Bayesian Statistical Methods and Data Analysis Project Report

Mass-Radius Relation of Exoplanets

Submitted by:

Guglielmo De Toma, Stefan Visnjic, Yannis Gschwind, Manish Prasad

Abstract

In this project, we conducted a Bayesian analysis of the mass-radius (M-R) relationship of exoplanets using the updated DACE Catalog [1]. Inspired by Müller et al. (2024) [3], we constructed a probabilistic model using a broken power law with two breakpoints, classifying exoplanets into three categories: small, intermediate, and large. From this model, we obtained parameter estimates for the mass-radius power laws in each planetary regime. These results were compared with theoretical predictions. For small terrestrial planets, our result of $R \propto M^{0.31}$ is consistent with the prediction of $R \propto M^{1/3}$. In contrast, for giant planets, we obtained an exponent value (≈ 0.00) that is notably larger than the theoretically expected value of $-1/3$. Moreover, our analysis revealed that simple measurement uncertainties are insufficient to explain the observed scatter in the data. To address this, we introduced intrinsic error as a parameter in the model, which significantly improved its ability to capture the scatter. Finally, we analyzed models with 0 to 4 breakpoints (incorporating the intrinsic error). Using statistics criteria, we compared the performance of each model and we obtained that the 2-breakpoint model is the best-fitting model.

1 Introduction

Exoplanets are planets located outside the Solar system. Since the detection of the first exoplanet in 1995 [2], scientists have discovered over 5500 exoplanets. Currently, we have reached a pivotal point in exoplanet research, transitioning from detection to characterization. In exoplanet characterization, two key parameters are the planetary mass and radius. However, these properties cannot typically be measured using the same method, and in many cases, only one of them is available. To address this limitation and enable a broader statistical overview of exoplanet populations, the mass-radius (M-R) relation plays a crucial role.

Numerous studies have explored the M-R relationship using observational data. A common approach is to model the M-R relation using broken power laws, reflecting the theoretical expectations for different planetary regimes. From planetary theory, we anticipate a scaling relation of $R \propto M^{1/3}$ for terrestrial-like planets and $R \propto M^{-1/3}$ for giant planets.

In this study, we performed a Bayesian analysis of the M-R relationship for exoplanets. Specifically, we aimed to 1) build a probabilistic model; 2) fit the model to the observational data; 3) evaluate the model's ability to describe the data and compare alternative models. Bayesian Statistics is particularly suited for this study because it accounts for uncertainties in both the data and the model parameters. Moreover, it enables the incorporation of prior information from planetary theory, provides tools for model comparison, and facilitates updates to the M-R relation as new data becomes available.

2 Methods

For our analysis, we used the DACE Catalog [1] from the University of Genève, which contains data for 760 exoplanets. The catalog only includes planets with relative measurement errors on the mass and radius smaller than 25% and 8%, respectively. The masses and radii in this data set are expressed in Earth units ($M_\oplus$ and $R_\oplus$). Our analysis was inspired by the approach described by Müller et al. (2024) [3], which models the M-R relation as a piecewise linear function on a log-log scale. For n breakpoints, the n-segmented piecewise linear relation can be parametrized as follows (see [3]):

$$\mu = m + b_1\xi + \sum_{i=2}^{n} b_i(\xi - \log_{10} P_{i-1})\, H(\xi - \log_{10} P_{i-1}), \tag{2.1}$$

where ξ is the logarithm of the mass ($\xi = \log_{10}(M/M_\oplus)$) (x-axis), μ is the prediction of the model ($\mu = \log_{10}(R/R_\oplus)$) (y-axis), m is the intercept on the y-axis, $\log_{10} P_i$'s are the breakpoints on the x-axis, $H(x)$ is the Heaviside function, and the slope of the i^{th} segment is: $tan(\theta_i) = \sum_1^i b_i$.

