Constraining Stellar Multiplicity with Approximate Bayesian Computation

Dietrich College Honors Thesis
April 2016

H. Eric Alpert
Carnegie Mellon University

Advisors:
Peter Freeman[1]
Carles Badenes[2]
Chad Schafer[1]

1 Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213
2 Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, PA 15260

Abstract

Since the beginning of modern astronomy, we have known of the existence of binary-star and multiple-star systems. However, we do not know the stellar multiplicity fraction, i.e., the proportion of stellar systems with two or more stars. Astronomers have argued that constraining the value of the multiplicity fraction will improve our theoretical understanding of the Universe. From understanding the birth and evolution of stars to improving the calibration of observational methods, constraining multiplicity will have profound impacts on the field of astrophysics. In this project we implement the advanced statistical algorithm Approximate Bayesian Approximation (ABC) to constrain the value of the stellar multiplicity fraction. The ABC algorithm uses a data simulation to derive a Bayesian posterior distribution without the calculation of the likelihood. We employ the Apache Point Observatory Galactic Evolution (APOGEE) data set of 97,313 stellar objects and the Apache Point Observatory and Kepler Astroseismology Science Constortium (APOKASC) data set of 1,916 stellar objects. We develop and test four forward models, of increasing complexity, to simulate the data-generating process as a function of multiplicity. Using the software package **`cosmoABC`**, our forward model and the APOGEE catalog, we derive a posterior distribution of stellar multiplicity. Using our third forward model, we constrain the value of the stellar multiplicity fraction with a 95% credible interval to 0.555 $\pm$ 0.051.

Acknowledgements

I wish to thank my thesis adviser, Dr. Peter Freeman, who introduced me to this fascinating topic. Through his constant guidance and support, he has displayed great commitment towards helping me succeed. It was an absolute pleasure and a great honor to work with him.

I am grateful to Dr. Carles Badenes and Dr. Chad Shafer, who have shared their expertise in their respective fields of astrophysics and statistics.

I would like to thank the faculty and staff of the Department of Statistics for providing me with a world-class training in statistics. I would also like to thank the faculty and staff of Carnegie Mellon University for creating an amazing and memorable undergraduate experience.

Most importantly, I wish to dedicate this thesis to my family. I could not have made it to where I am today without their unwavering love and support. They mean the world to me.

Contents

1 Introduction

Understanding the interaction between stars within a multiple-star system allows profound insights into our knowledge of the Universe. One key to improving our understanding is constraining the value of the stellar multiplicity fraction, the proportion of stellar systems with multiple stars. For example, observational constraints on stellar multiplicity have already challenged a predominant theory of binary system formation. For nearly a decade, the “capture" theory was the leading one explaining binary star formation (Shu et al., 1987). This theory states that stars form as single bodies through the collapse of gravitationally unstable pre-stellar cores. A star then gravitationally captures another star to form a binary system. However, early empirical evidence suggested that the stellar multiplicity fraction is at least 0.5 for pre-main sequence star populations (Mathieu, 1994). This result, combined with simulation studies that suggested the gravitational capture of a stellar body is inefficient and a rare occurrence (Tohline, 2002), contradicts the capture theory, and has led to the development of new theories.

The current leading theory for multiple star system formation invokes a fragmentation process in which the prestellar core breaks up into smaller fragments. The fragments then evolve into stars within a multiple star system (Mathieu, 1994). The steep curve of multiplicity as a function of mass suggests that cores with more mass tend to fragment into more pieces. There is also evidence that suggests solar-type stars are components of higher-ordered systems more often than low-mass stars. This result further supports a fragmentation mechanism for multiple-star system formation. As more information about stellar multiplicity is discovered, astronomers will be able to further improve on our knowledge of stellar formation.

In this project we will apply a newly developed statistical algorithm, Approximate Bayesian Computation (ABC; Pritchard et al., 1999), to constrain the value of the stellar multiplicity fraction. ABC uses a data simulation to derive a Bayesian posterior distribution without the calculation of the likelihood function. We develop a forward model to simulate data for a population of solar-type stellar systems as a function of stellar multiplicity with Monte Carlo simulations. We then utilize the software package **`cosmoabc`** (Ishida et al., 2015) to implement ABC to compare our simulated data to a survey of 97,313 objects from the APOGEE data set (Majewski et al., 2015). Through this process we constrain the value of stellar multiplicity for solar-type stars. We also explore how the stellar multiplicity varies as we add complexity to our forward models. We run several analyses with different forward models, increasing complexity by changing the assumptions modeling the physical systems. We conclude by discussing the advantages and disadvantages of our procedure.

1.1 Multiple-Star Systems

Our knowledge of stellar multiplicity is limited by our ability to efficiently detect multiple-star systems. Such systems are generally detected through the use of one or more of three methods and binary systems are classified by the method used to detect them.

Visual binary systems are ones where the component stars are bright and distinct enough to be detected individually. Visual binaries are hard to detect because they must be close to the Earth. As the distance between the Earth and the system increases, the apparent brightnesses of the stars decrease, and the angles between the component stars. Eventually we only observe a single point of light.

Spectroscopic binary systems are detected through spectroscopy. Spectroscopy breaks down the visual light of the stars into a spectrum, such as when light passes through a prism. As two stars in a multiple-star system orbit around each other, their distances from Earth shift slightly. As an object moves toward the Earth, its light waves compress due to a Doppler shift and observable features in the object's spectrum shift toward the blue end. Similarly, as an object moves away from the Earth its spectral features shift to the red end. Periodic shifts in the spectrum of a star system allow astronomers to infer the presence of multiple-stars.

Eclipsing binaries are detected by tracking the brightness of star systems. If a star passes in front of another star during its orbit, the brightness of the star system decreases. Astronomers track the luminosity of star systems over time and attempt to detect periodic dips in the luminosity to identify multiple-star systems. Eclipsing binaries are also rare because the inclination of the binary star's orbital plane relative to the Earth must be just right to ensure that an eclipse occurs along our line of sight.

Note that all three methods of detection require many followup observations so as to identify a periodic pattern of orbiting stars. This requirement makes it time consuming to assemble a large catalog of multiple-star systems. Our knowledge of empirical stellar multiplicity is also biased because these methods preferentially detect short-period systems. It is important to note that because we apply Approximate Bayesian Computation to analyze an entire population of observed stars, we can infer a stellar multiplicity fraction without having to detect any multiple-star systems within the population.

1.2 Stellar Multiplicity

Duchêne & Kraus (2013) summarize what astronomers currently know about stellar multiplicity. We know that about 25% of solar-type multiple-star systems are higher-ordered systems composed of three or more stars. These higher-ordered systems are typically hierarchical, meaning that they are composed of binary- or single-star subsystems. Within a higher-ordered system the ratio

between the longest period and the shortest period is $P_{Long}/P_{Short} \geq 5$. Also the mass ratios ($q = \frac{M_{Small}}{M_{Large}}$) for short-period subsystems are typically flat up to $q \approx 0.9$ with a mode at $q \approx 1$. The large-period subsystems typically have $q \leq 0.55$.

