Application of the Gaussian mixture model to examine the distribution of exoplanet population data

Mark Beyer Stjerne[1] and Ingrid Almquist Lien[1]

[1]Niels Bohr Institute, University of Copenhagen, DK-2100 Copenhagen, Denmark

Submitted March 21, 2025

Abstract

Gaussian mixture models (GMMs) offer an effective, unsupervised approach for identifying structure in multi-dimensional datasets, making them a valuable tool in exoplanet research where the true data distribution is often unknown. We begin by outlining the mathematical framework of GMMs and testing their effectiveness on synthetic data, where they successfully recover known clusters. When applied to exoplanet period-radius data, we observe clear clustering patterns. However, a two-dimensional Kolmogorov-Smirnov test shows that these clusters do not fully align with a Gaussian mixture, indicating a more complex underlying distribution. Despite this, the GMM provides a meaningful partitioning of the data, revealing structures of interest within the exoplanet population.

Keywords. Methods: statistical – Planets and satellites: fundamental parameters

1. Introduction

Since the launch of exoplanet search missions such as NASA's *Kepler* and *TESS* space telescopes, over 5,800 exoplanets have been observed and catalogued, a significant contribution in the ongoing search for Earth-like worlds. A notable feature of this dataset is the tendency for exoplanets to cluster within specific regions of parameter space, motivating further investigation into the underlying distribution of these populations. Gaussian mixture models offer a promising approach for clustering multi-dimensional data into probabilistically-defined components, making them well-suited for identifying potential structure in exoplanet populations.

In this work, we present the mathematical framework of GMMs, demonstrate their performance with synthetic data, and apply them to exoplanetary period-radius data to identify and characterize potential clusters, incorporating two outlier detection criteria, including the Mahalanobis distance. We determine the optimal number of clusters using the Bayesian Information Criterion and estimate bootstrapped confidence intervals for the resulting cluster parameters. Additionally, we discuss the Kolmogorov-Smirnov test in its one-dimensional form and present a two-dimensional extension, which we apply to assess how well the Gaussian clusters represent the underlying data distribution.

2. Theory

2.1. Gaussian mixture models

The Gaussian mixture model (GMM) is a probabilistic clustering technique, used to identify a set of K clusters in an M-dimensional parameter space that best represent the observed data distribution, assuming that each cluster consists of a multivariate Gaussian distribution (Press, 2007). This technique is categorized as an unsupervised learning method, as clusters are identified without requiring prior information about the data points' classifications.

The probability distribution $P(\boldsymbol{x}_n)$ of the data $\boldsymbol{x}$ (consisting of N data points) can be expressed as the weighted sum of the K Gaussian clusters:

$$P(\boldsymbol{x}_n) = \sum_{k=1}^{K} P(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) P(k), \tag{1}$$

where $k = 1 \dots K$. $P(k)$ is defined as the mixture weight, and $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean and covariance matrices of the k-th Gaussian cluster respectively. It follows that the density distribution of an individual Gaussian cluster is:

$$P(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)\cdot\boldsymbol{\Sigma}_k^{-1}\cdot(\boldsymbol{x}-\boldsymbol{\mu}_k)\right]}{(2\pi)^{M/2}\sqrt{|\boldsymbol{\Sigma}_k|}}. \tag{2}$$

where $|\boldsymbol{\Sigma}_k|$ is the determinant of $\boldsymbol{\Sigma}_k$. Using Bayes' theorem, the posterior probability (responsibility matrix) is given by:

$$P(k|\boldsymbol{x}_n) = \frac{P(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) P(k)}{P(\boldsymbol{x}_n)}. \tag{3}$$

The likelihood Λ describing the product of the probabilities of finding a point at each observed position is:

$$\Lambda = \prod_{i=1}^{N} P(\boldsymbol{x}_i). \tag{4}$$

The optimal values for the parameters are found by maximising the likelihood Λ. This process is akin to maximising the posterior probability of the parameters, assuming weak or non-informative priors. The parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and $P(k)$ can be determined from the data by means of the Expectation-Maximisation (EM) algorithm, which assumes no prior knowledge of clustering structures in the data:

1. Initial values for $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and $P(k)$ are assigned based on a chosen initialization method, such as random data point selection or k-means clustering.[1]
2. Expectation step (E-step): $P(\boldsymbol{x})$ and $P(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ are calculated using Equations 1 and 2.
3. Maximisation step (M-step): Model parameters are updated using:

$$\boldsymbol{\mu}_{k,\,\mathrm{new}} = \frac{\sum_{i=1}^{N} \boldsymbol{x}_i P(k|\boldsymbol{x}_i)}{\sum_{i=1}^{N} P(k|\boldsymbol{x}_i)}, \tag{5}$$

$$\boldsymbol{\Sigma}_{k,\,\mathrm{new}} = \frac{\sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \otimes (\boldsymbol{x}_i - \boldsymbol{\mu}_k) P(k|\boldsymbol{x}_i)}{\sum_{i=1}^{N} P(k|\boldsymbol{x}_i)}, \tag{6}$$

$$P(k)_{\mathrm{new}} = \frac{1}{N} \sum_{i=1}^{N} P(k|\boldsymbol{x}_i). \tag{7}$$

4. The EM steps are repeated until the total likelihood Λ (Equation 4) converges.

[1] It has been shown that poor initialization can lead to slow convergence or suboptimal solutions in the EM algorithm (Shireman et al., 2015).

