Understanding Planet Formation via Statistical Inference:

Characterizing Exoplanet Populations from the *Kepler* Space Science Mission

Megan Shabram

ABSTRACT

Results from the *Kepler* space science mission have exposed populations of planets that have orbital configuration and bulk densities quite different from the planets in our solar system. Ideas about how these populations have formed were developed to explain the formation of our solar system or “hot Jupiters”, and have yet to explain the plethora of “mini-Neptunes” and “super-Earths” detected by *Kepler*. To advance our understanding of planet formation, I propose the use of hierarchical Bayesian data analysis (HBA), as a framework for inferring population distributions of planetary system parameters, such as orbital eccentricity. With this implementation we will be able to search for new exoplanet populations and look for correlations between existing populations and their orbital and host star properties. This will allow us to address outstanding issues related to how planets form by matching dynamically evolved exoplanetary systems populations with information about the initial conditions of the system gained through knowledge of the host star properties.

1. The Importance of the Eccentricity Distribution and the Role of *Kepler*

Among the planetary system parameters of interest, the eccentricities of planetary orbits are a testimony to a planet’s dynamical history. Hence, constraining the eccentricity distribution for populations of planets is a way to directly probe planet formation mechanisms. NASA’s *Kepler* space science mission is invaluable for accomplishing this goal. *Kepler* is able to provide high precision broadband photometry for transit events where the light curves can be studied in extreme detail for many planetary systems. A transit event is the time dependant dimming of stellar light due to a planet passing between the star and the line of sight observation from Earth. The depths of these transit event light curves provide information about the radius of the planet relative the radius of the host star, allowing us to infer the planet’s size. Additionally, we can measure the transit duration, the ingress and egress duration, orbital period, and in some cases secondary eclipse. Combining these pieces of information using Kepler’s laws with information about the stellar mass and radius, we can learn about the orbital period, semi-major axis, the eccentricity of the planetary orbit, and the planet’s impact parameter (the distance between the center of the stellar disc and the chord that the planet traces as it crosses the stellar disc) (e.g. Winn 2010). This is described in the following relation:

$$D = \rho_{star}^{1/3}\sqrt{1-b^2}\left(a/r_{+}(e)\right)^{-1} \qquad (1)$$

where D is the transit duration, ρ_{star} is the stellar density, b is the impact parameter, a is the semi-major axis, and r_{+} is the distance of the planet from the star at the time of transit, which is a function of the orbital eccentricity, e. Additionally, multiple-transiting-planet systems in near-resonant orbital configurations (where orbital periods are related by the ratio of two small integers) are sensitive to variations in eccentricity. This can be detected in deviations of their orbits from strict periodicity, also known as transit timing variations. We can also use the information encoded in transit timing variations to obtain masses of planetary companions relative to the host star mass (e.g. Ford et al. 2012, Lithwick).

Constraining physical parameters, such as planet mass and eccentricity, for planets found by *Kepler* will provide opportunities to explore the implications of orbital parameters that are correlated with different planetary sub-populations and stellar properties. For instance, ***finding the eccentricity distribution as a function of planet size, stellar type, stellar metallicity, orbital period, effective temperature, and multiplicity can shed light on the connection between star formation and the formation of planetary companions.*** Essentially, the presence of more than one mode in the eccentricity distribution is indicative of more than one population and more than one planet formation scenario. If these modes are correlated with host star or planet properties, we can characterize them further and begin to paint a more detailed picture of the evolutionary processes that created these exoplanet populations. Additionally, this procedure can help disentangle otherwise degenerate formation scenarios. For example, the observed hot Jupiter population, consisting of heavily irradiated gas giants at small orbital separation (some on eccentric orbits) can be explained by planet-planet scattering (Rasio & Ford 1996, 2008) or gravitational instability at large orbital separation with migration (Boss 2000, migration ref). The objective of my research is to provide a more vigorous characterization of the current ensemble of exoplanets from *Kepler* and explore this parameter space to look for these types of informative correlations that can resolve current discrepancies in conflicting planet formation scenarios.

2. The Challenges of Characterizing the Eccentricity Distribution and The State of Exoplanet Population Analysis

Despite the remarkable photometric transit observations provided by *Kepler*, there are a number of outstanding issues that need to be addressed in order to obtain accurate constraints on the eccentricity distribution of planets around solar-like stars provided by *Kepler*. We are able to measure the planet radius relative to the host star radius with high precision, but uncertainty in stellar parameters for most stars ($T_{eff} \geq 5400K$) dominate the error budget for planet sizes and the eccentricity distribution (Moorhead et al. 2011), preventing a true realization of stand-alone planet properties. Additionally, measuring the impact parameter with accuracy is often difficult, and limits our ability to calculate the eccentricity directly from transit measurements alone (see equation 1). Moreover, even with the high SNR data from *Kepler*, interpreting orbital parameter distributions may still be limited by the sample size. Even more importantly, fitting for transit parameters for individual systems is non-linear, and the widely accepted reporting of best-fit parameter values is heavily biased when used to interpret information about the entire population. My research aims to address these limitations.