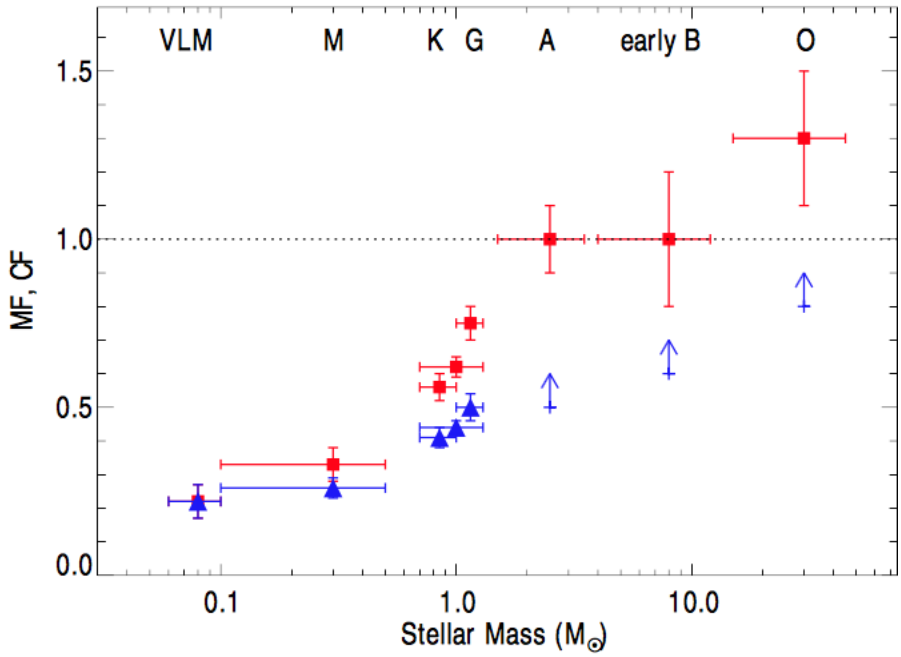

Figure 1: Distribution of f_m as a function of mass. (Reproduced from Duchêne & Kraus, 2013) The blue triangles represent the stellar multiplicity and the red squares represent the companion frequency, the ratio between the number of stars and the number of systems.

Astronomers also know that stellar multiplicity as a function of the primary mass is steep. A unit for the mass of a star is solar mass ($M_\odot$), the mass of the sun. Figure 1 shows the distribution of mass for each category of stellar masses. In particular, supersolar stars with mass between 1 and $1.3M_\odot$ have a multiplicity of $f_m = 50 \pm 4\%$. Also, stars with mass between 1.5 and 5 $M_\odot$ the stellar multiplicity is $f_m \geq 50\%$.

All previous studies of stellar multiplicity are based on observational data of multiplicity. This means that these studies use small sample sizes. Also these studies are subject to selection bias because most detected binary systems are short-period systems. In this project we use a large sample of about 100,000 objects and do not require classification of the type of system. Therefore, the possible systematic errors of the other studies do not apply to this one.

1.3 Significance

Stellar multiplicity impacts many problems in astrophysics. As previously mentioned, stellar multiplicity lies at the heart of stellar formation. Stellar multiplic-

ity also is important in models of stellar evolution. The proportion of multiple-star systems is an important parameter in simulated models of stellar evolution (Paxton et al., 2015).

Many interesting events occur when stars within a system interact. Around each star is a region where the gravitational force of the star exceeds the gravitational forces of its companion. This region is called the star's Roche lobe. If surface of a star extends pass the Roche lobe, that star will transfer matter to its companion. Mass transfer usually occurs when the larger star in a system finishes its main sequence (hydrogen burning) phase and expands in volume to being a giant star. When mass transfer occurs the system begins a common-envelope phase (Schreiber & Gaensicke, 2003). Following the common-envelope phase the mass donor star becomes a white dwarf star and the system becomes a cataclysmic variable. White dwarfs are very dense because they are about the size of the Earth while having masses $\gtrsim 0.5 M_{\odot}$. The cataclysmic variable could then evolve into a Type Ia supernova or X-ray binary system. Studying the birth rates of these objects requires knowing the stellar multiplicity fraction.

2 Approximate Bayesian Computation

Astronomers often use Bayesian methods to constrain the values of astronomical parameters in astronomical models. Bayesian statistics utilizes Bayes' Theorem in deriving the posterior distribution of a vector of parameters θ given data $\mathcal{D}$ as

$$\pi(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)\pi(\theta)}{P(\mathcal{D})}. \tag{1}$$

The calculation requires a prior distribution, $\pi(\theta)$, the likelihood function $P(\mathcal{D}|\theta)$, and a normalizing constant for a given data set, $P(\mathcal{D})$.

The prior distribution, $\pi(\theta)$, reflects our prior belief about the parameters' distributions of possible values. Astronomers often define the prior distribution as an astrophysical model derived from theory or empirical evidence. Using such a model allows the astronomer to update the model given new data, an approach unique to Bayesian theory. This approach further constrains the values of the parameters and improves the model.

The likelihood function, $P(\mathcal{D}|\theta)$, measures how likely an observer will observe the data $\mathcal{D}$ given the values of the parameters of interest θ. It can be the case that a likelihood function is not analytically expressible. In these cases a likelihood-free method of calculating the posterior distribution must be used. Approximate Bayesian Computation (ABC) is a new and powerful method for likelihood-free inference.

First implemented in 1998 by Pritchard et al., scientists use ABC in fields ranging from computational biology (Csilléry et al., 2010) to psychology (Turner & Zandt, 2012). Only recently have astronomers begun using ABC, for instance to estimate the luminosity function (Schafer & Freeman, 2012), study morpholog-

ical transformations of galaxies (Cameron & Pettitt, 2012), constrain estimates of cosmological parameters (Weyant et al., 2013), and constrain disk formation of the Milky Way (Robin et al., 2014). Furthermore, Ishida et al. (2015) and Akeret et al. (2015) have both published ABC software packages geared towards astronomical data analysis.

2.1 ABC Algorithm

The ABC algorithm features four elements: a forward model to generate simulated data, a prior distribution $p(\theta)$, a distance function between two distributions $\rho(\mathcal{D}_1, \mathcal{D}_2)$, and a sampling algorithm to update the prior distribution.