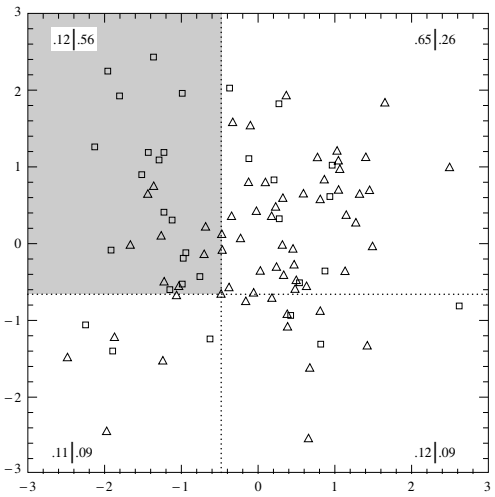

Figure 1: Example of 2D K-S test. Values in corners correspond to fraction of points from respective datasets within quadrant, i.e. (Triangle | Square). **Source:** Press (2007, p. 763).

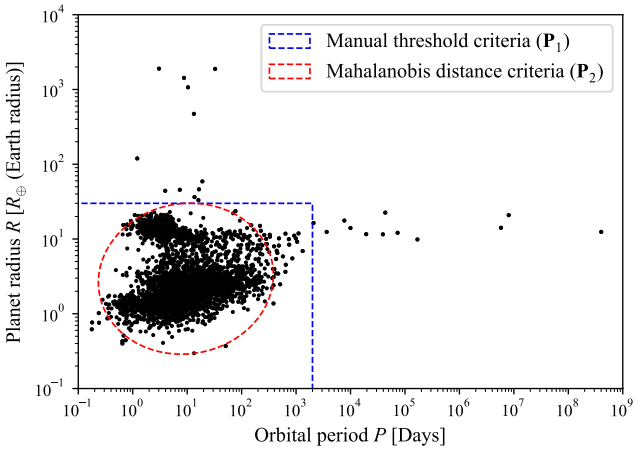

Figure 2: Orbital period-radius plot of 4,407 exoplanet candidates. Blue dashed line indicates outlier bounds based on manual selection criteria, and red dashed line indicates outlier bounds based on Mahalanobis distance criteria. The samples of exoplanets within the respective bounds are respectively denoted $\mathbf{P}_1$ and $\mathbf{P}_2$.

2.2. Two-dimensional Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (K-S) test is a non-parametric statistical test designed to compare a sample dataset with a reference probability distribution (one-sample test) or to compare two independent sample datasets (two-sample test; Press, 2007). It evaluates whether the distributions differ significantly by quantifying the maximum deviation between their cumulative distribution functions. The test in its one-dimensional formulation is commonly used for datasets characterized by a single variable. However, many datasets involve multi-dimensional parameter spaces, necessitating an extension of the K-S test to higher dimensions.

The two-dimensional K-S test generalizes the one-dimensional method to pairs of values, (x, y), which describe each data point. Unlike in one dimension, it is not trivial to define a unique cumulative probability distribution in two dimensions (Fasano & Franceschini, 1987). To address this, the test divides the plane for each data point (x_i, y_i) into four quadrants: $(x > x_i, y > y_i)$, $(x > x_i, y < y_i)$, $(x < x_i, y > y_i)$ and $(x < x_i, y < y_i)$. For each quadrant, the integrated probability (i.e. the fraction of data points contained within the quadrant) is computed, and the differences in these integrated probabilities between the two distributions are calculated. The K-S test statistic $\mathcal{D}$ is determined to be the largest absolute difference in corresponding integrated probabilities, ranging over all data points and their respective quadrants (Peacock, 1983).

Figure 1 shows a visualization of the quadrants for comparing two 2D data distributions made of 65 triangles and 35 squares. The dotted lines are centered on the triangle data point that maximizes the $\mathcal{D}$ statistic, with the maximum occurring in the upper-left quadrant. This quadrant contains 0.12 of all triangles and 0.56 of all squares, giving a $\mathcal{D}$ statistic value of 0.44. The p-value associated with the $\mathcal{D}$ statistic is then evaluated using the Kolmogorov probability distribution. Further clarification of the method is provided in Appendix A.

In applying the K-S test to data, the null hypothesis assumes the data comes from a specific distribution. However, when distribution parameters are estimated from the data (e.g., in a GMM), this creates a dependency between the data and the parameters, leading to a biased p-value (Press, 2007). To correct for this, a bootstrap test is recommended, where the K-S test is performed on datasets formed by resampling the data, to build an empirical distribution of p-values for a more reliable estimate of the null hypothesis and its significance.