To initiate the analysis, we adopted a model with two breakpoints as recommended by Müller et al. (2024) [3]. This choice reflects theoretical and empirical evidence suggesting three planetary regimes (small, intermediate, and large planets). Furthermore, we assumed that the data $\log_{10}(R_i/R_\oplus)$ is Gaussian distributed around the linear model, with standard deviations (σ_{y_i}) being the uncertainties in the mass measurements ($\sigma\left(\frac{R_i}{R_\oplus}\right)$) propagated into the log space:

$$\log_{10}\left(\frac{R_i}{R_\oplus}\right) \sim \mathcal{N}\left(\mu\left(m, b_1, b_2, b_3, P_1, P_2, \xi_i\right), \sigma_{y_i}^2\right), \qquad \text{where } \sigma_{y_i} = \frac{\sigma\left(\frac{R_i}{R_\oplus}\right)}{\left(\frac{R_i}{R_\oplus}\right)} \cdot \ln 10. \tag{2.2}$$

Therefore, the parameters of the model are: $m, b_1, b_2, b_3, P_1, P_2$. Uninformative uniform priors were chosen for all parameters to minimize bias in the results. Past studies ([3]) were taken into account to choose the intervals of the uniform distributions.
After defining the model, we fitted it by maximizing the posterior probability distribution, yielding the Maximum A Posteriori (MAP) estimates for the model parameters. Then, to gain insights into the posterior probability distribution and quantify the uncertainties associated with the parameters, we sampled the posterior distribution using the Markov Chain Monte Carlo (MCMC) method, implemented by the emcee Python package. The number of parameters is equal to 6. The number of walkers was set to $n_{walker} = 2 \cdot n_{param} + 2$ and we chose 32000 steps per walker. The resulting chains were thinned by sampling every half auto-correlation time steps. To avoid the influence of the burn-in phase, the initial steps, approximately ten times the auto-correlation time, were discarded. Finally, we evaluated the model's quality by generating random samples from both the predictive model distribution and the posterior predictive distribution. These samples were compared against the observed data to evaluate the model's ability to reproduce the data.
Subsequently, we extended this Bayesian workflow to models with varying numbers of breakpoints, ranging from 0 to 4. To determine the best-fitting model, we evaluated the Bayesian Information Criterion (BIC), which was also used in Müller et al. [3]. For a Gaussian likelihood, the BIC can be computed as follows:

$$\text{BIC} = N \ln\left(\frac{1}{N}\sum_i (x_i - \hat{x}_i)^2\right) + n_{param} \ln N, \tag{2.3}$$

with x_i being the data points, $\hat{x}_i$ the predicted points from the MAP model, N the number of data points, and n_{param} the number of parameters. In addition to the BIC, the Deviance Information Criterion (DIC) and Widely Applicable Information Criterion (WAIC) were computed. All these metrics balance goodness-of-fit with model complexity by favouring models with higher likelihood while penalizing those with an excessive number of parameters, thus helping to mitigate overfitting.

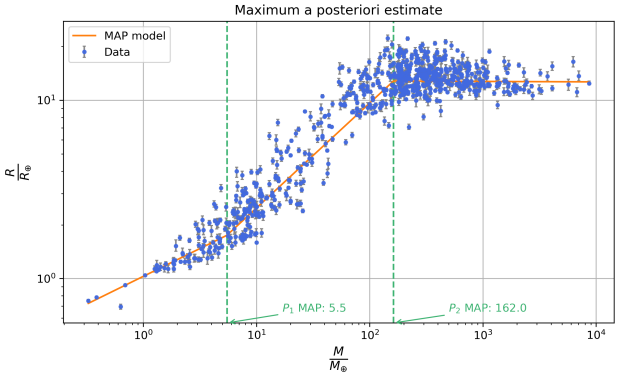
3 Results and Discussion

Table 1 shows the Maximum A Posteriori (MAP) parameter values and Figure 1a the corresponding model. Qualitatively, the MAP model captures the average behaviour of the data well. Figure 1b displays the chains from the Markov Chain Monte Carlo (MCMC) sampling, illustrating that the parameter distributions reach stationarity after the burn-in phase. The walkers effectively explore the parameter space without getting trapped in local minima. The parameters' auto-correlation times ranged between 50 and 105 steps.
After discarding the burn-in phase and applying thinning (see Methods), the mean and standard deviation of the parameters were computed. We observed that the mean values are in good agreement with the MAP parameter values. Then, these values were transformed back to the physical power law. The results are presented in Eq. (3.1):