The ABC algorithm samples a parameter value θ^* from the prior $\pi(\theta)$. Using θ^* we simulate a data set $\mathcal{D}_s$ and calculate the distance $\rho(\mathcal{D}, \mathcal{D}_s)$ between the observed and simulated data. We accept θ^* if $\rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon$ for some defined tolerance $\epsilon > 0$. The underlying principle of ABC is that as $\epsilon \to 0$,

$$P(\theta|\rho(\mathcal{D}, \mathcal{D}_s) \leq \epsilon) \xrightarrow{d} \pi(\theta|\mathcal{D}). \tag{2}$$

The unique element of the ABC algorithm is using a distance metric ρ to compare the distributions between observed and simulated data. To improve the efficiency of the ABC algorithm, the comparison often occurs between summary statistics of the observed and simulated data, $\rho = \rho(s(\mathcal{D}), s(\mathcal{D}_s))$. The summary statistic should preserve the necessary information to constrain the parameters. Ideally the statistic is a sufficient statistic meaning that $p(\mathcal{D}|s, \theta)$ is not a function of θ. Fulfilling this condition ensures that the posterior distribution given the statistic will equal the posterior given the data, $\pi(\theta|s(\mathcal{D})) = \pi(\theta|\mathcal{D})$. Much research focuses on deriving approximations for sufficient statistics and deriving summary statistics that preserve necessary information (Weyant et al., 2013).

Many sampling algorithms are available for ABC calculations. One of the earliest proposed methods was the rejection algorithm outlined in Algorithm 2.1(Akeret et al., 2015). Rejection sampling becomes inefficient as $\epsilon \to 0$ because the rejection rate increases towards infinity. To improve the sampling efficiency of ABC, Markov Chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC) methods are used (see Weyant et al., 2013). for details of these sampling algorithms). In this project I use population Monte Carlo (PMC) which is based on SMC (see Algorithm 2.2).

2.2 Population Monte Carlo ABC

We outline the algorithm implemented in **`cosmoABC`** in Algorithm 2.2. In the first step of PMC ABC, the algorithm samples M values called "particles" from the prior distribution, i.e., $\theta_i \sim \pi(\theta)$ for $i \in [M]$. Here $M > N$, where N is the sample size of the posterior distribution. We simulate a data set, $\mathcal{D}_{s,i}$, using

Algorithm 2.1 Rejection Sampling ABC

```
N ← Sample Size
ϵ ← Rejection Criteria
i ← 0
while i < N do
    Draw θ* from prior π(θ)
    With θ* generate D*
    if ρ(D, D*) ≤ ϵ then
        Store θ*
        i ← i + 1
    end if
end while
```

the forward model for each particle and then calculate its distance from the observed data $\rho_i = \rho(\mathcal{D}, \mathcal{D}_{s,i})$. We accept $\theta_1, \theta_2, ..., \theta_N$, where θ_i is the datum with the i^{th} smallest distance. These N particles are the first particle system, $S_{t=0}$. We then assign the particles weights $W_0^j = 1/N$ with $j \in [N]$, and define the distance threshold $\epsilon_{t=1}$ as the p^{th} quantile of the distances, where $p \in [0, 1]$ is a user-defined quantile. The initial iteration concludes by calculating the sample covariance matrix C_0 from $S_{t=0}$.

In subsequent iterations of the PMC algorithm we use importance sampling to update the prior distribution $p(\theta|S_{t-1}, W_{t-1})$. We draw particles from the distribution of $S_{t=0}$ with weights $W_{t=0}$, $\theta^* \sim p(\theta|S_{t-1}, W_{t-1})$. We accept θ^* if $\rho^* \leq \epsilon_t$. We repeat this sampling until N particles are accepted. For each particle system S_t, for $t \geq 1$, the weights are defined as

$$W_t^j = \frac{\pi(\theta_t^j)}{\sum_{i=1}^{N} W_{t-1}^i p(\theta_t^j|\theta_{t-1}^i, C_{t-1})}, \tag{3}$$

where $p(\theta_t^j|\theta_{t-1}^i, C_{t-1})$ is a normal distribution pdf centered around θ_{t-1}^i with variance C_{t-1}.

This process iterates until the algorithm reaches the convergence criterion, $N/K \leq \Delta$, where K is number of draws required to draw an accepted particle and Δ is a user-defined convergence criterion. See Algorithm 2.2 for details.

2.3 Distance Estimation: Kullback-Leibler Divergence

Within our analyses we implement an empirical Kullback-Leibler (KL) divergence estimation to determine the distance between two distributions. The KL diver-

Algorithm 2.2 PMC ABC implemented in **`cosmoABC`** (Ishida et al., 2015)

$\mathcal{D} \leftarrow$ Observed Data
$t \leftarrow 0$
$K \leftarrow M$
for i = 1, . . ., M **do**
    Draw θ^* from prior $\pi(\theta)$
    With θ^* generate $\mathcal{D}^*$
    Calculated distance $\rho^* = \rho(\mathcal{D}, \mathcal{D}^*)$
    Store $S_{init} \leftarrow \{\theta^*, \rho^*\}$
end for
Sort elements in S_{init} by ρ
Keep the N values of θ^* with lowest distances in $S_{t=0}$
$C_{t=0} \leftarrow$covariance matrix of S_0
for $L = 1, \ldots, N$ **do**
    $W_1^L \leftarrow 1/N$
end for
while $N/K > \Delta$ **do**
    K $\leftarrow$ 0
    t $\leftarrow$ t+1
    $S_t \leftarrow$ []
    $\epsilon_t \leftarrow 25^{th}$-quantile of distances in S_{t-1}
    while len(S_t) $< N$ **do**
        K $\leftarrow$ K + 1
        Draw θ_0 from S_{t-1} with weights W_{t-1}
        Draw θ^* from $N(\theta_0, C_{t-1})$
        With θ^* generate $\mathcal{D}^*$
        Calculate distance $\rho = \rho(\mathcal{D}, \mathcal{D}^*)$
        if $\rho \leq \epsilon_t$ **then**
            $S_t \leftarrow \{\theta^*, \rho, K\}$
            $K \leftarrow 0$
        end if
    end while
    for J = 1,. . ., N **do**
        $W_t^J \leftarrow$ equation (3)
    end for
    $W_t \leftarrow$ Normalized weights
    $C_t \leftarrow$ weighted covarience matrix from $\{S_t, W_t\}$
end while

Parameter	Description
$\mathfrak{D}$	Observed data
$\mathfrak{D}_s$	Simulated data
M	Number of particles in first iteration
S	Particle System
N	Number of particles in S
t	Iteration index
K	Number of draws index
W	Importance weights
ϵ	Distance threshold
Δ	Convergence criterion
θ	Model parameters

Table 1: Parameter descriptions for Algorithm 2.2.

gence between two continuous probability distributions P and Q with probability density functions $p(x)$ and $q(x)$ is defined as

$$D(P||Q) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx.$$

The KL divergence by definition is also the expected log-likelihood ratio of P and Q.

We must note that the KL divergence is not a true distance metric because generally $D(P||Q) \neq D(Q||P)$ and also because it does not satisfy the triangle equality. However, its properties are ideal for comparing two continuous distributions. The divergence between P and Q is only zero if and only if $P = 0$. Also the divergence increases when the values of $p(x)$ and $q(x)$ diverge from each other. We also choose to use this metric because as we shall see below, it best captures the differences within the long tail of the response variable in the stellar multiplicity analysis compared to other commonly used distance measures.