3. Method

3.1. Description

We use data from the NASA Exoplanet Archive (Akeson et al., 2013), taken as of March 19, 2025, which contains 38,157 referenced exoplanet parameter measurements and uncertainties for 5,856 exoplanet candidates. For this study, we focus on identifying clusters in the parameter space of orbital period (P) and planetary radius (R). To refine the dataset, we calculate the aggregate mean for the period and radius measurements of each exoplanet, excluding entries with undetermined values for either parameter. This results in a dataset of 4,407 exoplanets, whose period-radius distribution is shown in Figure 2, where clear clustering tendencies can be observed.

Looking at Figure 2, we identify outlying exoplanets with radii around $R \sim 10^1\ R_\oplus$ (Earth radius) and periods ranging from $P \sim 10^3 - 10^9$ days, as well as exoplanets with $P \sim 10^1$ days and radii greater than $R \sim 5 \times 10^1\ R_\oplus$. These outliers could introduce noise that affects the GMM's ability to fit meaningful clusters, potentially leading to suboptimal convergence and distorted cluster parameters. To mitigate this, we formulate two outlier criteria, categorising the exoplanet data into two datasets. In one, we define parameter bounds within which we perform the GMM clustering, considering only data points in the region $0 \leq P \leq 2 \times 10^3$ days and $0 \leq R \leq 3 \times 10^1\ R_\oplus$. We denote this sample $\mathbf{P}_1$, which includes 4383 exoplanets (24 are excluded). A second method makes use of the Mahalanobis distance, a statistical measure commonly used for detecting multivariate outliers in a dataset. It quantifies the distance between an observation and the mean of a distribution, accounting for the correlations between variables through the covariance matrix. Unlike the Euclidean distance, which treats each dimension independently, the Mahalanobis distance adjusts for the variability and correlation structure of the data.

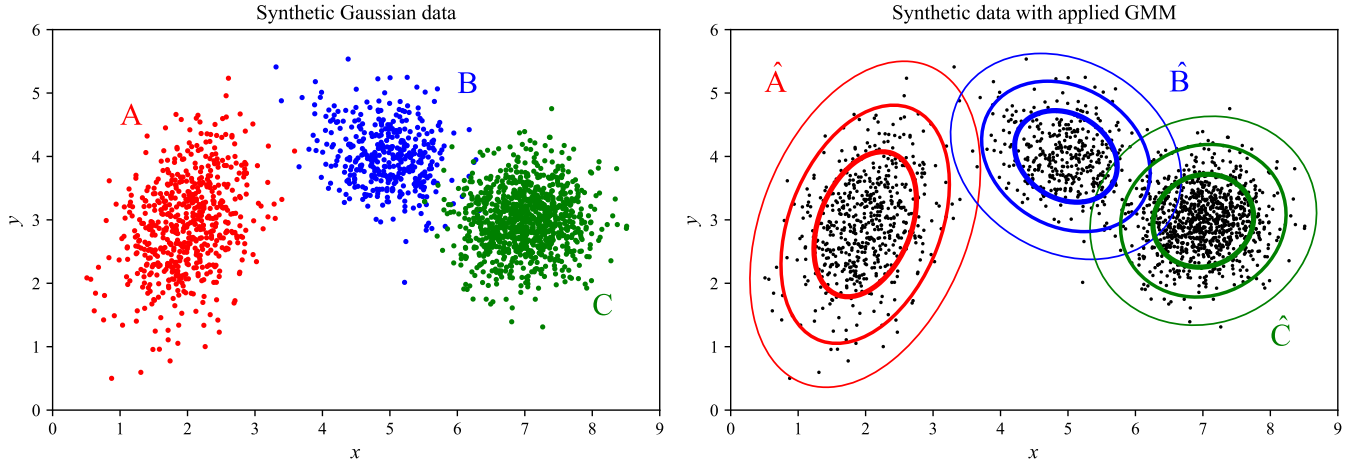

Figure 3: Left: Example iteration of synthetic data drawn from multivariate Gaussian clusters described by parameters in Table 1, with colour and label indicating source cluster k. **Right:** Synthetic data with overlaid 1σ, 2σ, and 3σ confidence ellipsoids (thicker line for the 1σ region, followed by progressively thinner lines for the 2σ and 3σ regions) representing the component Gaussians after applying the GMM.

