Bayesian Analysis of Kepler's Third Law Discovery and Bacteremia Classification

Aliaksandr Hubin, Georg Heinze, Michael Kammer, Mariana Nold

ROeS, Graz, September 2025

Research Question

- Derive physical laws using statistical techniques based on the Open Exoplanet Catalogue Tables [5].
- Concentrate on the 3rd Kepler's law with the response variable `semimajoraxis`.
- Validate the gravitational constant G estimated from the discovered law.

Details and Background

- Open Exoplanet Catalogue: Live Database on discovered extra-solar planet.

- Read the most uptodate data using:

```
data =
read.csv("https://raw.githubusercontent.com/
OpenExoplanetCatalogue/oec_tables/master/comma_separated
open_exoplanet_catalogue.txt")
```

Dataset and Relevant Variables

Name	Value
Rows	5,414
Columns	25
Discrete columns	7
Continuous columns	18
All missing columns	0
Missing observations	46,203
Total observations	135,350

Table 1: Basic Statistics

Overview of the Variables

Field	**Description**
name	Primary identifier of the planet
binaryflag	Binary flag [0=no known stellar binary companion; 1=P-type binary (circumbinary); 2=S-type binary; 3=orphan planet (no star)]
mass	Planetary mass [Jupiter masses]
radius	Radius [Jupiter radii]
period	Period [days]
semimajoraxis	**Semi-major axis [Astronomical Units]**
eccentricity	Eccentricity
periastron	Periastron [degree]
longitude	Longitude [degree]
ascendingnode	Ascending node [degree]
inclination	Inclination [degree]
temperature	Surface or equilibrium temperature [K]
age	Age [Gyr]
discoverymethod	Discovery method
discoveryyear	Discovery year [yyyy]
lastupdate	Last updated [yy/mm/dd]
system_rightascension	Right ascension [hh mm ss]
system_declination	Declination [+/-dd mm ss]
system_distance	Distance from Sun [parsec]
hoststar_mass	Host star mass [Solar masses]
hoststar_radius	Host star radius [Solar radii]
hoststar_metallicity	Host star metallicity [log relative to solar]
hoststar_temperature	Host star temperature [K]
hoststar_age	Host star age [Gyr]
list	A list of lists the planet is on