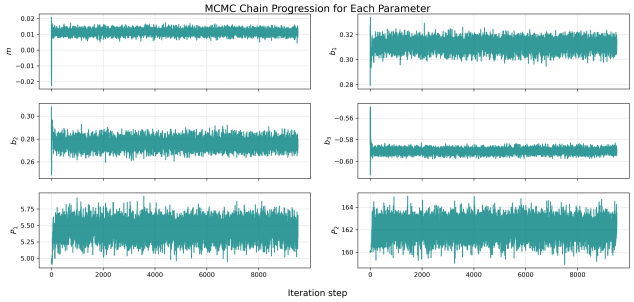
$$\frac{R}{R_\oplus} = \begin{cases} (1.027 \pm 0.004)\frac{M}{M_\oplus}^{(0.312\pm0.004)} & \text{if } \frac{M}{M_\oplus} < (5.456 \pm 0.005) \\ (0.643 \pm 0.004)\frac{M}{M_\oplus}^{(0.588\pm0.006)} & \text{if } (5.456 \pm 0.005) < \frac{M}{M_\oplus} < (162.111 \pm 0.002) \\ (12.975 \pm 0.157)\frac{M}{M_\oplus}^{(-0.003\pm0.007)} & \text{if } \frac{M}{M_\oplus} > (162.111 \pm 0.002) \end{cases}. \tag{3.1}$$

Parameter	m	b_1	b_2	b_3	P_1	P_2
MAP Value	0.012	0.312	0.275	-0.591	5.481	161.986

Table 1: MAP parameter values.



(a) MAP model (orange) fitted to the mass-radius observations of the exoplanets (blue) with corresponding y-axis error bars (grey). The locations of the breakpoints are marked with dotted lines (green).

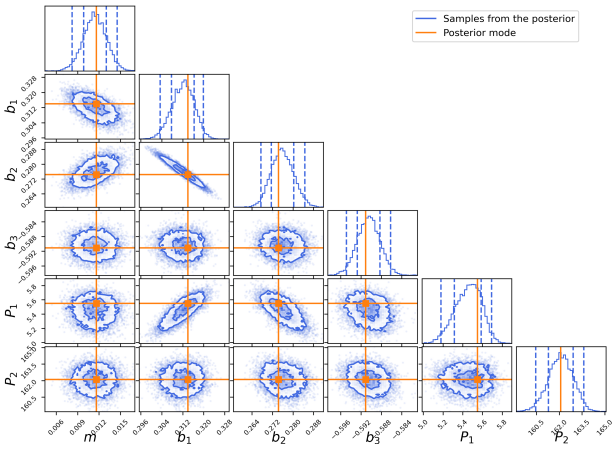


(b) MCMC chains of the six parameters after thinning. The first 1000 disregarded steps are also shown.

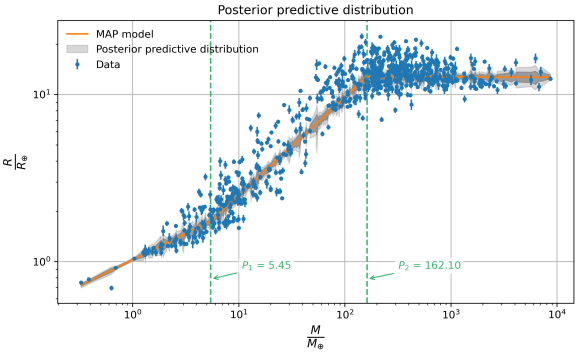
From these results, some physical interpretations can already be discussed. For example, the mass-radius (M-R) relationship of small terrestrial planets is theoretically expected to follow a power-law exponent of approximately 1/3. Our result of ≈ 0.31, is consistent with this prediction, validating our model's consistency with planet formation theory for small planets. In contrast, for giant planets, we obtained an exponent value (≈ 0.00) that is significantly larger than the theoretically expected value of $-1/3$.