3 Data

Within out analysis we use two data sets of stellar objects. The Apache Point Observatory Galactic Evolution Experiment (APOGEE) provides spectroscopic information for over 150,000 stellar objects. The joint project between the Apache Point Observatory and the Kepler Astroseismology Science Consortium (APOKASC) provides astroseismic analysis on 1,916 stellar objects.

3.1 APOGEE Catalog

To constrain the stellar multiplicity fraction, we make use of the Apache Point Observatory Galactic Evolution Experiment Data Release 12 (APOGEE; Majewski et al., 2015). The project has collected over 500,000 spectra for over 150,000 stellar objects within the Milky Way galaxy. Using these spectra, the project has derived and cataloged stellar parameters such as the stellar effective temperature (T_{eff}), luminosity (L), surface gravity ($\log g$), and measures of metallicity, the proportion of elements heavier than helium.

A key characteristic of the APOGEE catalog is its temporal component. Each star is the subject of multiple observations. The catalog reports the time and the stellar radial velocity v_r for each observation, i.e., the velocity at which the star is moving away from the Earth. Given these measurements astronomers calculate the maximum change in radial velocity ΔRV_{max}:

$$\Delta RV_{max} = \max v_v - \min v_r. \tag{4}$$

The APOGEE data pipeline classifies the quality of each observation. We filter the catalog and keep all stellar objects classified as "good" with at least two individual observations. The final data set we use for our analyses contains 97,313 stellar systems.

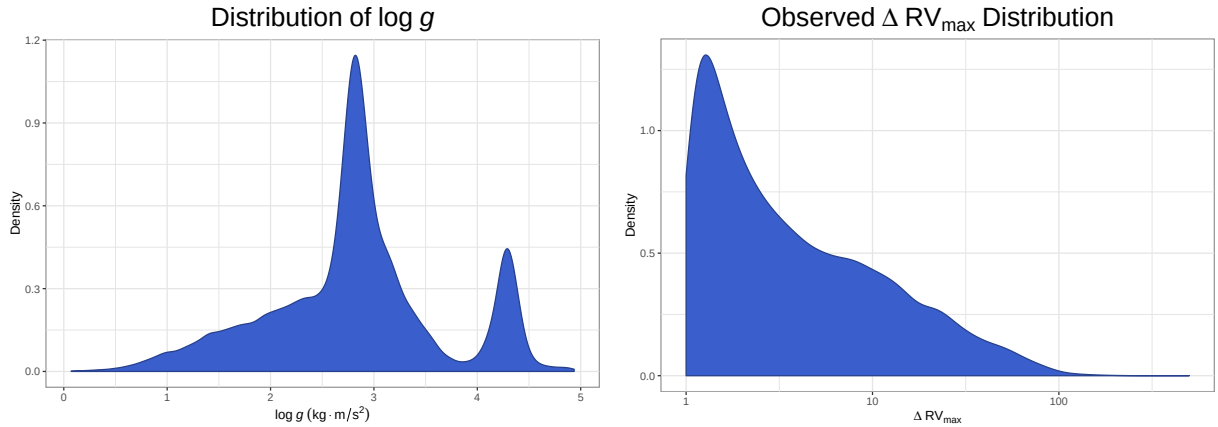

Figure 2: Left: Empirical distribution of $\log g$. Right: Empirical distribution of ΔRV_{max}

3.2 APOKASC Catalog

We also used the distribution of masses provided in the APOKASC catalog (Pinsonneault et al., 2014). The APOGEE project provides spectroscopic analysis of the stellar objects and the Kepler Astroseismology Science Consortium (KASC) provide astroseismic analysis, adding estimates of stellar mass. The catalog contains data on 1916 stellar objects. Figure 3 shows the distribution of the masses. The stellar objects are mostly red giants with masses between 0.8 and 3.2 $M_{\odot}$.

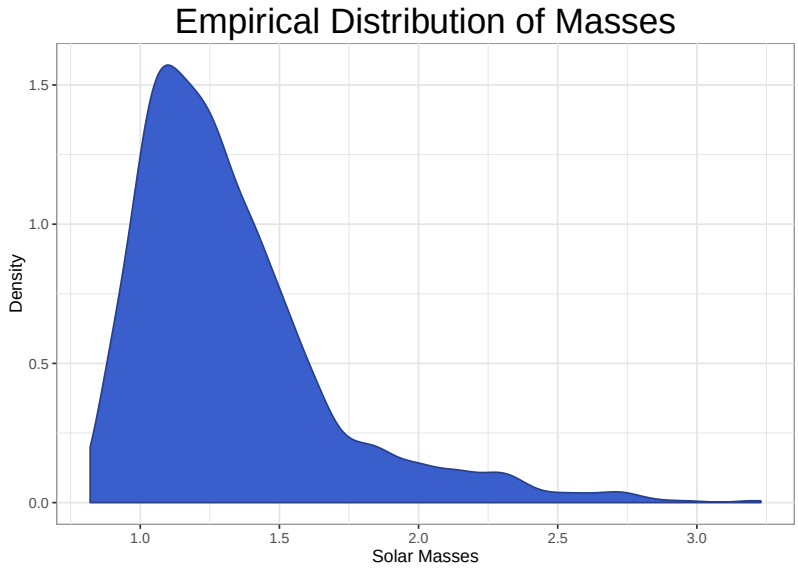

Figure 3: Empirical distribution of masses from APOKASC (Pinsonneault et al., 2014).

4 Data Simulation

We develop a forward model that takes a stellar multiplicity value, f_m as an input and returns a vector of predicted ΔRV_{max} values with size n_{sys} = 97,313, the size of the APOGEE catalog. We create four versions of the forward model to accommodate changes to our assumptions of the physical processes.

Parameter	Definition
f_m	Stellar multiplicity fraction
n_{sys}	Number of stellar systems within a simulated data set
X_{bin}	Indicator variable determining if system in binary
i	Inclination
q	Mass ratio
ω	Argument of pericenter
m	Primary mass
$\log g$	$\log_{10}$g
a	Orbital axis lengths
p	Periods
e	Eccentricity
$v_r(t)$	Radial velocity of system at time t
ΔRV_{max}	Maximum change in radial velocities
$\epsilon_{\Delta RV_{max}}$	Error of ΔRV_{max}

Table 2: Table of variables used in models with definitions.

4.1 Model 1

The algorithm begins by using the input parameter f_m to determine the number of binary stars within the simulated catalog. We define $X_{bin,\,i} \sim Bernoulli(f_m)$ for $i \in [n_{sys}]$ where $X_{bin,\,i} = 1$ if simulated stellar system i is binary and 0 if not. By definition $\sum_{i=1}^{n_{sys}} x_{bin,\,i} \sim Binomial(f_m, n_{sys})$.

