ASTM21: Project 2

Robin Emanuelsson

November 2019

Introduction

In the last 20 years over 3000 exoplanets have been detected. This number of samples is enough and the probability of detecting these as a function of different parameters, such period, eccentricity, is sufficient. This data might give us some critical clues in determining the mechanism of planet formation. In this report we will use RV data (not Kepler data since that data determines our parameter of interest indirectly) from the website `http://exoplanets.org` to determine the shape parameters of a beta distribution that is assumed to be the correct distribution of eccentricities for some interval of orbital periods.

Theory

The eccentricities, in some interval of periods, of the exoplanets are assumed to be given by the beta distribution

$$P(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{1}$$

where x, in this case, is the eccentricity data point, $0 \leq x \leq 1$ and α and β are the shape parameters $\alpha, \beta > 0$ and $\Gamma(x)$ is the gamma function.
This is a reasonable assumption since the eccentricity is naturally bounded between 0 and 1, just like the beta distribution. To get statistical model under which the observed data is most probable we need estimate the shape parameters.

Maximum Likelihood Estimators

One way of estimating the parameters is by optimizing some likelihood function. If n data points, x_i, are measured independently the likelihood function is defined by

$$L(\alpha,\beta|\mathbf{x}) = \prod_{i=1}^{n} P(x_i|\alpha,\beta) \tag{2}$$

$$= \prod_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{3}$$

The log of this function is often more useful and is

$$\ln(L(\alpha,\beta|\mathbf{x})) = n\ln(\Gamma(\alpha+\beta)) - n\ln(\Gamma(\alpha)) - n\ln(\Gamma(\beta)) \tag{4}$$

$$+ (\alpha-1)\sum_{i=1}^{n}\ln(x_i) + (\beta-1)\sum_{i=1}^{n}\ln(1-x_i) \tag{5}$$

$$= l(\alpha,\beta|\mathbf{x}) \tag{6}$$

To optimize this we need to set the derivatives with respect to each of the parameters to zero, that is

$$\partial_\alpha l(\alpha,\beta|\mathbf{x}) = \frac{n\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} - \frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^{n}\ln(x_i) = 0 \tag{7}$$

$$\partial_\beta l(\alpha,\beta|\mathbf{x}) = \frac{n\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)} - \frac{n\Gamma'(\beta)}{\Gamma(\beta)} + \sum_{i=1}^{n}\ln(1-x_i) = 0 \tag{8}$$

There is no closed form solution to these equation so we will solve them with the Newton-Raphson method.
We set $\theta = (\alpha, \beta)$ and iterate according to

$$\theta_{i+1} = \theta_i - J^{-1}g \tag{9}$$

where g is

$$g_1 = \psi(\alpha) - \psi(\alpha+\beta) - 1/n\sum_{i=1}^{n}\ln(x_i) \tag{10}$$

$$g_2 = g_1 = \psi(\beta) - \psi(\alpha+\beta) - 1/n\sum_{i=1}^{n}\ln(1-x_i) \tag{11}$$

here $\psi(x)$ is the digamma function $\frac{\Gamma'(x)}{\Gamma(x)}$.
J is the Jacobin and is equal to

$$\begin{bmatrix} \psi'(\alpha)-\psi'(\alpha+\beta) & -\psi'(\alpha+\beta) \\ -\psi'(\alpha+\beta) & \psi'(\beta)-\psi'(\alpha+\beta) \end{bmatrix} \tag{12}$$

where $\psi'(x)$ is the trigamma function $\frac{\Gamma''(x)}{\Gamma(x)} - \frac{\Gamma'(x)}{\Gamma(x)}^2$.

