

where $\rho' = \frac{V_{12}}{\sigma_1\sigma_2}$ is the correlation, V is the covariance of the two variables and μ_x is the mean log(density) and μ_y is the mean ESI. An additional condition being used in the EM algorithm is the normalization condition:

$$\sum_{i=1}^{k} w_i = 1 \tag{9}$$

In this study, we use the GMM method for the $k=2$ and $k=3$ lognormal fits to the data followed by information theory based model comparison methods to assess the best fit amongst these two models.

3.1. Model Comparison

Once we have obtained the best-fit parameters for each model, we need to select the optimum model from all the possibilities being considered. Naively, the simplest way to do model comparison would be by carrying out likelihood comparison between the competing models and choosing the model with the highest likelihood as the best model. However, the maximization of likelihood could lead to an overfitting of the model to the data with additional degrees of freedom and hence we need a more robust and accurate criterion, which will penalize the use of extra free parameters. This can be done by using the Information criterion tests, such as Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC), which are commonly used in Astrophysics literature [13, 14, 15, 16, 17] (and references therein). These information criteria-based methods provide a way to penalize the excess free parameters and determine the best model accordingly.

3.1.1. AIC:

Akaike Information Criteria or AIC [18] penalizes lightly the excess free parameters and is defined as:

$$AIC = 2p + 2\ln L \tag{10}$$

where p is the number of free parameters in the model and L is the likelihood. The AIC defined in Eq. 10 is valid when the ratio N/p is very large i.e. > 40. For a ratio less than this, a first order correction is included and the modified expression is given by:

$$AIC = 2p + 2\ln L + \frac{2p(p+1)}{N-p-1} \tag{11}$$

Throughout our data, the ratio is greater than the value prescribed and hence we do not account for this correction in our study. The preferred model is the one with a lower value of AIC and the efficacy of this hypothesis is determined using the quantity:

$$\Delta AIC_i = AIC_i - AIC_{min}, \tag{12}$$

where ΔAIC_i value corresponds to the preference of the model i over the model with the lower AIC value and hence is the null hypothesis. The confidence in the model can be determined by the magnitude of the ΔAIC value. Although one cannot formally calculate p-values from ΔAIC, one usually uses qualitative strength of evidence rules to judge the efficacy of a given model [13, 17, 19]. As pointed out by Liddle [17], the value for the best model will be, $\Delta AIC_i = 0$. Now, if $0 < \Delta AIC_i < 2$, then we can say that we have a weak or no statistical evidence to reject the ith model over the null hypothesis. $2 < \Delta AIC_i < 6$ implies that the model has only weak support and has evidence against this model. For models with $\Delta AIC_i > 6$ there exists strong evidence against the model and $\Delta AIC_i > 10$ implies a very strong or decisive evidence against the ith model. These rules can be applied directly for the BIC criterion (next subsection) as well.

3.1.2. BIC:

Bayesian Information Criterion or BIC was used by Schwarz [20] and is used to penalize the free parameters much more harshly than the AIC criterion and is defined as:

$$BIC = p\ln N + 2\ln L \tag{13}$$

Again, the preferred model is the one with the lower values of BIC and is taken as the null hypothesis for further determining the significance of different models.

$$\Delta BIC_i = BIC_i - BIC_{min} \tag{14}$$

Similar to the significance test for the AIC criterion, the ΔBIC_i value acts as the significance measure for the BIC test and follows the same values as for AIC. The only difference being that according to BIC criterion, the penalty for a model with extra number of free parameters is harsher compared to AIC.