The mass ratio for a system is the ratio between the secondary and primary star within a stellar system, i.e., $q = \frac{M_2}{M_1}$. Following Raghavan et al. (2010) we sample the mass ratio for system i as $q_i \sim Uniform(0.1, 1)$.

The inclination of a stellar system is the angle between the reference plane (the plane containing the primary star and the Earth) and the orbital plane of that system. See Figure 4. For system i the inclination is defined as $I_i \sim Uniform(0, \pi)$.

The argument of pericenter determines the angle from the vector pointing towards the pericenter, and the ascending node, the line where the orbital plane and reference plane intersect. We define the argument of pericenter for system i as $\omega_i \sim Uniform(-\pi, \pi)$.

We perform rejection sampling to generate values for orbital semimajor axes in astronomical units (AU). We first used least-squares density estimation to fit a Gaussian distribution to the histogram of orbital axes in Figure 13 of Raghavan et al. (2010). We determined that the orbital axes followed a distribution such that $\log_{10}(a) \sim N(1.533, 1.713)$. We also define the minimum orbital axis as $a_{min} = 1.5 r_\odot$. We draw a value from this distribution and compare it to the minimum value. If the drawn value is smaller than the minimum we reject the

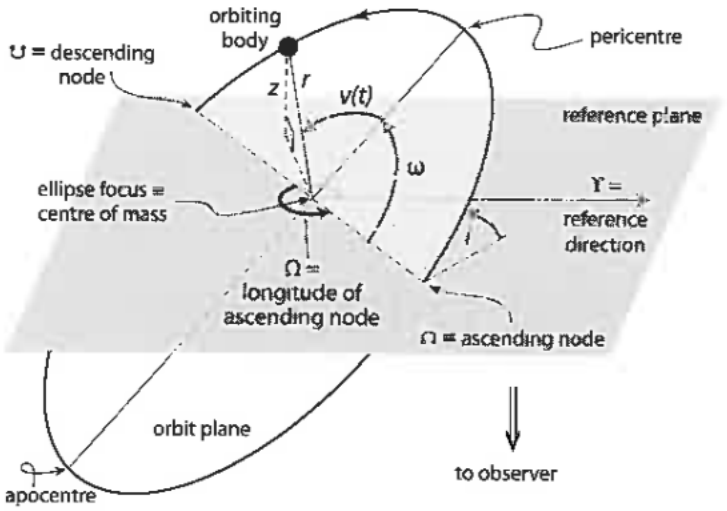

Figure 4: Diagram displaying definitions of the celestial mechanics simulated in model. (Reproduced from Perryman, 2011) The inclination is the angle between the reference plane and the orbital plane. The argument of pericenter is the angle on the orbital plane between the ascending node and the pericenter.

value and draw again. Algorithm 4.1 outlines the process of generating a value for an orbital axis length.

Algorithm 4.1 Rejection sampling of orbital axis lengths.

while No accepted a **do**
  Generate $a_{temp} \sim N(1.533, 1.713)$
  Calculate $a_{\min}(a_{temp}, m, \log_{10}(g))$
  if $a_{\min} < a_{temp}$ **then**
    Accept a_{temp}
  end if
end while

We then use Kepler's Third Law to calculate the period given the orbital semi-major axis length a, primary mass m and mass ratio q:

$$\frac{p^2}{a^3} = \frac{4\pi^2}{Gm(1+q)}. \tag{5}$$

We next calculate the radial velocities. For each system in the APOGEE catalog, data multiple observations were collected. The data set includes the time of each observation. We let $t_{i,j}$ be the time between 1^{st} and j^{th} observation for the i^{th} system. By definition $t_{i,1} = 0$. Then the radial velocities for system i at each time t is

$$v_r(t) = \frac{kqa}{1+q}\sin(i)\sin(\omega + kt), \tag{6}$$

where $k = 2\pi/p$. Note that the model generates a vector of radial velocities for system i of size n_i where n_i is the number of observations that system in the APOGEE catalog.

The distribution of $\Delta RV_{\max}$ contains a main core of values followed by a long right tail as seen in Figure 2. The core of the distribution represents random noise surrounding a value of zero for single-star systems. In a single-star system the measurements of radial velocity should be very small because only binary companions can create a large radial velocity. We fit a Gaussian distribution around this core and find that the error follows $\epsilon_{RV} \sim N(0.135, 0.15)$ km/s, for $\epsilon > 0$. We only want to sample positive error because the core is truncated at 0. We first calculate the probability that a value generate from $N(0.135, 0.15)$ is positive using the normal CDF: $p = 1 - F(0)$. Using this value we generate a random value $u \sim Uniform(0, 1 - F(0))$. Finally we calculate the inverse CDF of u to calculate the error, $\epsilon = F^{-1}(u)$.

We combine the error and radial velocities to calculate $\Delta RV_{\max}$:

$$\Delta RV_{\max} = X_{bin}\left[\max(v_r) - \min(v_r)\right] + \epsilon. \tag{7}$$

4.2 Model 2

Model 2 builds off of Model 1 by changing the primary mass assumption. In Model 1 we assume that the mass is $m = 1M_{\odot}$. In Model 2 we sample from the empirical mass data from the APOKASC catalog (Pinsonneault et al., 2014). We perform kernel density estimation to obtain the estimated distribution, $\hat{f}(x)$, of stellar masses. We determine the bandwidth using the heuristic of Scott (2008):

$$h = n^{-\frac{1}{4+d}} \tag{8}$$

where n is the number of data points and d is the dimensionality of the data. The mass data has dimensionality $d = 1$. We then sample the primary masses of stellar systems $m_i \sim \hat{f}_m(x)$.

4.3 Model 3

In Model 3 we constrain the value of the minimum orbital semimajor axis, a_{min}. In Models 1 and 2 we assume that $a_{min} = 3r_{\odot}$. We again perform kernel density estimation on the empirical data of $\log_{10} g$ from the APOGEE catalog to obtain the estimated distribution, $\hat{f}_{\log_{10}(g)}(x)$ (see Figure 2). We sample from this estimated distribution: $\log_{10}(g)_i \sim \hat{f}_{\log_{10}(g)}(x)$.

To calculate the minimum orbital axis a_{min} we must first obtain the radii of

the primary stars, r. We make use of Newton's Theory of Gravitation:

$$g = \frac{GM}{r^2}. \tag{9}$$

Here M is the stellar mass, r is the stellar radius, g is the force of gravity on the stellar surface and G is Newton's gravitational constant. Using the generated masses and $\log_{10}(g)$ we solve for the radius of the primary star as

$$r_i = \sqrt{\frac{Gm_i}{g_i}}. \tag{10}$$

We now must calculate the radius of the Roche lobe around the primary star, r_L. We make use of Eggleton's approximation (Eggleton, 1983):

$$\frac{r_{L,\,i}}{a_i} = \frac{0.49q_i^{-2/3}}{0.6q_i^{-2/3} + \log_{10}\left(1 + q_i^{-1/3}\right)}. \tag{11}$$

It must hold that the radius of the primary star is less than the radius of the Roche lobe. Otherwise Roche lobe overflow will occur. Therefore

$$r < r_L \;\Rightarrow\; a > \frac{ra}{r_L} \;\Rightarrow\; a_{min} = \frac{ra}{r_L}.$$

4.4 Model 4

In previous models, we assume circular orbits where the eccentricity is $e = 1$. In Model 4 we change this assumption and sample the eccentricity from an empirical distribution. We use the empirical data from Figure 14 in Raghavan et al. (2010) to derive the conditional distribution of eccentricity given the period. For each system in our model we sample from this conditional distribution.