Synthetic cluster parameters				**Fitted GMM cluster parameters (1σ CI)**			
Cluster k	μ_k	Σ_k	P_k	Cluster k	μ_k	Σ_k	P_k
A	$\begin{pmatrix}2\\3\end{pmatrix}$	$\begin{pmatrix}0.25 & 0.125\\0.125 & 0.5\end{pmatrix}$	0.3	Â	$\begin{pmatrix}1.993^{+0.027}_{-0.022}\\3.000^{+0.029}_{-0.032}\end{pmatrix}$	$\begin{pmatrix}0.248^{+0.013}_{-0.013} & 0.122^{+0.018}_{-0.012}\\0.1215^{+0.018}_{-0.012} & 0.497^{+0.032}_{-0.025}\end{pmatrix}$	$0.300^{+0.000}_{-0.001}$
B	$\begin{pmatrix}5\\4\end{pmatrix}$	$\begin{pmatrix}0.25 & -0.075\\-0.075 & 0.25\end{pmatrix}$	0.2	B̂	$\begin{pmatrix}5.009^{+0.031}_{-0.034}\\3.994^{+0.027}_{-0.025}\end{pmatrix}$	$\begin{pmatrix}0.255^{+0.022}_{-0.020} & -0.077^{+0.017}_{-0.016}\\-0.077^{+0.017}_{-0.016} & 0.244^{+0.021}_{-0.018}\end{pmatrix}$	$0.201^{+0.003}_{-0.003}$
C	$\begin{pmatrix}7\\3\end{pmatrix}$	$\begin{pmatrix}0.25 & 0\\0 & 0.25\end{pmatrix}$	0.5	Ĉ	$\begin{pmatrix}7.001^{+0.018}_{-0.014}\\2.994^{+0.02}_{-0.015}\end{pmatrix}$	$\begin{pmatrix}0.246^{+0.014}_{-0.012} & 0.001^{+0.008}_{-0.009}\\0.001^{+0.008}_{-0.009} & 0.250^{+0.011}_{-0.011}\end{pmatrix}$	$0.499^{+0.003}_{-0.003}$

Table 1: Comparison of true synthetic cluster parameters (left) vs. GMM-estimated cluster parameters (right). Uncertainties are obtained from 1σ confidence intervals from 100 bootstrap trials.

The Mahalanobis distance is computed as:

$$D = \sqrt{(x-m)^T C^{-1} (x-m)}, \tag{8}$$

where x is the vector of the observation, m is the mean vector of the independent variables, and C is the covariance matrix of the variables. A larger D indicates that the point is farther from the centroid of the distribution, suggesting that it may be a multivariate outlier (Kim, 2000). We apply this technique, considering points within the 97.5% confidence interval as non-outliers. We denote this sample $\mathbf{P}_2$, which includes 4333 exoplanets (74 are excluded). The GMM is applied on both $\mathbf{P}_1$ and $\mathbf{P}_2$ to determine the optimal clusters for the exoplanet data, given the choice of outlier criteria. The bounds of both outlier criteria are shown in Figure 2.

We utilize the GMM implementation from the Python package `scikit-learn` (Pedregosa & et al., 2012), initialising the cluster parameters using k-means. Additionally, we choose a covariance structure where each cluster is allowed its own general covariance matrix, enabling each cluster to have distinct shapes, orientations, and variances. For the GMM, 100 iterations of the EM algorithm are performed, or until convergence is achieved. The convergence criterion is monitored and reported by the model. An example plot of the natural log likelihood convergence through the EM-algorithm is provided in Appendix B. To select the best model, the GMM is refitted 100 times, and the parameters of the model with the highest likelihood given the data are stored.

To determine the optimal number of clusters, a balance must be reached between the likelihood of the fitted GMM and the complexity of the model, characterized by the number of clusters. This balance can be quantified using the Bayesian Information Criterion (BIC), which combines the model likelihood with a penalty for increased model complexity:

$$\mathrm{BIC} = k\ln(n) - 2\ln(\hat{\mathcal{L}}), \tag{9}$$

where k represents the number of free parameters in the model, n is the number of data points, and $\hat{\mathcal{L}}$ denotes the maximum likelihood of the model. Among a set of candidate models, the one with the lowest BIC is preferred, as the first term, $k\ln(n)$, penalizes excessive parameter usage, discouraging overfitting. However, despite the BIC favouring more complex models, we prioritize a parsimonious model, opting for a model with fewer components even if it results in a slightly higher BIC.

3.2. Validation

To test the GMM, we generate synthetic data from a two-dimensional Gaussian mixture model with three clusters, totaling 2,000 points. Each cluster has its own mean, covariance, and mixture weight, which represents the fraction of points assigned to that cluster relative to the entire dataset. This is illustrated in Figure 3.

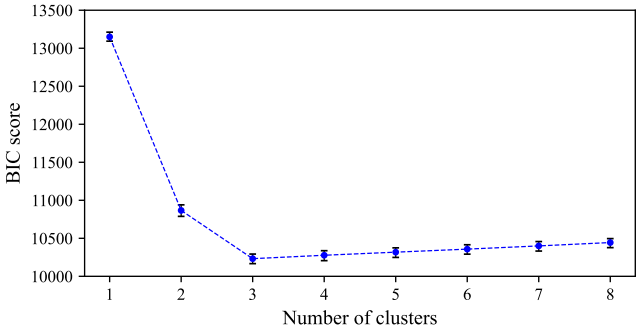

Figure 4: BIC scores for GMM applied to synthetic Gaussian data in Figure 3. Errorbars are generated using 1σ confidence intervals from 100 bootstrap trials.

We apply the GMM to the synthetic data as described in Section 3.1 and display the 1σ, 2σ, and 3σ confidence ellipsoids of the Gaussian components on top of the data in Figure 3. To determine the optimal number of Gaussian components, we calculate the BIC score for the fitted GMM with increasing cluster components, shown in Figure 4. Uncertainties in the BIC are estimated using 100 bootstrap trials, with resampling from the synthetic data distribution. We see that the lowest BIC is reached for a GMM fitted with $k = 3$ clusters, with no significant improvement observed beyond this. We conclude that the data is best described by three clusters, as anticipated.