The corner plot (see Figure 2a) shows the correlation between the parameters. Notably, parameters b_1 and b_2 exhibit significant correlation. This can be interpreted as follows: as the slope of one segment of the piecewise line increases, the following segment's slope decreases to fit the data. Conversely, parameters b_2 and b_3 show weaker correlation, potentially due to higher scatter in the third segment compared to the first.

Figure 2b shows the posterior predictive distribution (PPD). In the PPD plot, the percentile bands (2.5%, 16%, 84% and 97.5%) do not overlap with all data points given their corresponding uncertainties. This indicates that the model struggles to fully capture the scatter of the data around the MAP fit.



(a) Corner plot of the six parameters. The outermost subplots are the 1D histogram of the parameter distributions, while the inner subplots show the 2D histograms representing the joint distributions between pairs of parameters. The orange lines mark the location of the posterior mode. The blue lines divide the 2D histograms into regions corresponding to the 2.5%, 16%, 84%, and 97.5% quantiles.



(b) Posterior predictive distribution, shown in grey tones. The colors have two different intensities representing the 2.5%, 16%, 84% and 97.5% quantiles. The MAP model (orange), the breakpoints (green) and the data points (blue) are also shown.

Figure 2

This discrepancy suggests that the M-R relationship is not as narrow as we assumed (the variance of the likelihood was only given by the uncertainty in the measurements of radii). The relation has a significant *intrinsic scatter*. Two main factors contributing to this scatter are: 1) Planetary radius depends not only on mass but also on additional variables, such as stellar irradiation and planetary age. 2) Within each segment of the piecewise power law, exoplanet compositions can vary considerably.

To address this, we introduced an *intrinsic error* parameter in the likelihood. This term effectively represents three parameters, one for each segment of the power law. Their prior distributions were set to uniform distributions, with intervals determined based on the scatter of the data around the MAP model. As an initial simplification, we fixed the breakpoint parameters (P_1, P_2) to the MAP estimates obtained from the previous model (see Table 1). After incorporating intrinsic scatter, we repeated the Bayesian analysis as described in the Methods section. The updated PPD results (Figure 3) show improved agreement between the percentile bands and observed data, suggesting that the model now better describes the data's variability.

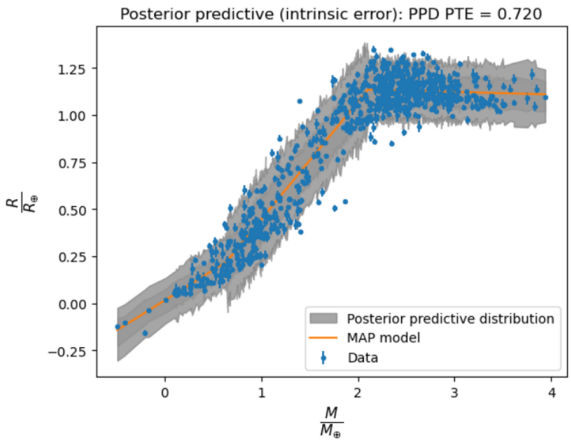


Figure 3: Posterior predictive distribution of the model with intrinsic error. The 2.5%, 16%, 84%, and 97.5% quantiles of the distribution are displayed in shades of grey. The MAP model (orange) and the data points (blue) are also shown.

$$\frac{R}{R_\oplus} = \begin{cases} (1.032 \pm 0.036) \left(\frac{M}{M_\oplus}\right)^{(0.320 \pm 0.030)} & \text{if } \textit{small} \text{ exoplanets} \\ (0.655 \pm 0.021) \left(\frac{M}{M_\oplus}\right)^{(0.627 \pm 0.008)} & \text{if } \textit{intermediate} \text{ exoplanets} \\ (14.581 \pm 0.890) \left(\frac{M}{M_\oplus}\right)^{(-0.013 \pm 0.010)} & \text{if } \textit{giant} \text{ exoplanets} \end{cases} . \tag{3.2}$$