Method of Moments Estimators

Another method of estimating the parameters is with the method of moments.
The first and second moment of the beta distribution is

$$E(X) = \frac{\alpha}{\alpha+\beta} \tag{13}$$

$$E(X^2) = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)} \tag{14}$$

which means that the variance is

$$Var(X) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} \tag{15}$$

We can apprximatly set the sample mean $\bar{X} = 1/n\sum_{i=1}^{n}x_i$ and sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{X})^2$ to bea equal to the mean and variance

$$\bar{X} = \frac{\alpha}{\alpha+\beta} \tag{16}$$

$$S^2 = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)} \tag{17}$$

Rearranging this in terms of of α and β we get and estimation of the parameters

$$\alpha = \bar{X}\left(\frac{\bar{X}(1-\bar{X})}{S^2} - 1\right) \tag{18}$$

$$\beta = (1-\bar{X})\left(\frac{\bar{X}(1-\bar{X})}{S^2} - 1\right) \tag{19}$$

The method of moments is very straight forward but in dealing with small data sizes, which we will do, its not very accurate. What we will do is to use this method to get a good initial guess for the Newton-Raphson method[1].

Confidence Bound

The value we get with the estimators is usually not the true value. By using confidence bounds, we obtain a region within which these values are likely to occur a certain percentage of the time. This gives us measure of the usefulness tool of the data and the accuracy of the resulting estimates. This range of plausible values is called a confidence interval. The confidence region is given by the likelihood ratio bounds

$$-2\ln\left(\frac{L(\theta)}{L(\hat{\theta})}\right) \geq \chi^2_k \tag{20}$$

where $L(\theta)$ is the likelihood for unknown parameters, $L(\hat{\theta})$ is the likelihood at the estimated parameter and $\chi^2_{a;k}$ is the chi-squared statistic and k degrees of freedom of the parameter.

Results

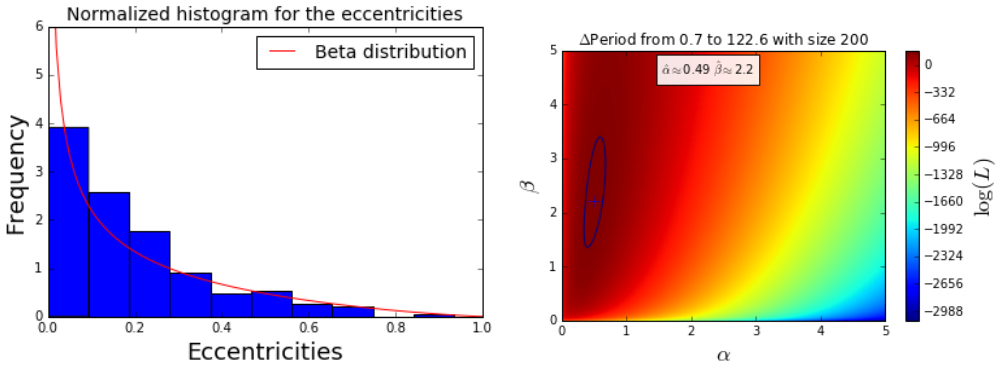

Figure 1: Figure to the right: A histogram over the eccentricities compared to a beta distribution with the estimated parameters. Figure to the left: A contour plot showing the value for the log-likelihood. The encircled region in blue shows the 90% confidence region and the plus marker is the position of the estimated parameters and is estimated to be $\hat{\alpha} \approx 0.49$ $\hat{\beta} \approx 0.2$.
For both figure an interval of orbital periods from 0.7 to 122.6 with 200 data points in between was used.

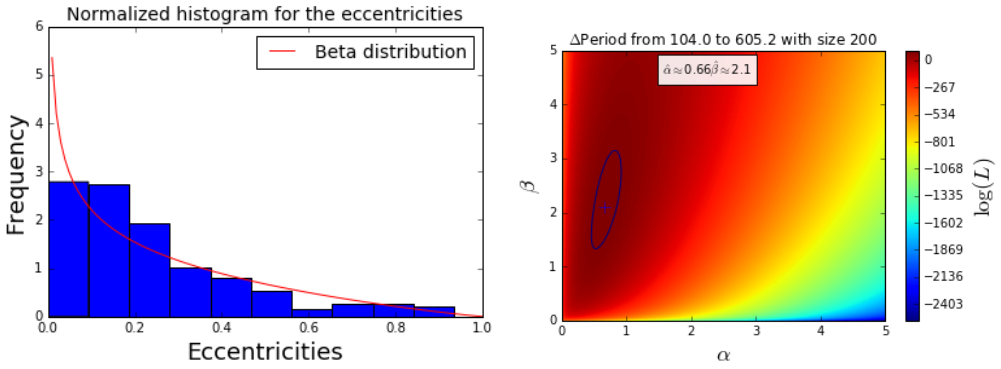

Figure 2: Figure to the right: A histogram over the eccentricities compared to a beta distribution with the estimated parameters. Figure to the left: A contour plot showing the value for the log-likelihood. The encircled region in blue shows the 90% confidence region and the plus marker is the position of the estimated parameters and is estimated to be $\hat{\alpha} \approx 0.66$ $\hat{\beta} \approx 0.21$.
For both figure an interval of orbital periods from 104 to 605.2 with 200 data points in between was used.