The focus of research being done to characterize the *Kepler* sample has largely been to constrain the occurrence rates of various planetary size regimes relative to the planets in our solar system. In particular, many studies have examined the occurrence rates of "hot Jupiters," "mini-Neptunes", and "super-Earths", and how these might correlate with orbital separation (e.g. Fressin et al. 2013, Howard et al. 2012). Other work has generated histograms of the number of planets in these mass regimes verses semi-major axis, orbital period, planet radius, and stellar effective temperature to learn about parameter distributions for these populations. However, these have mainly been first-pass analyses with the aim to catalog the *Kepler* objects of interest (Borucki et al. 2011). Furthermore, a prescription for using transit durations to constrain the eccentricity distribution for terrestrial planets was outlined in Ford et al. (2008) and executed for the *Kepler* sample in Moorhead et al. (2011), with conclusions limited by uncertainties in stellar radii for the hottest stars in the sample. Moorhead et al. (2011) found the distribution of transit durations is broader than predicted for planets on low eccentricity orbits. However, this discrepancy could be the result of incorrect assumptions about stellar densities. In short, to accurately constrain the eccentricity distribution for exoplanets with transit light curves from the *Kepler* data set, firm constraints on stellar densities are needed. To address this challenge, there have been follow-up observations to improve reported stellar parameters of planet host stars using both asteroseismology and high-resolution spectroscopic observations. This will allow an improved determination of stellar parameters using stellar evolution models such as the Yonsei-Yale stellar evolution model described in Demarque et al. 2004 and Batalha et al. 2011. Even with this improvement, significant uncertainties in stellar radius and density remain.

3. New Method in Light of these Challenges

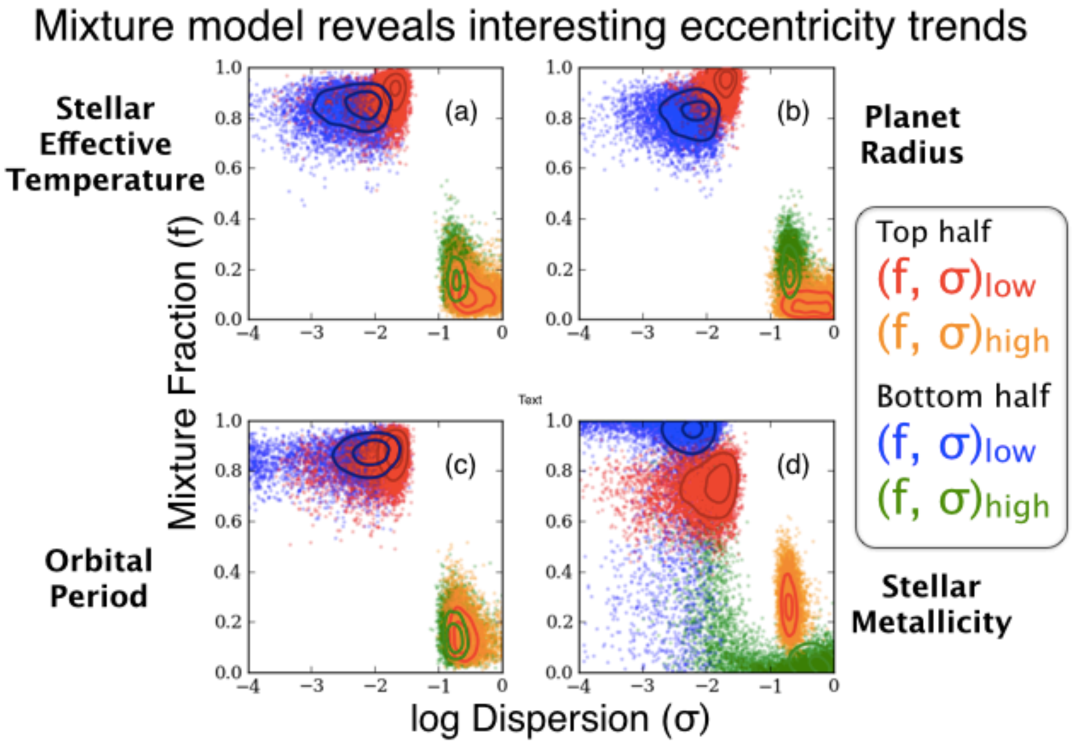
Accounting for biases in the sample is extremely important for inferring parameter distributions of entire planet populations. The cornerstone of my approach is the method of hierarchical Bayesian data analysis (HBA). The major advantage of this method is the incorporation of the full error distribution information of each measurement that goes into the calculation of the population's distribution, removing biases that can arise from using merely best-fit parameter values to obtain population distributions. With the advancement in computing resources and facilities, Bayesian parameter estimation for individual planetary systems has started to become common in the field of exoplanets. With the growing application of reporting probabilistic parameter estimates instead of best-fit values for extrasolar planets (e.g. Ford 2005, Eastman), the ability to infer more accurate and less biased parameter distributions for entire planet populations as a function of stellar and planetary properties has become a reality. This is because probabilistic parameter estimates are distributions that convolve the true parameter value with its uncertainty distribution, which may be unique to each measurement or observed planetary system. So far, the utilization of HBA has largely been overlooked in exoplanet population analysis to date (Hogg et al. 2010).