To include the eccentricity in the radial velocity calculations we must first approximate the eccentric anomaly, $E(t)$. The mean anomaly of a system with period p at time t is:

$$(t) = kt = \frac{2\pi t}{p} = E(t) - e\sin(E(t)). \tag{12}$$

We use Newton's method to approximate the value of $E(t)$ for each time lag t. We then calculate the true anomaly $\nu(t)$:

$$\nu(t) = 2\texttt{arctan2}\left(\sqrt{1+e}\sin\frac{E(t)}{2}, \sqrt{1-e}\cos\frac{E(t)}{2}\right), \tag{13}$$

where `atan2` is the arctan function. Then we calculate the radial velocities for

each system:

$$v_r(t) = \frac{ka}{\sqrt{1-e^2}} \sin(i)[\sin(\omega + \nu(t)) + \sin(\omega)]. \tag{14}$$

We then calculate the ΔRV_{max} as in equation (7).

5 ABC Analyses

5.1 Prior Distribution

When determining the prior distribution to use, we first consider the known information on the parameter of interest. The stellar multiplicity fraction is a proportion and thus $f_m \in [0, 1]$. We choose to use a standard uniform distribution for our prior distribution.

$$\theta_{prior} \sim Uniform(0, 1) \tag{15}$$

5.2 Distance Function

To implement the KL divergence in ABC, we must make an empirical estimate of the KL divergence of the true distributions. For computational efficiency we use a histogram estimator to estimate the empirical distributions of the observed data $\mathcal{D}$ and a simulated data set $\mathcal{D}_s$. The empirical density function is defined as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{1}_{(x_0, x_0+h]}(x_i), \tag{16}$$

where n is the number of observations in each data set, h is the histogram bin width and the function $\mathbb{1}_{(x_0,x_0+h]}(x_i)$ is an indicator function which is 1 when $x_0 \leq x_i \leq x_0 + h$ and 0 otherwise. We interpret the empirical density function $\hat{f}_n(x)$ as the proportion of observations within the bin that contains the point x. To determine the optimal bin width h we use the heuristic of Freedman & Diaconis (1981):

$$h^* = 2\frac{IQR(\mathcal{D})}{n^{1/3}}, \tag{17}$$

where IQR($\mathcal{D}$) is the interquartile range of the data set $\mathcal{D}$. We further scale the bin width by a factor of 0.5 to capture the variation in the long tail. Using the scaled heuristic we obtain a bin width of $h \approx 6.5$ which corresponds to approximately 1000 bins in total over the observed data.

Our estimator for the KL divergence is then

$$\rho = \hat{D}(\mathcal{D}||\mathcal{D}_s) = \sum_{i=1}^{m} \hat{p}(x_i) \log\left(\frac{\hat{p}(x_i)}{\hat{q}(x_i)}\right), \tag{18}$$

where m is the total number of bins given as $m = \max\{\mathcal{D}, \mathcal{D}_s\}/h$, x_i is a point within bin i, and $\hat{p}$ and $\hat{q}$ are the estimated empirical density functions of $\mathcal{D}$ and $\mathcal{D}_s$ respectively.

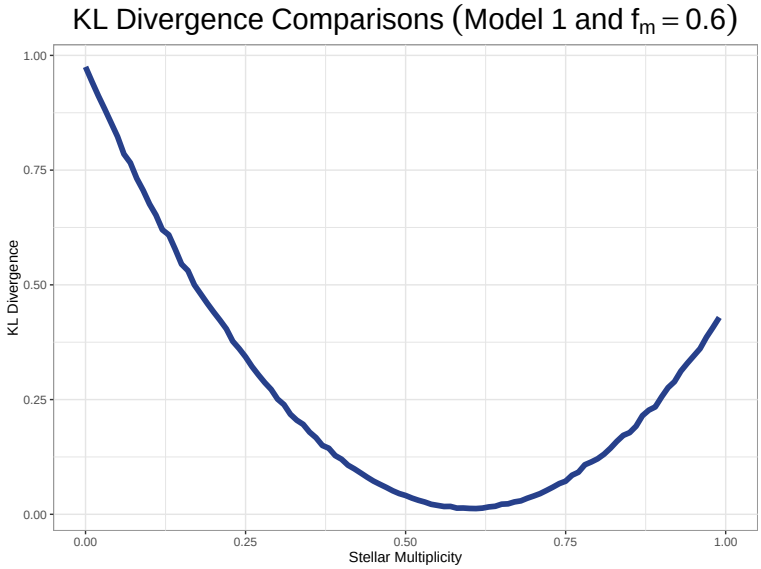

Figure 5: KL divergence measurements for simulated data of Model 1 relative to a simulated data set of $f_m = 0.6$.

Figure 5 shows the KL divergences of simulated data using Model 1. We use a simulated catalog built with $f_m = 0.6$ as our baseline and use our KL divergence metric to compare simulated data of different stellar multiplicities. We see that the KL divergence at 0.6 is near zero, which is ideal for a distance metric. This result suggests that the forward model produces "similar" catalogs for the same value of stellar multiplicity. We also see the KL divergence increases as the stellar multiplicity diverges from 0.6.

5.3 ABC Specifications

We perform four analyses within this project. Table 3 summarizes the model modules used in each. Our first analysis utilizes the simplest model: we assume a constant mass $m = 1M_\odot$, and a minimum orbital axis length $a = 3r_\odot$. In our second model we implement mass sampling. In the third analysis we sample $\log_{10} g$ and use it to constrain the range of the orbital axis lengths. In Analysis 4 we sample eccentricities.

Our implementation of ABC through **`cosmoABC`** requires user-defined input values. These input values are displayed in Table 4. The reader should refer to Algorithm 2.2 to understand how the sampling method applies these values. We choose $N = 100$ as the number of particles in the posterior distribution because it provides sufficient information to define the posterior. This number is also small enough such that we can perform an analysis in a computationally

Operation	Model 1	Model 2	Model 3	Model 4
Determine Binary Stars	✓	✓	✓	✓
Generate Mass Ratio	✓	✓	✓	✓
Generate Inclination	✓	✓	✓	✓
Generate Argument of Pericenter	✓	✓	✓	✓
Generate Primary Mass	✗	✓	✓	✓
Generate $\log g$	✗	✗	✓	✓
Calculate Orbital Axis Minimum	✗	✗	✓	✓
Generate Orbital Axis	✓	✓	✓	✓
Generate Eccentricity	✗	✗	✗	✓
Calculate Period	✓	✓	✓	✓
Calculate Radial Velocities	✓	✓	✓	✓
Generate Error	✓	✓	✓	✓
Calculate ΔRV_{max}	✓	✓	✓	✓

Table 3: Modules used for each Model.