The result consists of 3 pairs of plots. The first plot, in each pair, is of a normalized histogram over the the eccentricities for the intervals [0.7,122.6], [104,605.2], [602,5000] respectively. There is also a curve for the beta distribution with the estimated shape parameters, $\hat{\alpha}, \hat{\beta}$, for reference. The other plot, in each pair, is of a contour plot of the value of the log-likelihood function over different values of the shape parameters. In this plot there is also marker where the estimate shape parameters lays. Around this point is the 90% confidence region.

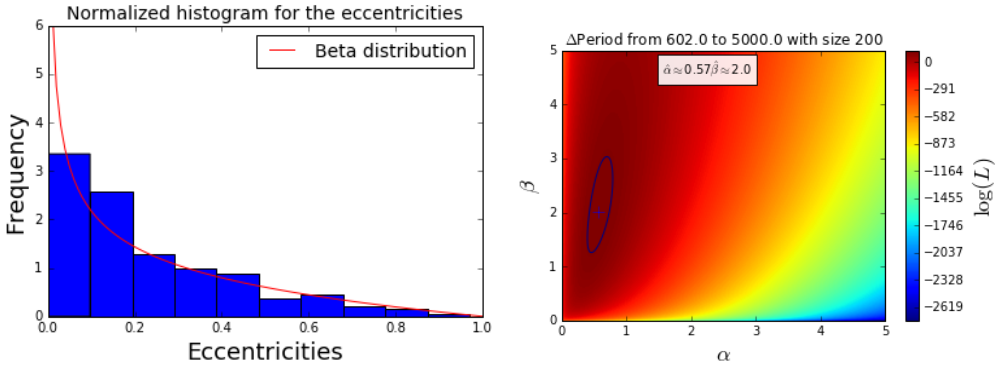

Figure 3: Figure to the right: A histogram over the eccentricities compared to a beta distribution with the estimated parameters. Figure to the left: A contour plot showing the value for the log-likelihood. The encircled region in blue shows the 90% confidence region and the plus marker is the position of the estimated parameters and is estimated to be $\hat{\alpha} \approx 0.57$ $\hat{\beta} \approx 0.2$.
For both figure an interval of orbital periods from 602 to 5000 with 200 data points in between was used.

Discussion

The histograms compared to the beta distribution with the estimated shape parameters in all the cases shows good a good agreement which indicates that the estimated shape parameters are not far from the true value. The same can be said for the contour plot where it is clear that the estimated values were in the region of the lowest value (the deep red region) and was always within the confidence region.
However in 2013 D. M. Kipping published a paper on 'Parametrizing the exoplanet eccentricity distribution with the Beta distribution' and found that $\hat{\alpha} = 0.867^{+0.044}_{-0.044}$ and $\hat{\beta} = 3.03^{+0.017}_{-0.16}$[2]. The reason for this large discrepancy is not clear. One source of error is that some data points are measured as zero. This is not very likely so a minimum of $e_{min} = 0.0001$ was used, since the data had a precision of that magnitude for some of the zero values.. This however should not impact the result significantly.
The different pairs of plot differed by their interval of orbital periods. All of them had the same number of data points but as we can see the ranges of these intervals are not the same. This is because the periods were not uniformly distributed. The further increases the accuracy of the result one could take into account this distribution of periods to interpolate some better e_{min} and maybe choose more clever intervals.

References

[1] C.B. Owen, 'Parameter Estimation for the Beta Distribution', Brigham Young University, Department of Statistics (2008)

[2] D.M. Kipping, 'Parametrizing the exoplanet eccentricity distribution with the Beta distribution', Harvard University, Department of Astronomy, (2013)