4. RESULTS:

4.1. *1D classification*

We first describe our results for the one-dimensional classification using only the density. We apply the techniques and methods described in the previous sections to the exoplanet catalog, generated by filtering the data from the NASA Exoplanet archive and the Extrasolar Planet encyclopedia as mentioned earlier. For the density function, we find the best-fit model parameters for k lognormal distributions according to the density from Eq. 6. Each distribution is characterized by its mean, standard deviation and the weight of the distribution indicating the number of planets that have been classified under that particular distribution. We apply the GMM routine to the density functions after varying the number of Gaussians from 1 to 14.
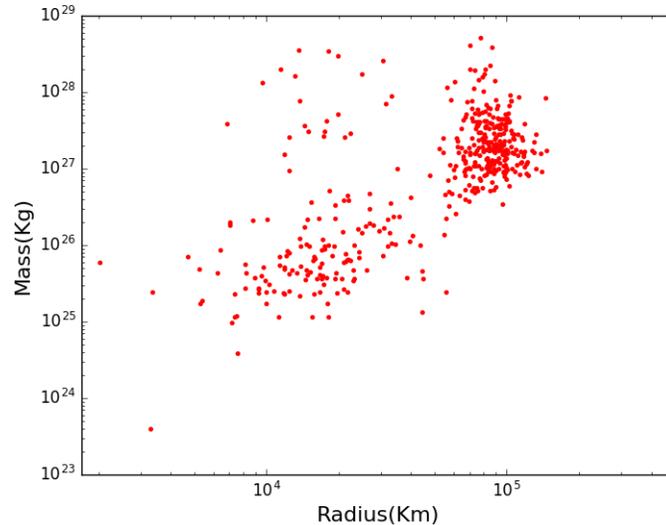
FIGURE 1. A scatter plot of the mass versus radius for the sample dataset encompassing all the 450 planets used in the study.
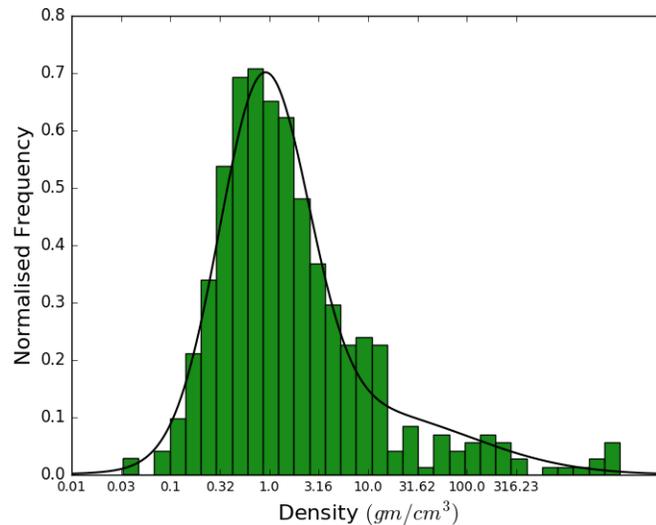

FIGURE 2. The GMM based fit for the density of the exoplanets using the best-fit parameters from Eq. 6 for $k = 2$. Details of the fit can be found in Tab. 1.

The scatter plot in Fig. 1 shows all the selected planets for the study as a function of their mass and radius. The distribution looks clustered in certain areas with lots of outliers. The density distribution of 450 exoplanets with their histograms can be found in Fig. 2 and Fig. 3 for the 2-Gaussian and 3-Gaussian fits respectively and we can see intuitively that no difference can be discerned by eye from the two figures. Both the models fit well the distribution of the density function, and hence we have to rely on quantitative model comparison tests that have been carried
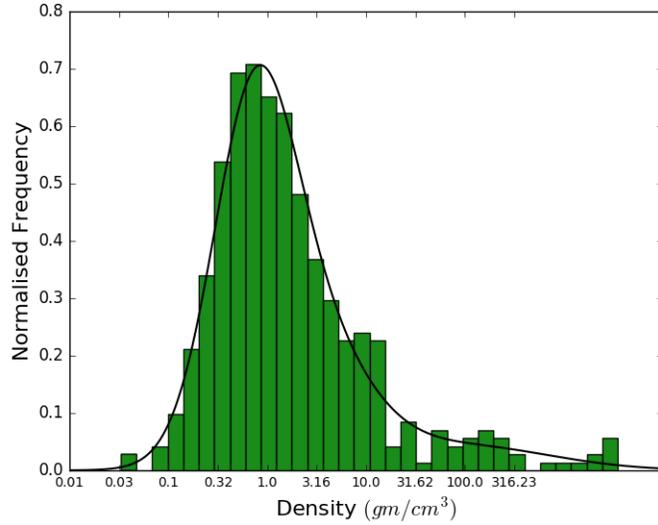
FIGURE 3. The GMM based fit for the density of the exoplanets using the best-fit parameters from Eq. 6 for $k = 3$. Details of the fit can be found in Table 1.
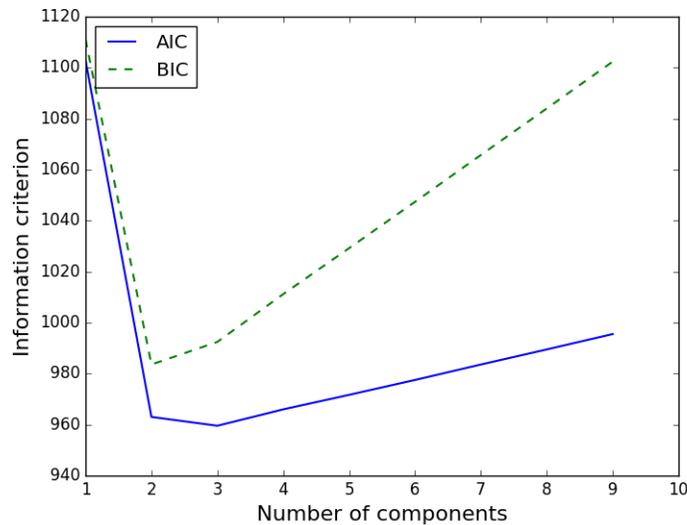