Quantitatively, the Probability to Exceed (PTE[1]) value for the updated PPD is 0.720, confirming the model's ability to generate replicated data consistent with observations. Moreover, while the mean parameter estimates (see Eq. 3.2) remained consistent with those of the model that only accounted for measurement uncertainty (Eq. 3.1), the introduction of intrinsic scatter increased parameter uncertainties by approximately an order of magnitude.
We extended the Bayesian workflow to models with varying numbers of breakpoints (0–4), incorporating intrinsic scatter. (Breakpoints were fixed using MAP estimates from the models without intrinsic error terms.)
We compared their performance using the DIC, WAIC, and BIC criteria (see Methods). As shown in Table 2, both WAIC and DIC slightly favored the model with 4 breakpoints, while BIC preferred the model with 2 breakpoints. Since it is not clear which model is the best based solely on these criteria, we still favored the simplest model, i.e., the model with 2 breakpoints.

N breakpoints	0	1	2	3	4
DIC	-614.6	-1441.3	-1474.3	-1482.2	-1488.4
WAIC	-614.2	-1441.2	-1473.7	-1480.8	-1487.6
BIC	-600.1	-1418.1	-1441.5	-1441.0	-1437.3

Table 2: Model performance for varying number of breakpoints.

4 Conclusion

In this project, we aimed to study the dependency between mass and radius of exoplanets and to investigate the transitions between different planetary regimes. To achieve this, we conducted a Bayesian analysis on the

[1]The PTE PPD is the probability that the chi-squared (χ^2) value for the replicated data (y_{rep}) from the posterior predictive distribution exceeds the chi-squared value for the observed data (y): $Pr(\chi^2(y_{rep}, \theta) \geq \chi^2(y, \theta)|y)$, where θ stands for the parameters.

mass-radius relationship of exoplanets, using the updated DACE Catalog [1]. A probabilistic model using a broken power law with two breakpoints (see Müller et al. (2024) [3]) was constructed, allowing the classification of exoplanets into three categories: small, intermediate, and large. The initial model accounted for only the measurement uncertainties in the observed radii. We estimated the power-law parameters for each planetary regime and compared the resulting mass-radius power-law exponents for the three regimes with theoretical predictions. For small terrestrial planets, our result of ≈ 0.31 is consistent with the prediction of $1/3$. In contrast, for giant planets, we obtained an exponent value (≈ 0.00) that is substantially larger than the theoretically expected value of $-1/3$.

Our analysis of the posterior predictive distribution revealed that, measurement uncertainties alone were not sufficient to explain the observed scatter in the data. To address this, we introduced intrinsic error as a parameter in the model, which significantly improved its ability to capture the scatter (PPD PTE = 0.720).

We analyzed models with 0 to 4 breakpoints, incorporating intrinsic scatter. Model performance was evaluated using DIC, WAIC, BIC. While WAIC and DIC slightly favored the 4-breakpoint model, BIC supported the 2-breakpoint model. Due to the non-univocal results of the criteria, we chose the simplest of the top three models: the model with 2 breakpoints.

Further exploration could focus on incorporating measurement errors in planetary masses, which were assumed to be exact in this analysis. Furthermore, extending the model to account for additional factors, such as stellar irradiation or planetary age, could refine our understanding of the mass-radius relationship. Finally, applying this approach to larger datasets or newly discovered exoplanets would also help validate and extend the findings presented here.

References

[1] DACE - Data Analysis Center for Exoplanets. *DACE Exoplanet Archive*. Accessed: December 2024. URL: https://dace.unige.ch/exoplanets/.

[2] Michel Mayor and Didier Queloz. "A Jupiter-mass companion to a solar-type star". In: *nature* 378.6555 (1995), pp. 355–359.

[3] Müller, Simon et al. "The mass-radius relation of exoplanets revisited". In: *A&A* 686 (2024), A296. DOI: 10.1051/0004-6361/202348690. URL: https://doi.org/10.1051/0004-6361/202348690.