Currently, it is common practice to look at a histogram (possibly smoothed) of best-fit individual planetary system parameter measurements and deduce information about the population. This method is known to result in biased population parameters. HBA is a method that can infer the true posterior distribution for a population parameter by including the information contained in the measurement uncertainty. Hogg et al. (2010) have outlined this reasoning for the RV equation for star-companion systems. With their generated data set, they are able to prove that HBA outperforms simple histograms of best-fist parameters. Survey science is fundamentally statistical, as data are always measured within the limits of instrumental sensitivity, and HBA is a robust approach to accommodate this. We have generalized this approach to work with *Kepler* data.

There are many advantages to using the HBA method. ***With HBA, it is possible to get high precision results even if each object's parameters are measured with high uncertainty.*** This type analysis is increasingly necessary where results are combinations of large numbers of input parameters drawn from many different sources. I have already implemented an HBA model for the eccentricity distribution. As a first step, I focused on sub-populations that are particularly well constrained, in particular, a sample of hot Jupiters from *Kepler*. In the long run, I plan to implement this method for the entire *Kepler* sample.

In order to properly infer population distributions for extrasolar planets, analysis routines will need to include the following: **(1)** a robust statistical approach that can account for measurement uncertainty when analyzing population properties, in particular hierarchical Bayesian data analysis (HBA), **(2)** measurements of physical parameters from transit light curves and their uncertainties, and **(3)** stellar parameters of extrasolar planet host stars and realistic uncertainties.

With my initial application of hierarchical Bayesian data analysis, I am able to obtain robust models for the eccentricity distribution from a limited, yet well constrained, sample (B.-O. Demory *in prep*). As a second step in a comprehensive analysis of the *Kepler* exoplanets, I have looked at how the eccentricity distribution correlates with planet radius, orbital period and stellar effective temperatures.



The figure above shows the posterior probability distribution for the dispersion, σ, of a mixture of two normal distributions (centered at zero and truncated at 1) and the fraction of planets that come from each mixture component, *f*. We test if the data provide evidence for two populations in the sample by comparing one and two component Gaussian mixture models. We use MCMC to sample from our hierarchical model and find that a mixture of two Gaussian is favored by that data and see evidence of two population in panels (a), (b), and (c) (The first cluster is composed of red and blue, and the second cluster is composed of orange and green). Additionally we have divided our sample in half by high and low values (to maintain statistical power). We apply a hierarchical Bayesian model to small (blue and green) and large (red and orange) halves of the *Kepler* Hot Jupter occultation data, sorted by (a) stellar effective temperature, (b) planet radius, (c) orbital period, and (d) stellar metallicity. The blue and red points represent samples of the posterior distribution of σ_{low} and f_{low}, the dispersions for a two-component mixture of Gaussians and their associated mixture fractions for the smaller half of the sorted data. The orange and green points represent samples of the posterior distribution of σ_{high} and f_{high}, the dispersions for a two-component mixture of Gaussians and their associated mixture fractions for the larger half of the sorted data. The data are plotted with the vertical axis representing the mixture fraction and the horizontal axis representing the dispersion. The contours plotted over the sampled posterior represent multi-variate kernel density estimates (Shabram and Ford 2014, *in prep*). Now draw your attention to panel (d), where the data is sorted by metallicity. The bottom half of the sorted data (blue and green) show mixture fractions that are consistent with 1 and 0, which means that this sample is better fit with a one component model. The larger half of the sample (red and orange) however is still well fit using a two component model. Therefore we find that higher metallicity host star systems are more dynamically complex than lower metallcity star systems. Dawson reference here ***Applying the HBA method to the eccentricity distribution of a small sample of well constrained hot Jupiters is the first step in my project to correlate orbital system parameters and stellar host star properties with populations of planets from the entire Kepler sample, where the uncertainties are, for the most part, large.*** In short, I have begun my studies with the focus on understanding the eccentricity distribution of planet sub-populations. I have demonstrated the method to be able to directly constrain eccentricities for both individual planetary systems and the population of planets. Studying the role of eccentricity is the first in my plan to study the role of multiple parameters in the planet formation process.