Parameter	Value	Description
M	200	Number of particles in first particle system
N	100	Number of partcles in particle systems
q_ϵ	0.25	Quantile for distance threshold
Δ	0.05	Convergence criteria

Table 4: Values of parameters used in implementation of ABC for all analyses. See Algorithm 2.2 for the specific implementation of **`cosmoABC`**.

efficient manner; for instance, Analysis 3 took about 200 CPU-hours. For the initial particle system, we double that number. The distance threshold at each iteration t is the smallest quartile of the distances in iteration $t-1$. The analysis will end if for any particle in particle system t, $N/K < \Delta$, where K is the number of values sampled before the algorithm accepts the particle. This means that the algorithm converges when it takes at least $N/\Delta = 2000$ draws of the parameter to find an acceptable parameter.

6 Results and Discussion

Figure 6 shows simulated distributions of ΔRV_{max} for a fixed value of stellar multiplicity for each model. We see that these distributions exhibit a main core with a long tail, similar to the empirical data in Figure 2. Analysis 1 produces the smoothest curve because the model has fewer free parameters relative to the other analyses. In Analysis 2, we sample from the empirical mass distribution. This sampling introduced more noise to the model as seen in the curve. However, the distribution tends to follow the distribution produced in Model 1.

In Analysis 3, we tighten the range of values for the orbital axis length. This change in assumption impacts the distribution. The ΔRV_{max} values tend to be

less in model 3 compared to Models 1 and 2.

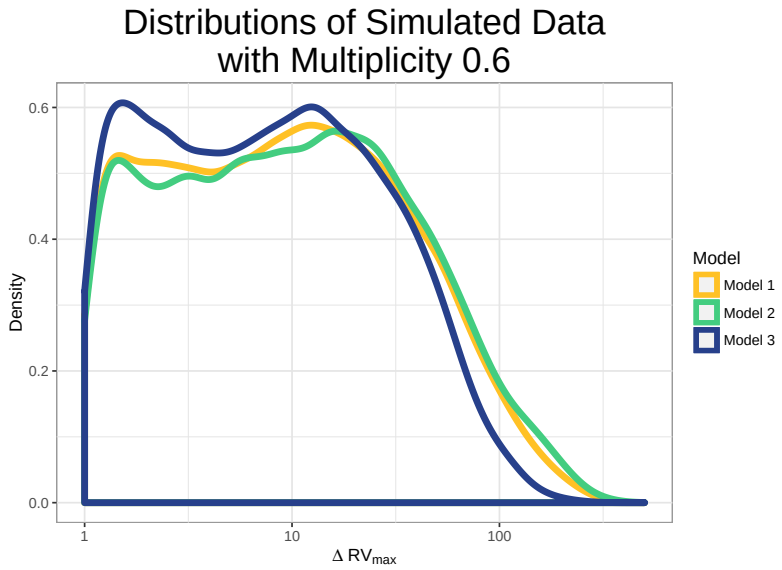

Figure 6: Distributions of data generated from the three forward models for a fixed stellar multiplicity $f_m = 0.6$.

Figure 7 shows the posterior distributions of Analyses 1, 2, and 3. Table 6 presents the posterior mean, standard deviation, median and 95% credible intervals. We see that all analyses provide similar results. All three analyses derive posterior distributions with a centered around 0.55. Also Analysis 2 displays the lowest standard deviation and Analysis 3 has the highest standard deviation. It makes sense that Analysis 3 has the highest standard deviation because Model 3 has the most free parameters of the first three models. The standard deviation of Analysis 2 is puzzling because Model 2 features an additional free parameter compared to Model 1. We would intuitively expect that Analysis 1 would have the smallest posterior standard deviation.

Analysis	Mean	Standard Deviation	Median	CI Low	CI High
1	0.552	0.017	0.548	0.520	0.585
2	0.549	0.010	0.546	0.529	0.569
3	0.555	0.026	0.548	0.504	0.607

Table 5: Summary statistics for all posterior distributions.

The observational studies reviewed in Duchêne & Kraus (2013) show that the stellar multiplicity for solar type stars is $f_m = 0.50 \pm 0.04$. We found that the stellar multiplicity with a 95% credible interval is $f_m = 0.55 \pm 0.051$, for analysis 3. Both ranges overlap suggesting that our results provide validity for the observational studies. The discrepancy between the centers could be due to selection bias in the observational studies. Also our analysis uses masses within

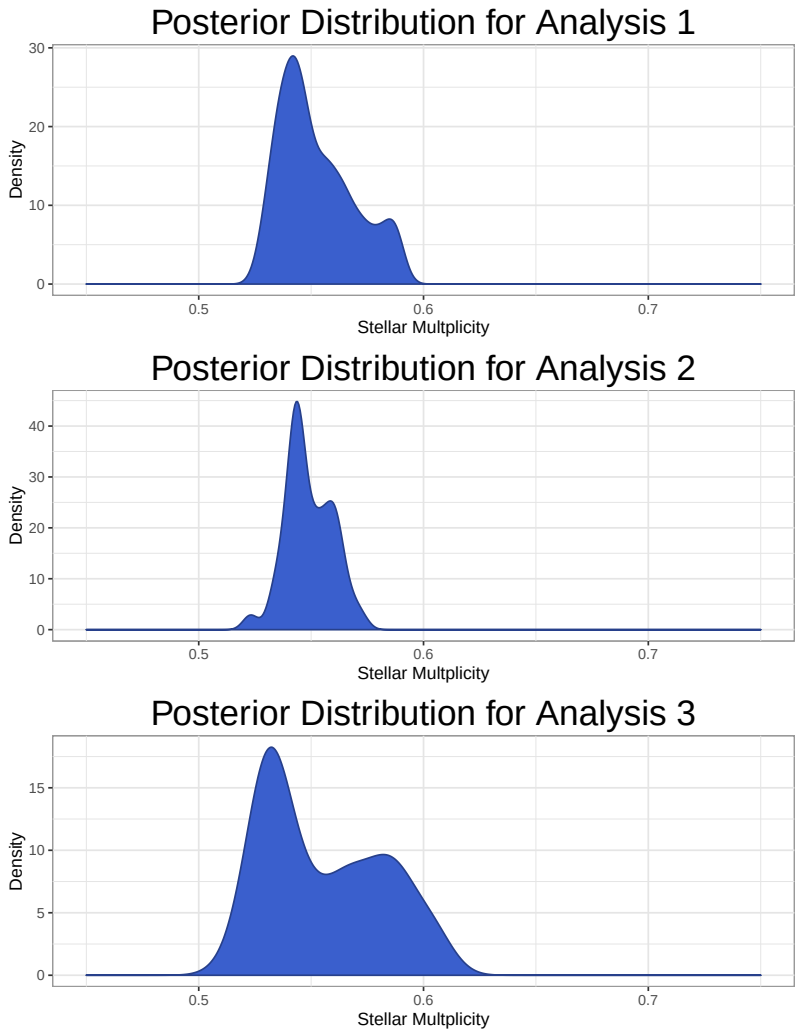

Figure 7: Posterior distributions for Analyses 1–3, from top to bottom respectively.

the range of 0.80 to 3.25 $M_{\odot}$, where Duchêne & Kraus (2013) reference results for stars with masses between 1.5 and 5 $M_{\odot}$.