IDA

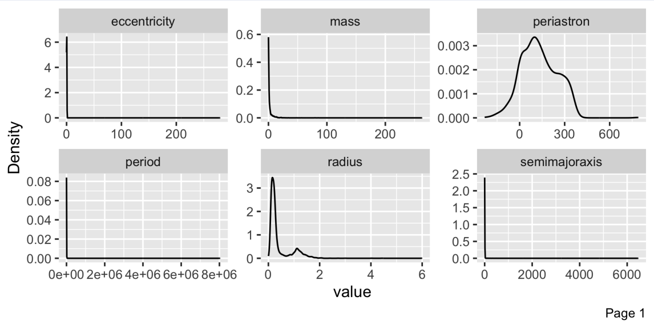


Figure 1: Density plots of the continuous variables

IDA

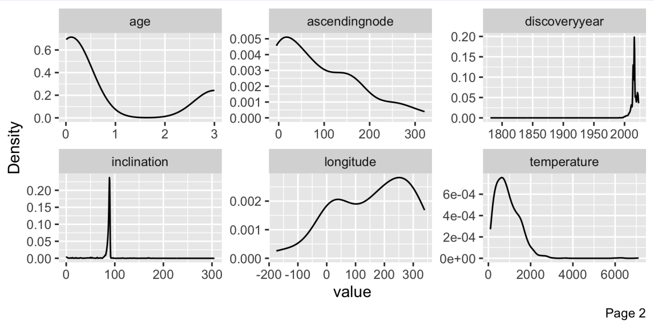


Figure 2: Density plots of the continuous variables

IDA

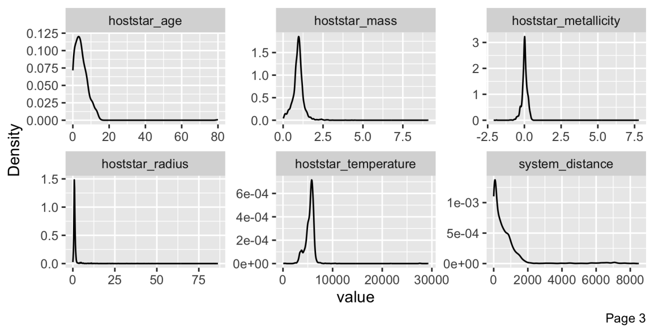


Figure 3: Density plots of the continuous covariates

IDA

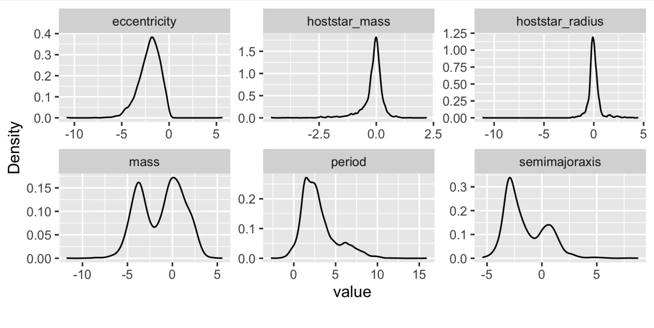


Figure 4: Density plots of the log of selected continuous covariates

IDA

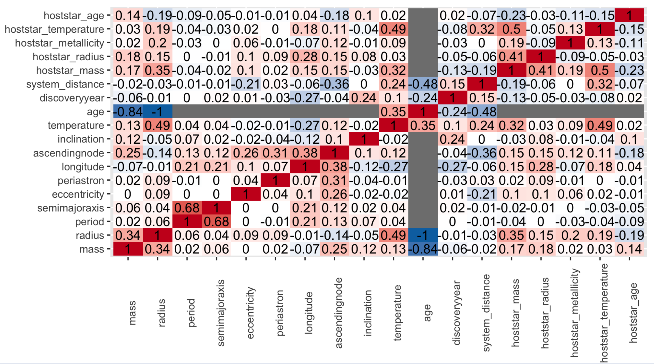


Figure 5: Pearson correlations between the continuous covariates. Covariate age is missing in 99.93% of the cases.

Statistical Issues Addressed

- Model selection and interpreting parameters in an *M-Closed* setting.
- Nonlinearities are important.
- Can one derive the true law statistically?
- Does statistical estimates' uncertainty cover the true gravitational constant G?
- Can one *"beat"* the true law with a simple additive predictive model?

Missing data pattern

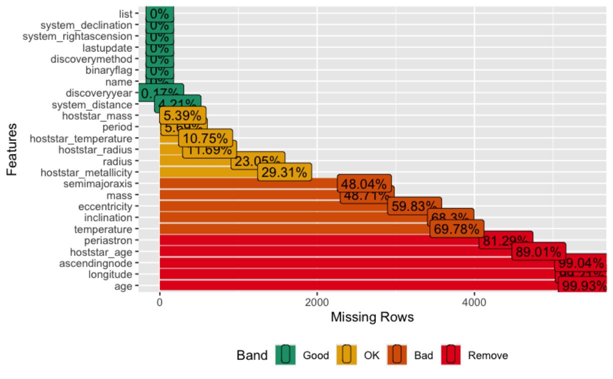


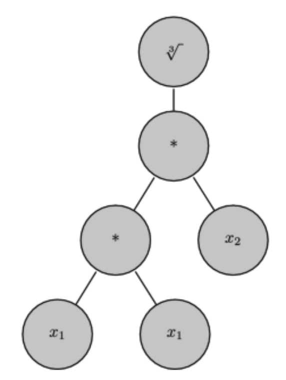
Figure 6: Missing data patterns. It was decided to select only physics relevant columns with reasonably few missing data.