4. The Significance of this Research and the NASA Objectives

In summary, I have created the framework that will allow us to uncover new sub-classes of planet populations and place constraints on planet formation mechanisms such as core accretion, gravitational instability, planet-planet scattering and migration. By exploring unbiased parameter distributions as a function of stellar properties, environment, and planetary mass radius regimes, there is the potential to revolutionize the classification of exoplanetary systems. Also, it will be possible to incorporate data from future transit survey missions, such as NASA's small explorer program's Transiting Exoplanet Survey Satellite (TESS), in a self-consistent way. ***Thus, my research directly supports NASA's present and future astrophysics missions.***

There are many other questions that this research can help answer. After further exploring the eccentricity distribution of the *Kepler* sample, I will extend this method to explore the distribution of other planetary system parameters such as planet size, mass, period, and semi-major axis. I also plan to look for correlations of these parameters with stellar parameters to connect star and planet formation processes. Additionally, I will look at planetary systems in near-resonant configurations and compare them to systems that are not in order to look for trends that may expose unique formation histories. By applying the HBA method to models with several population parameters it is possible to deconstruct the processes of planet formation and evolution. Understanding these relationships is important for cosmo-chemistry, the field of extrasolar planets, understanding the formation of our solar system, and astrobiology. My research is also grounds for advancing other technical fields.

The trajectory of my research has the potential to be very broad. In addition to understanding planet formation mechanisms and classifying planetary populations, I am also interested in applying our method to exoplanet composition and structure. As a postdoc I hope to connect the study of orbital dynamics to the study of exoplanet characterization, so as to qualify degenerate composition and structure models for reported bulk densities of planets by making the case of the relationship between the rigidity of a planets interior, the effect of tides, and its orbital parameters. Understanding the composition and structure of exoplanetary interiors will allow detailed comparison with our solar system planets, and address big-picture questions such as: how common are areas that are suitable for Earth-like life? We will never know the true character of these planets unless we remove this degeneracy. An NESSF would support my effort to generalize HBA to study (XYZ, TTV, EB near here) in particular and to lay the groundwork for future HBA studies addressing increasingly ambitious questions.

By characterizing the formation and evolution of planetary system populations, my proposed research will help to *pioneer a future in space exploration* beyond our solar system while *advancing scientific discovery,* which is congruent with the NASA mission objectives. Additionally, my research aims to "*Advance scientific knowledge of the origin and evolution of the solar system, and the potential for life elsewhere*" as well as "*search for Earth-like planets*" which is in direct alignment with the science objectives of the NASA Science Mission Directorate (SMD).

Add section about relevance to NASA objectives. (page 8 of solicitation: break down sentence one by one and explain).

- Academic qualification of student

ask eric to add stuff about proposal to letter if needed.

-**Quality of proposed research** (How well thought through is this, details that show you have thought about it and its well thought through). Lay it out in space I have so that someone from a different area can appreciate the connection. Don't expect them to infer what we do is relevant and important. Address each of things, make it easy. For example, how does it test theories of planet formation, see dawson et al. for and example. Higher ecc. Systems wont have fully damped so we can see more systems around higher metallicties stars that have significant eccentricities.

-Relevance to NASA priorities

relevance to Kepler, future K@ TESS, other large surveys. List the ones other than Kepler, can be non NASA, (GAIA)., PLATO

- Soundness of approach or feasibility (for first test we will have a sample size of this and it will have a power of this. We have done preliminary analysis of preliminary data sets and we are able to distinguish between models for data sets of this size. We did something and we can tell you something meaningful. In the future we will use a larger data set and we will get that data set from

- Understanding of research area (keep references regular, cite people other than our own group,

- REFERENCES -

Borucki et al. 2011 ApJ 728 117
Boss, Alan P., 2000, ApJ, 536, 2, L101-L104
Carter, J., et al. 2008, ApJ, 689, 499
Demarque et al. 2004 ApJS, 155, 2, 667-674
Ford, E. B., 2005, ApJ, 129, 3, 1706-1717
Ford, E. B., et al. 2008, ApJ, 678, 1407
Ford, E. B., & Rasio, F. A. 2008, ApJ, 686, 621
Ford, E. B., et al. 2012, ApJ, 756, 185
Fressin, F., et al. 2013, ArXiv e-prints 1301.0842
Hogg, D. W., et al. 2010, ApJ, 725, 2166
Howard, A. W., et al. 2012, ApJS, 201, 15
Moorhead, A. V., et al. 2011, ApJS, 197, 1
Rasio, F., & Ford, E. B. 1996, Science, 274, 954
Winn, J., 2010, eprint arXiv:1001.2010