Table 1 presents the source parameters of the Gaussian data, along with the fitted GMM parameters and their 1σ confidence intervals derived from 100 bootstrap trials. All of the source parameters fall within the confidence intervals of the fitted GMM parameters, confirming that the GMM accurately recovers the underlying Gaussian components.

To assess whether the data follows a multivariate Gaussian distribution, we apply the 2D K-S test using the Python package `ndtest`[2]. We conduct a bootstrap analysis with 200 trials, resampling data from the fitted GMMs and calculating the p-values for each sample. Figure 5 shows the resulting uniform distribution of p-values between 0 and 1, which supports the null hypothesis, indicating that the data in Figure 3 is drawn from a multivariate Gaussian distribution, as expected.

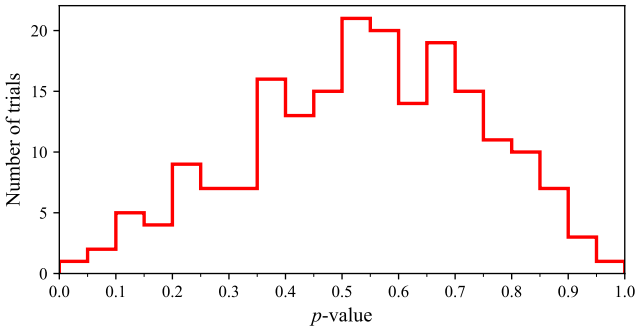

Figure 5: Distribution of p-values from 200 bootstrap trials of the 2D K-S test, comparing fitted GMM to synthetic data (Figure 3).

We conclude that the GMM is capable of detecting Gaussian structures in multidimensional data. This sets up the foundation for applying this clustering technique to examine the underlying distribution of real-world data.

[2] Written by Zhaozhou Li, https://github.com/syrte/ndtest

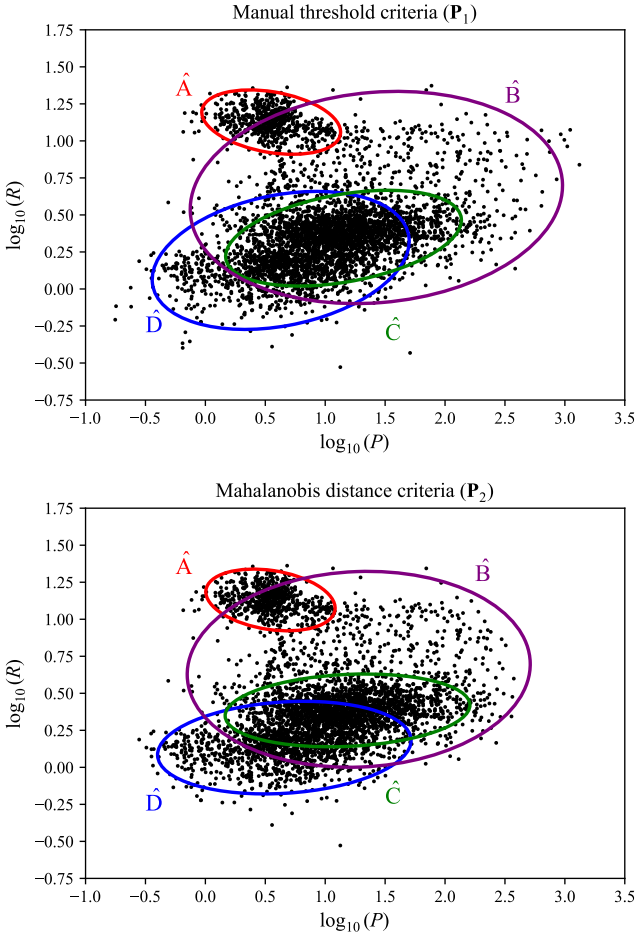

Figure 6: Top: Dataset $\mathbf{P}_1$ with 2σ confidence ellipsoids from a fitted GMM, with optimal number of clusters $k = 4$. **Bottom:** Dataset $\mathbf{P}_2$ with 1σ confidence ellipsoids from a fitted GMM, with optimal number of clusters $k = 4$. GMM-fitted cluster parameters for both datasets ($\mathbf{P}_1$ and $\mathbf{P}_2$) are provided in Appendix C.

4. Results

To apply the GMM to the exoplanet data and simplify plotting, we define our parameter space as the base-10 logarithms of the exoplanet radius and period, i.e., $x = \big(\log_{10}(P), \log_{10}(R)\big)$. We then apply the GMM as described in Section 3.1 to the datasets $\mathbf{P}_1$ and $\mathbf{P}_2$.