Kepler Analysis: Methodology Overview

- **Data**: 939 complete observations on semimajoraxis,mass, radius, period, eccentricity, hoststar_mass, hoststar_radius, hoststar_metallicity, hoststar_temperature, binaryflag from Open Exoplanet Catalogue; split into 639 training, 300 test.
- **Model Chosen**:
- Bayesian Generalized Nonlinear Models (BGNLM) [3] with both Jeffreys and g-priors [4].
- **Also tried**:
  1. Bayesian Linear Regression (BLR) with Jeffreys prior [2].
  2. Bayesian Fractional Polynomials (BFP) [1].
- **Evaluation**: Posterior inclusion probabilities (PIP), predictive R^2, coverage for G.
- **Tool**: `FBMS` `cran.r-project.org/web/packages/FBMS/`.

Bayesian Generalized Nonlinear Models (BGNLM)

- **Response:** $Y_i|\mu_i, \phi \sim \mathfrak{f}(y|\mu_i, \phi)$.
- **Model:**
  $\mathrm{h}(\mu_i) = \beta_0 + \sum_{j=1}^{q} \gamma_j \beta_j F_j(\boldsymbol{x}_i, \boldsymbol{\alpha}_j)$,
  where F_j are functional trees.
- **Feature constraints:** Limited depth, limited set of algebraic operators $(+, *, g_1(\cdot), ..., g_k(\cdot))$ allowed.
- **Priors:** Encourage parsimony, with complexity $c(F_j)$ based on the number of algebraic operators regularizing prior inclusions.
- **More: Florian Frommlet** in GS-8.



*Functional tree:

$F = \sqrt{x_1^2 x_2} \Rightarrow a \propto \sqrt{P^2 M}$

Kepler Analysis: Bayesian Linear Regression (BLR)

- **Model**: Linear regression with Jeffreys prior, 5000 MCMC iterations (stability checks over 20 repetitions).
- **Results**: Effect sizes positive for period, mass, eccentricity; negative for hoststar_metallicity. $R^2 = 0.953$ **(train)**, 0.964 **(test)**.

Feature	**PIP**
period	1.000
mass	1.000
eccentricity	0.999
hoststar_metallicity	0.539
hoststar_mass	0.038

Table 3: BLR Posterior Inclusion Probabilities

- **Note**: Low PIP for hoststar_mass is unexpected given its role in Kepler's Law.

Kepler Analysis: Bayesian Fractional Polynomials (BFP)

- **Model**: BFP with transformations (p0, p2, p3, p05, pm05, pm1, pm2, p0p0, p0p05, p0p1, p0p2, p0p3, p0p05, p0pm05, p0pm1, p0pm2), 20 chains, 10 cores.
- **Results**: Non-linear terms improve fit. $R^2 = 0.998$ **(train), 0.997 (test)**.

Feature	**PIP**
period	1.000
*p*0*p*05(period)	0.999
*pm*2(hoststar_metallicity)	0.986
*pm*2(mass)	0.986
eccentricity	0.986
p0p1(hoststar_mass)	0.986
radius	0.981

Table 4: BFP Posterior Inclusion Probabilities

- **Note**: Misses hoststar_mass interaction critical for Kepler's Law.

Kepler Analysis: Bayesian Generalized Nonlinear Models (BGNLM)

- **Model**: Transformations (e.g. sin, exp_dbl, log, troot, p3), Jeffreys/g-priors, 64 parallel chains.
- **Results**: $R^2 = 1.000$ **(train)**, 1.000 **(test)**.

Feature	**PIP**
$troot((period^2 \cdot hoststar\_mass))$ (Jeffreys)	1.000
$troot((period^2 \cdot hoststar\_mass))$ (g-prior)	1.000

Table 5: BGNLM Posterior Inclusion Probabilities

- **Note**: Exactly recover Kepler's Law functional form: $a \propto (P^2 M)^{1/3}$.