FIGURE 4. The AIC and BIC values as a function of the number of Gaussian components used to fit the density of exoplanets. AIC shows a preference for two components, whereas BIC shows a preference for three.

TABLE 1

MODEL COMPARISON PARAMETERS FOR EXOPLANETS, WHEN CLASSIFIED ACCORDING TO DENSITY. THE FIRST COLUMN DENOTES THE NUMBER OF COMPONENTS USED FOR THE FITTING. THE SECOND COLUMN IS THE BEST-FIT VALUE FOR THE LOGARITHM OF THE DENSITY (EXPRESSED IN gm/cm^3), THE THIRD COLUMN IS ITS STANDARD DEVIATION, AND THE FOURTH COLUMN INDICATES THE NUMBER OF PLANETS CLASSIFIED IN EACH CATEGORY. THE NEXT TWO COLUMNS INDICATE THE AIC AND BIC VALUE FOR BOTH THE HYPOTHESIS. THE LAST TWO COLUMNS INDICATE THE ABSOLUTE VALUE OF Δ (AIC/BIC) BETWEEN THE MODEL WITH MINIMUM VALUE AND THE NEXT HIGHEST VALUE. AS WE CAN SEE, AIC PREFERS 3 COMPONENTS, WHEREAS BIC PREFERS 2. HOWEVER, ΔAIC AND ΔBIC ARE BOTH < 10 AND HENCE THEIR SIGNIFICANCE (COMPARED TO THE VALUE WITH HIGHER AIC/BIC) IS MARGINAL.

k	μ	σ	w_i	AIC	BIC	$\Delta(AIC)$	$\Delta(BIC)$
2	0.88	0.20	322	965.7	**986.2**		
	9.69	1.08	128			5.6	6.6
3	0.71	0.17	225				
	2.03	0.36	175	**959.6**	992.8		
	88.0	0.82	50				

NOTE. — In the table the preferred value for every test is highlighted in bold.

out on the data, viz. the AIC and the BIC test. As seen in Fig. 4, the BIC test indicates that the 2-component model is the optimum model as it has the minimum BIC value followed by the 3-component model, which has a larger

value than the two component model. This trend is different from the AIC test, as the AIC has a minimum for three Gaussians, indicating that this is the best model, followed by the two-component model. These results if compared to the previous attempts at one-dimensional classification done by OR16 are very similar in both, the two Gaussian and the three Gaussian models proposed in this paper. From the 2-component model, the mean density values are at 0.88 gm/cm^3 and 9.69 gm/cm^3, with each class containing 322 and 128 exoplanets respectively. The inferred mean values of the density for the 3-component model are at 0.71 gm/cm^3, 2.03 gm/cm^3 and 88.1 gm/cm^3 with 225, 175, and 50 in each of the classes respectively. In the previous study by OR16, the mean density values are at 0.7 gm/cm^3 and 6.3 gm/cm^3 with 320 and 106 respectively in each class for 2 components and at 0.71 gm/cm^3, 6.9 gm/cm^3 and 29.1 gm/cm^3 with 340, 80, and 7 exoplanets respectively in each class for the 3 components.