We find the optimal number of clusters for both datasets by inspecting the BIC scores, shown in Figure 7. For dataset $\mathbf{P}_1$, we observe a substantial decrease in the BIC score when increasing the number of clusters from $k = 1$ to $k = 2$, as well as from $k = 2$ to $k = 3$ and $k = 3$ to $k = 4$. However, beyond $k = 4$, the change in BIC is negligible, indicating diminishing returns in model complexity. Based on this, we select $k = 4$ as the optimal number of clusters for $\mathbf{P}_1$. For dataset $\mathbf{P}_2$, we see a similar trend in the BIC score, continuing to decrease from $k = 1$ until $k = 4$. Since a lower BIC score indicates a better balance between model fit and complexity, we determine that $k = 4$ is the most appropriate choice for both $\mathbf{P}_1$ and $\mathbf{P}_2$. Our use of BIC as a guiding tool for determining the optimal number of clusters reflects a subjective decision, acknowledging that the true underlying distribution of the data cannot be definitively determined. Therefore, while BIC provides a useful heuristic, its application in this context remains open to interpretation and potential inaccuracies.

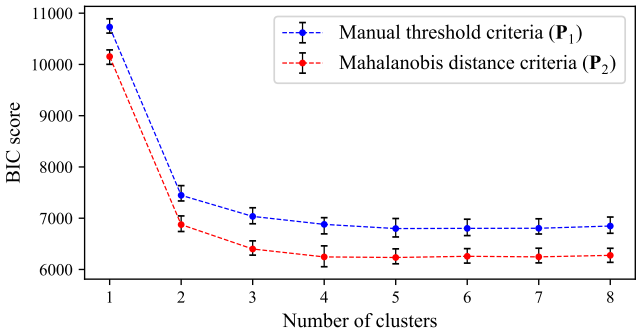

Figure 7: BIC scores for GMM applied to exoplanet data, calculated for both $\mathbf{P}_1$ and $\mathbf{P}_2$. Errorbars are generated using 1σ confidence intervals from 100 bootstrap trials.

Figure 6 shows the two datasets with overlaid 2σ confidence ellipsoids for the fitted GMM cluster components. Parameters for these cluster components are found in Appendix C. In examining the shape of the confidence ellipsoids in Figure 6, we observe that the choice of outlier detection method influences the configuration of the Gaussian components. Specifically, the confidence ellipsoids for clusters in similar regions exhibit distinct covariance matrices, depending on the outlier criterion applied. We see that the manual threshold criterion permits more data points to remain outside a well-defined boundary, resulting in clusters that are more loosely defined by the GMM. In contrast, the Mahalanobis distance criterion removes a greater number of peripheral data points, leading to a more constrained definition of cluster boundaries. However, as the clusters determined by the model remain spatially similar, it is not possible to conclude that one outlier method is definitively superior in accurately defining the GMM components within the dataset.

We verify the nature of the exoplanet data distribution by performing a two-dimensional K-S test as described in Section 3.2. Specifically, we sample data points from the fitted GMM and compare them with the exoplanet data distribution by computing the corresponding p-values. To ensure statistical robustness, we conduct 200 bootstrap trials of the 2D K-S test for these two distributions.

Figure 8 presents the resultant distribution of p-values for both datasets. For $\mathbf{P}_1$, we observe that all trials fall within the $0 \leq p \leq 0.05$ bin, suggesting that the null hypothesis—that the data is drawn from the same distribution as the fitted GMM samples—is unlikely. Similarly, for $\mathbf{P}_2$, the majority of p-values fall in the $0 \leq p \leq 0.05$ bin, decreasing until no trials yield a $p > 0.2$. This indicates that there is little statistical support for the assumption that the dataset follows Gaussian clusters.

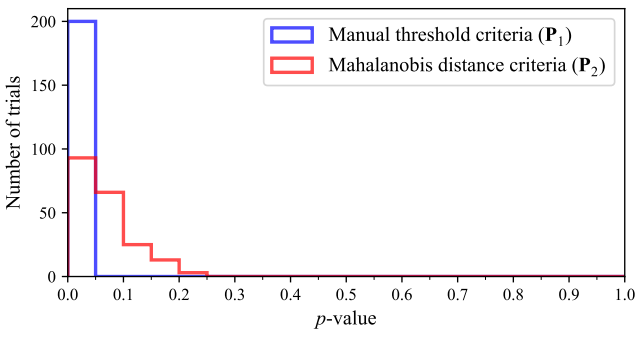

Figure 8: Distribution of p-values from 200 bootstrap trials of the 2D K-S test, comparing the fitted GMM to the source exoplanet period-radius data.

This result contrasts with our validation on synthetic data, where the GMM successfully recovered the underlying cluster structure, yielding a uniform p-value distribution consistent with the null hypothesis. The failure to obtain a similar p-value distribution for the observed exoplanet period-radius data suggests that it is not fully described by a multivariate Gaussian distribution. However, despite this limitation, the GMM still provides an optimal partitioning of the data into distinct components, which may capture meaningful structures even if the clusters themselves are not strictly Gaussian.

5. Discussion

While the GMM proved effective in locating clusters in data that is known to be Gaussian distributed, its application to exoplanet data reveals a potential limitation. Specifically, the identified Gaussian clusters identified may be artifacts arising from approximating a non-Gaussian data distribution. This can be influenced by selection effects or biases in the observed data, which may lead to artificial clustering. An approach to address this caveat when using GMMs is to compare the identified clusters with existing observational evidence from other studies (e.g., Lee et al., 2012). In doing so, one can assess whether the fitted clusters reflect genuinely Gaussian distributions or are simply approximations from model limitations.