We also performed an ABC analysis for Model 4. Figure 8 shows the derived posterior distribution.

We see that our analysis derived a bi-modal posterior distribution. This result is sufficiently different from the previous results so as to call into question the construction of simulating algorithm. We must perform further investigation to perform an analysis with the eccentricity.

7 Future Work

In this project we develop a framework to constrain stellar multiplicity for solar-type stars. In future analyses we can expand on this framework. As more ob-

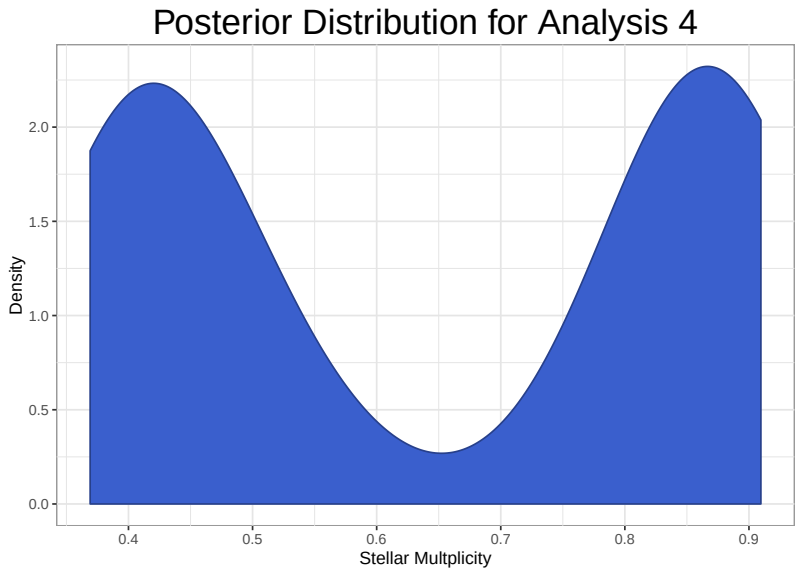

Figure 8: Posterior distribution derived in Analysis 4.

served data becomes available we can modify our forward model to simulate systems different from those containing solar-type stars. This will improve our understanding of stellar multiplicity as a function of mass.

We can also continue to improve Model 4. In Analysis 4 our results failed to provide insight on the value of stellar multiplicity. We believe this result could be due to a modeling error. In further investigations we hope to fix this error and improve the results of Analysis 4.

To further constrain the results obtained within this project we can investigate improvements to the ABC process. Analysis 3 with 100 particles and a sample of about 100,000 stellar systems requires about 200 CPU-hours. With more computing resources we can expand the size of the particle system, and tighten our acceptance and convergence criteria. These changes may improve the constraints we can place on stellar multiplicity.

We can also use different parameters for comparison. We used $\boldsymbol{\Delta RV_{max}}$ as our only parameter for comparison between simulated and observed data. We can utilize different parameters provided in the APOGEE data set such as measures of metallicity and eccentricity over inclination, e/i. We can also use an approximation of a sufficient statistic to compare the multivariate distribution of multiple parameters against the observed data.

Finally, we can perform joint analyses on a vector of parameters such as mass and stellar multiplicity. With this joint distribution we will then be able to derive the conditional distribution of stellar multiplicity as a function of mass.

8 Conclusion

We developed a framework to constrain the stellar multiplicity fraction with Approximate Bayesian Computation. Our analysis utilized the Apache Point Observatory Galactic Evolution Experiment catalog, the largest sample of stellar objects with resolution allowing efficient detection of stellar multiplicity. We built a forward model to simulate the data-generating process of stellar systems. Using this forward model we implemented ABC to constrain the value of stellar multiplicity for solar-type stars. From our third analysis we determined that the stellar multiplicity for solar-type stars is $f_m = 0.555 \pm 0.051$ with 95% credible intervals.

References

Akeret, J., Refregier, A., Amara, A., Seehars, S., & Hasner, C. 2015, J. Cosmol. Astropart. Phys., 2015, 043

Cameron, E., & Pettitt, A. N. 2012, Monthly Notices of the Royal Astronomical Society, 425, 44

Csilléry, K., Blum, M. G., Gaggiotti, O. E., & François, O. 2010, Trends in Ecology & Evolution, 25, 410

Duchêne, G., & Kraus, A. 2013, Annual Review of Astronomy and Astrophysics, 51, 269

Eggleton, P. P. 1983, ApJ, 268, 368

Freedman, D., & Diaconis, P. 1981, Probability Theory and Related Fields, 57, 453

Ishida, E., Vitenti, S., Penna-Lima, M., et al. 2015, Astronomy and Computing, 13, 1

Majewski, S. R., Schiavon, R. P., Frinchaboy, P. M., et al. 2015, ArXiv e-prints, arXiv:1509.05420

Mathieu, R. D. 1994, Annual Review of Astronomy and Astrophysics, 32, 465

Paxton, B., Marchant, P., Schwab, J., et al. 2015, ApJS, 220, 15

Perryman, M. 2011, Exoplanet Handbook (Cambridge University Press)

Pinsonneault, M. H., Elsworth, Y., Epstein, C., et al. 2014, ApJS, 215, 19

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., & Feldman, M. W. 1999, 16, 1791

Raghavan, D., McAlister, H. A., Henry, T. J., et al. 2010, ApJS, 190, 1

Robin, A. C., Reylé, C., Fliri, J., et al. 2014, Astronomy & Astrophysics, 569, A13

Schafer, C. M., & Freeman, P. E. 2012, in Lecture Notes in Statistics (Springer Science + Business Media), 3–19

Schreiber, M. R., & Gaensicke, B. T. 2003, Astronomy and Astrophysics, 406, 305

Scott, D. W. 2008, Multivariate Density Estimation (John Wiley & Sons, Inc.)

Shu, F. H., Adams, F. C., & Lizano, S. 1987, Annual Review of Astronomy and Astrophysics, 25, 23

Tohline, J. E. 2002, Annual Review of Astronomy and Astrophysics, 40, 349

Turner, B. M., & Zandt, T. V. 2012, Journal of Mathematical Psychology, 56, 69

Weyant, A., Schafer, C., & Wood-Vasey, W. M. 2013, ApJ, 764, 116