From Tab. 1, we see that for the AIC test, the best model preferred is the 3-Gaussian model but there is sufficient confidence shown in both, the 2-Gaussian model and surprisingly the 4-Gaussian model (see Fig. 4) while rejecting all other models by a huge margin. As described in Sect. 3.1.1, the intervals of Δ_i are well within the range of not having sufficient evidence to reject the 2-component and 4-component models over our null hypothesis of a 3-component model. The BIC test prefers the 2-component model and has weak confidence in the 3-Gaussian model while rejecting all other models by a significant margin and hence rejecting the 4-component model as well from further consideration. Therefore, the results from the two information criterion tests do not agree. However, ΔAIC and ΔBIC are both less than 10 between the two and three component model, so no one model among these is decisively favored between the two.

4.2. *2D Classification:*

We now proceed to a 2-dimensional GMM based classification using both the logarithm of the density and ESI. We use the combined data from ESI and density using the datasets specified earlier in the manuscript and perform a two-dimensional GMM analysis. We consider only the planets that have measured values for all the quantities required for the calculation of ESI. A total of 450 exoplanets were analyzed for a range of lognormal components.
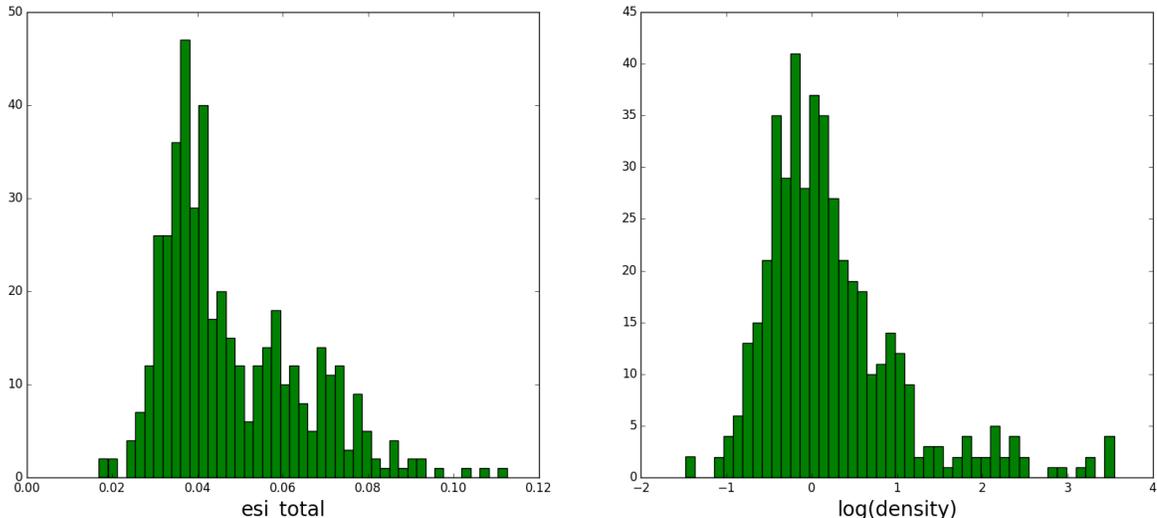

FIGURE 5. The histograms for the density (expressed in gm/cm^3) and the total ESI shown individually to see the general trend of the distribution.

As we can see from Fig. 7, we get a result that is similar to the one we saw in the above case of 1-D classification, where the 3-component component was preferred only with AIC, albeit with marginal significance, using only the density as a parameter. From this two-dimensional analysis using the total ESI along with the density, we have both AIC and BIC preferring the 3-component distribution over all the other ones and by a substantial margin. The best-fit values of the parameters along with their covariance, as well as the ΔAIC and ΔBIC values for the two and three component distributions can be found in Tab. 2

The AIC and BIC tests both point to definitive evidence for one model (three components) and give concordant results. From the statistical confidence measures ΔAIC and ΔBIC, we can assert our confidence in the hypothesis of the three-Gaussian model over all other model fits. We find that for the next preferred models in the analysis, the $\Delta AIC = 14$ and $\Delta BIC = 22$, which is significant enough to reject the respective models in favor of our null hypothesis with strong confidence.