Future works could explore alternative outlier detection methods to improve data preprocessing for GMMs, as well as testing the convergence of GMMs across various datasets. Exploring other model selection criteria, such as the Akaike Information Criterion (Akaike, 1974), could provide more reliable methods for determining an optimal distribution model. Furthermore, testing non-Gaussian mixture models, such as Dirichlet Process Mixture Models (Li et al., 2019) or Student's t-mixture models (Gerogiannis et al., 2009), could offer a more flexible approach to better capture the complex underlying distributions in exoplanet data.

6. Conclusion

In this work, we established the mathematical framework of Gaussian mixture models and the two-dimensional Kolmogorov-Smirnov (K-S) test, demonstrating their effectiveness through validation on synthetic Gaussian data. We applied these methods to exoplanet period-radius data, and identified an optimal Gaussian mixture representation of the data. However, the K-S test indicates that the data is not consistent with a purely Gaussian origin, suggesting that the underlying distribution of exoplanet populations may be more complex than initially assumed. These findings emphasise the need for further research into more complex models that could better describe the true structure of multivariate exoplanetary datasets.

Acknowledgements

This research has made use of the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with the National Aeronautics and Space Administration under the Exoplanet Exploration Program.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Akeson, R. L., et al. (2013). The NASA Exoplanet Archive: Data and Tools for Exoplanet Research. *Publ. Astron. Soc. Pac.*, *125*, 989. https://doi.org/10.1086/672273

Fasano, G., & Franceschini, A. (1987). A multidimensional version of the kolmogorov–smirnov test. *Monthly Notices of the Royal Astronomical Society*, *225*(1), 155–170. https://doi.org/10.1093/mnras/225.1.155

Gerogiannis, D., Nikou, C., & Likas, A. (2009). The mixtures of student's t-distributions as a robust framework for rigid registration. *Image and Vision Computing*, *27*(9), 1285–1294. https://doi.org/https://doi.org/10.1016/j.imavis.2008.11.013

Kim, M. (2000). Multivariate outliers and decompositions of mahalanobis distance. *Communications in Statistics-theory and Methods - COMMUN STATIST-THEOR METHOD*, *29*, 1511–1526. https://doi.org/10.1080/03610920008832559

Lee, K. J., et al. (2012). Application of the gaussian mixture model in pulsar astronomy - pulsar classification and candidates ranking for the fermi 2fgl catalogue. *Monthly Notices of the Royal Astronomical Society*, *424*(4), 2832–2840. https://doi.org/10.1111/j.1365-2966.2012.21413.x

Li, Y., Schofield, E., & Gönen, M. (2019). A tutorial on dirichlet process mixture modeling. *Journal of Mathematical Psychology*, *91*, 128–144. https://doi.org/https://doi.org/10.1016/j.jmp.2019.04.004

Peacock, J. A. (1983). Two-dimensional goodness-of-fit testing in astronomy. *Monthly Notices of the Royal Astronomical Society*, *202*, 615–627. https://doi.org/10.1093/mnras/202.3.615

Pedregosa, F., & et al. (2012). Scikit-learn: Machine learning in python. *CoRR*, *abs/1201.0490*. http://arxiv.org/abs/1201.0490

Press, W. H. (2007). *Numerical recipes: The art of scientific computing*. Cambridge University Press.

Shireman, E., Steinley, D., & Brusco, M. (2015). Examining the effect of initialization strategies on the performance of gaussian mixture modeling. *Behavior Research Methods*, *49*. https://doi.org/10.3758/s13428-015-0697-6

Appendix A: Equations for 2D K-S test

To determine the significance level (p-value) for the two-dimensional K-S test statistic $\mathcal{D}$ in the one-sample case, one can use the approximation:

$$\mathrm{P}(\mathcal{D} < \text{ Observed}) = Q_{KS}\left(\frac{\mathcal{D}\sqrt{N}}{1+\left(0.25-\frac{0.75}{\sqrt{N}}\right)\sqrt{1-r^2}}\right), \quad (10)$$

where Q_{KS} is the Kolmogorov probability function, N is the sample size, and r is the coefficient of correlation. For the two-sample case with sample sizes N_1 and N_2, N is defined for Equation 10 such that:

$$N = \frac{N_1 N_2}{N_1 + N_2}. \quad (11)$$

These formulas are accurate when the sample size N is greater than 20 and the resulting p-value is less than 0.20. For p-values greater than 0.20, the significance result is still valid, indicating that the data and model (or the two datasets) are not significantly different. However, the accuracy of the p-value may diminish at higher significance levels. In the case of perfect correlation ($r \to 1$), the two-dimensional K-S test reduces to a one-dimensional test, as both data sets lie along a perfect straight line, and the formulas for the two-dimensional test become equivalent to the one-dimensional case (Press, 2007).