Estimating the gravitational constant

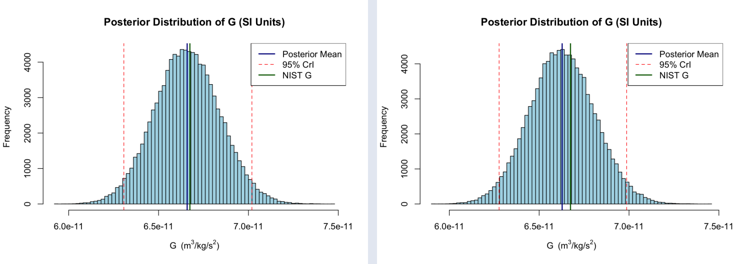


Figure 7: Posterior distribution for G under Jeffreys prior (left) and g-prior (right). g-prior induces shrinkage on the regression coefficient, hence posterior mean is slightly shifted to 0 as compared to the objective Jeffreys prior.

Kepler Analysis: Model Comparison and Conclusions

Model	Train R^2	Test R^2	Captures Kepler's Law	Covers true G
BLR	0.953	0.964	No	NA
BFP	0.998	0.997	Partial	NA
BGNLM (Jeffreys)	1.000	1.000	Yes	Yes
BGNLM (g-prior)	1.000	1.000	Yes	Yes

Table 6: Model Performance

Conclusions:

- BGNLM excels, recovering $a \propto (P^2M)^{1/3}$ without any preliminary preparing the data, etc. Also covering the true G through credible intervals.
- This is achieved with no prior knowledge of physics whatsoever and minimal statistical efforts.

Frequentist Robustness Check I

- Symbolic regression:
  - A symbolic regression fitter was programmed
  - Uses same catalogue of operators as BNGLM model
  - run on training data
- Estimation of G constant (with CI)
- Stability investigation:
  - ... random number seed
  - ... data (bootstrap)
- Validation on test set

Frequentist Robustness Check II

- Symbolic Regression identified correct model:
  `cbrt((hoststar_mass * (period * period)))`
- *G* constant estimation:
  - 95% CI covered true value if using a bootstrap interval and/or log transformation (residual analysis!)
- Stability:
  - Random number seed: correct model 69% of replications
  - Bootstrap: correct model in 74% of replications
- Validation: Perfect calibration, $R^2_{test} = 0.9999712$

Conclusion Frequentist vs Bayesian Kepler's Law Recovery

- Both Bayesian (BGNLM) and frequentist symbolic regression recover Kepler's 3rd law accurately.
- BGNLM recovers the exact law and covers the true G without log transform or residual checks, thus was in this sense slightly more robust for a lazy statistician.
- Frequentist method achieves high stability and uses bootstrap for uncertainty quantification.
- Bayesian approach offers principled inference; frequentist relies on resampling.
- Similar results, but in the frequentist a novel custom implementation of Symbolic regression was needed as standard ones failed.

References

[1] Aliaksandr Hubin, Georg Heinze, and Riccardo De Bin. **"Fractional Polynomial Models as Special Cases of Bayesian Generalized Nonlinear Models"**. In: *Fractal and Fractional* 7.9 (2023), p. 641.

[2] Aliaksandr Hubin and Geir Storvik. **"Mode jumping MCMC for Bayesian variable selection in GLMM"**. In: *Computational Statistics & Data Analysis* 127 (2018), pp. 281–297.

[3] Aliaksandr Hubin, Geir Storvik, and Florian Frommlet. **"Flexible Bayesian nonlinear model configuration"**. In: *Journal of Artificial Intelligence Research* 72 (2021), pp. 901–942.

[4] Yingbo Li and Merlise A Clyde. **"Mixtures of g-priors in generalized linear models".** In: *Journal of the American Statistical Association* 113.524 (2018), pp. 1828–1845.

[5] Hanno Rein. **"A proposal for community driven and decentralized astronomical databases and the Open Exoplanet Catalogue".** In: *arXiv preprint arXiv:1211.7121* (2012).