5. CONCLUSIONS

In this manuscript, we have undertaken a classification of the exoplanet catalog using clustering based on the logarithm of the planet density (similar to a recent analysis in OR16 [1]), followed by a 2-dimensional analysis using
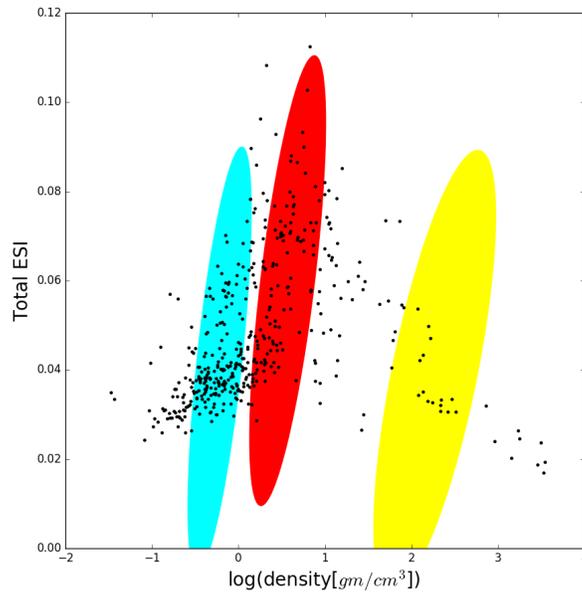
FIGURE 6. The scatter plot of the distribution using the two components, log (density) and total ESI. The three ellipses represent the 1σ confidence level region for the 3-component model, which are centered at the means of the distribution acquired from best-fit of Eq. 6 and Eq. 8.
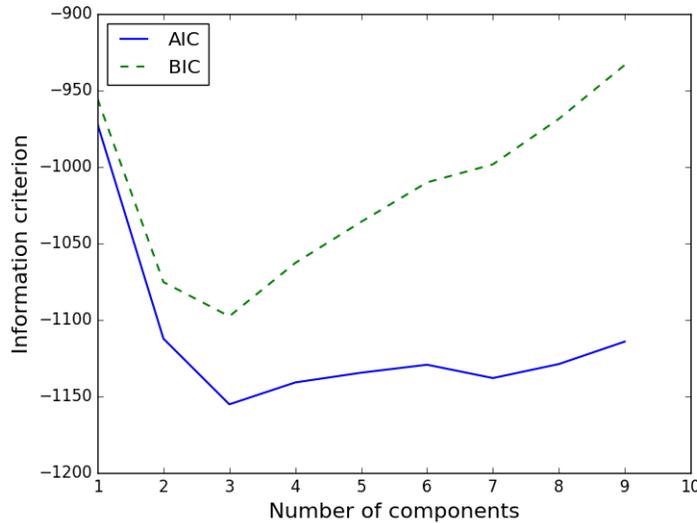

FIGURE 7. AIC and BIC values for the two dimensional GMM analysis (as a function of log(density) and ESI) over the combined data. Both AIC and BIC attain a minimum value for three components.

both the log of density and Earth Similarity Index (ESI) [2] for each of the exoplanets. We use Gaussian Mixture Model to classify the data for both the one-dimensional and two-dimensional classifications based on log(density) and {log(density), ESI} respectively. For both of these classifications, we determine the best-fit parameters for each model using the EM algorithm. We then use information theoretic criterion, such as AIC and BIC to determine the optimum number of free parameters. Our results are as follows:

1. For the one-dimensional approach, our analysis does not provide a conclusive evidence between a two-component and a three-component model, since neither of the information criterion tests cross the threshold (> 10) needed for decisive evidence. As stated in Tab. 1, the ΔAIC test weakly favors the three component Gaussian model, whereas the ΔBIC test weakly favors the two component Gaussian model. The 2 Gaussian model has the mean values of the density at 0.88 gm/cm^3 and 9.69 gm/cm^3, whereas the corresponding values for the 3 Gaussian model are located at 0.71 gm/cm^3, 2.03 gm/cm^3 and 88.1 gm/cm^3.

TABLE 2
Results from model comparison for 2D classification of exoplanets using logarithm of the density and ESI, along with results from information theoretic criterion. The explanation for the first two columns is same as in Tab. 1. The third column denotes the covariance matrix. Other columns are same as in Table 1. For this case, both AIC and BIC point to decisive evidence for three components.

k	μ	Σ		w_i	AIC	BIC	$\Delta(AIC)$	$\Delta(BIC)$
2	(-0.063, 0.046)	0.24 0.005	0.005 0.0012	332	**-1112.2**	**1075.2**		
	(1.07, 0.052)	0.902 -0.01	-0.01 0.0013	118			43.0	22.4
3	(-0.22, 0.042)	0.17 0.0036	0.0036 0.012	270	-1155.2	-1097.7		
	(0.57, 0.06)	0.33 0.0056	0.0056 0.0013	143				
	(2.27, 0.04)	0.68 -0.011	-0.011 0.0013	37				

2. The two-dimensional classification on the other hand provide robust and consistent results from both the tests. As is summarized in Tab. 2, both the tests give decisive evidence for the three component Gaussian model with ΔAIC and ΔBIC > 10 in both the cases.