Appendix B: EM algorithm convergence

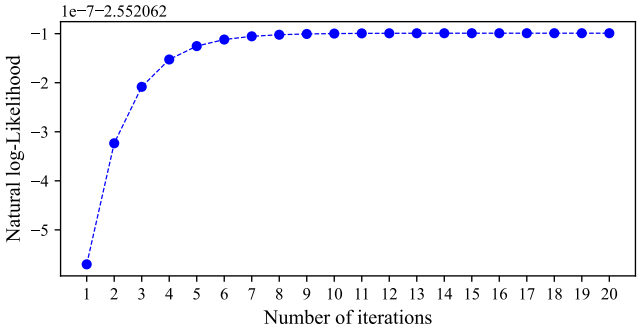

Figure 9: Natural log likelihood at each iteration step of the EM algorithm in one application of the GMM to data in Figure 3.

Appendix C: Fitted exoplanet cluster parameters

Data	k	μ_k	Σ_k	P_k
$\mathbf{P}_1$	$\hat{\mathrm{A}}$	$\begin{pmatrix} 0.550^{+0.018}_{-0.011} \\ 1.127^{+0.006}_{-0.005} \end{pmatrix}$	$\begin{pmatrix} 0.056^{+0.008}_{-0.006} & -0.007^{+0.001}_{-0.002} \\ -0.007^{+0.001}_{-0.002} & 0.008^{+0.001}_{-0.000} \end{pmatrix}$	$0.148^{+0.009}_{-0.007}$
	$\hat{\mathrm{B}}$	$\begin{pmatrix} 1.420^{+0.198}_{-0.086} \\ 0.618^{+0.041}_{-0.042} \end{pmatrix}$	$\begin{pmatrix} 0.404^{+0.033}_{-0.096} & 0.027^{+0.013}_{-0.011} \\ 0.027^{+0.013}_{-0.011} & 0.084^{+0.013}_{-0.008} \end{pmatrix}$	$0.158^{+0.016}_{-0.019}$
	$\hat{\mathrm{C}}$	$\begin{pmatrix} 1.163^{+0.039}_{-0.047} \\ 0.349^{+0.042}_{-0.026} \end{pmatrix}$	$\begin{pmatrix} 0.163^{+0.011}_{-0.018} & 0.015^{+0.010}_{-0.010} \\ 0.015^{+0.010}_{-0.010} & 0.016^{+0.004}_{-0.007} \end{pmatrix}$	$0.409^{+0.028}_{-0.041}$
	$\hat{\mathrm{D}}$	$\begin{pmatrix} 0.640^{+0.048}_{-0.049} \\ 0.190^{+0.055}_{-0.055} \end{pmatrix}$	$\begin{pmatrix} 0.192^{+0.016}_{-0.013} & 0.022^{+0.009}_{-0.011} \\ 0.022^{+0.009}_{-0.011} & 0.033^{+0.021}_{-0.016} \end{pmatrix}$	$0.288^{+0.037}_{-0.034}$
$\mathbf{P}_2$	$\hat{\mathrm{A}}$	$\begin{pmatrix} 0.548^{+0.029}_{-0.017} \\ 1.130^{+0.006}_{-0.010} \end{pmatrix}$	$\begin{pmatrix} 0.051^{+0.011}_{-0.006} & -0.005^{+0.001}_{-0.003} \\ -0.005^{+0.001}_{-0.003} & 0.007^{+0.001}_{-0.001} \end{pmatrix}$	$0.146^{+0.010}_{-0.008}$
	$\hat{\mathrm{B}}$	$\begin{pmatrix} 1.330^{+0.305}_{-0.063} \\ 0.645^{+0.048}_{-0.064} \end{pmatrix}$	$\begin{pmatrix} 0.324^{+0.022}_{-0.167} & 0.003^{+0.013}_{-0.010} \\ 0.003^{+0.013}_{-0.010} & 0.074^{+0.008}_{-0.013} \end{pmatrix}$	$0.147^{+0.022}_{-0.019}$
	$\hat{\mathrm{C}}$	$\begin{pmatrix} 1.160^{+0.032}_{-0.054} \\ 0.379^{+0.012}_{-0.044} \end{pmatrix}$	$\begin{pmatrix} 0.168^{+0.012}_{-0.036} & 0.006^{+0.013}_{-0.003} \\ 0.006^{+0.013}_{-0.003} & 0.011^{+0.008}_{-0.002} \end{pmatrix}$	$0.429^{+0.022}_{-0.040}$
	$\hat{\mathrm{D}}$	$\begin{pmatrix} 0.633^{+0.045}_{-0.043} \\ 0.140^{+0.079}_{-0.023} \end{pmatrix}$	$\begin{pmatrix} 0.180^{+0.013}_{-0.014} & 0.009^{+0.010}_{-0.005} \\ 0.009^{+0.010}_{-0.005} & 0.017^{+0.025}_{-0.004} \end{pmatrix}$	$0.280^{+0.040}_{-0.026}$

Table 2: GMM-estimated cluster parameters for exoplanet data in Figure 6. Uncertainties are obtained from 1σ confidence intervals from 100 bootstrap trials.