REFERENCES

[1] A. Odrzywolek and J. Rafelski, "Classification of exoplanets according to density," *ArXiv e-prints* (Dec., 2016) , `arXiv:1612.03556 [astro-ph.EP]`.

[2] J. M. Kashyap, "Quantitative indexing and Tardigrade analysis of exoplanets," *ArXiv e-prints* (Feb., 2017) , `arXiv:1702.03678 [astro-ph.EP]`.

[3] K. Rice, "The Detection and Characterization of Extrasolar Planets," *Challenges* **5** (Sept., 2014) 296–323.

[4] J. I. Lunine, B. Macintosh, and S. Peale, "The detection and characterization of exoplanets," *Phys. Today* **62** no. 5, (2009) 46–51.

[5] V. F. Weisskopf, "Of atoms, mountains, and stars - A study in qualitative physics," *Science* **187** (Feb., 1975) 605–612.

[6] S. Kulkarni and S. Desai, "Classification of Gamma-Ray Burst durations using robust model-comparison techniques," *Astrophys. Space Sci.* **362** no. 4, (2017) 70, `arXiv:1612.08235 [astro-ph.HE]`.

[7] D. e. a. Schulze-Makuch, "A Two-Tiered Approach to Assessing the Habitability of Exoplanets," *Astrobiology* (Dec., 2011) . `http://online.liebertpub.com/doi/abs/10.1089/ast.2010.0592`.

[8] J. M. Rodríguez Mozos and A. Moya, "Statistical-likelihood Exo-Planetary Habitability Index (SEPHI)," *ArXiv e-prints* (July, 2017) , `arXiv:1707.07986 [astro-ph.EP]`.

[9] J. R. Bray and J. T. Curtis, "An ordination of the upland forest communities of southern Wisconsin," *Ecological monographs* **27** no. 4, (1957) 325–349.

[10] Ž. Ivezić, A. Connolly, J. Vanderplas, and A. Gray, *Statistics, Data Mining and Machine Learning in Astronomy.* Princeton University Press, 2014.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)* (1977) 1–38.

[12] Bovy Jo, D. W. Hogg, and S. T. Roweis, "Extreme deconvolution: Inferring complete distribution functions from noisy, heterogeneous and incomplete observations," *Annals of Applied Statistics* **5** (June, 2011) , `arXiv:0905.2979 [stat.ME]`.

[13] K. Shi, Y. F. Huang, and T. Lu, "A comprehensive comparison of cosmological models from the latest observational data," *MNRAS* **426** (Nov., 2012) 2452–2462, `arXiv:1207.5875`.

[14] D. L. Shafer, "Robust model comparison disfavors power law cosmology," *Phys. Rev. D* **91** no. 10, (May, 2015) 103516, `arXiv:1502.05416`.

[15] S. Desai and D. W. Liu, "A search for evidence of solar rotation in Super-Kamiokande solar neutrino dataset," *Astroparticle Physics* **82** (Sept., 2016) 86–92, `arXiv:1604.06758 [astro-ph.HE]`.

[16] S. Ganguly and S. Desai, "Statistical Significance of Lorentz invariance violation from GRB 160625B," *ArXiv e-prints* (June, 2017) , `arXiv:1706.01202 [astro-ph.IM]`.

[17] A. R. Liddle, "How many cosmological parameters?," *MNRAS* **351** (July, 2004) L49–L53, `astro-ph/0401198`.

[18] K. P. Burnham and D. R. Anderson, "Multimodel inference understanding AIC and BIC in model selection," *Sociological methods & research* **33** no. 2, (2004) 261–304.

[19] A. R. Liddle, "Information criteria for astrophysical model selection," *MNRAS* **377** (May, 2007) L74–L78, `astro-ph/0701113`.

[20] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics* **6** no. 2, (1978) 461–464.