On the Relation Between Intrinsic and RV-Traced Exoplanet Observations

Cail Daley

[1]Department of Astronomy and Van Vleck Observatory, Wesleyan University, Middletown, CT, 06459; cdaley@wesleyan.edu

ABSTRACT

The true exoplanet distribution across a given parameter (planet mass, semi-major axis, etc.) is not, of course, perfectly traced by the distribution derived from radial velocity observations. This is due to many factors and biases, in particular the bias towards high-mass planets close to their host star. Assuming we understand these biases well, it should be possible to find the true exoplanet distribution that corresponds to the distribution traced by RV techniques. While such a goal falls outsides the scope of this project, I would like to take inspiration from this concept and explore the relations between intrinsic distributions and those traced by RV techniques. This would involve creating synthetic radial velocity curves for hundreds/thousands of systems, varying the number of planets, planet masses and separations, etc. Then for each system, I will use some combination of nested sampling and MCMC methods to fit the data, allowing the distribution derived from RV data to be compared to the 'true' data. I would like to emphasize that the goal of this project does not take the form of an attempt to directly learn about the intrinsic exoplanet distribution in our universe, but rather to learn about the relation between RV-observed and intrinsic distributions regardless of the form the distribution takes.

Subject headings: methods: statistical, numerical — techniques: radial velocities — planets and satellites: detection, fundamental parameters

1. Introduction

The primary focus of exoplanetary science during its two decades of existence has been the identification and characterization of individual exoplanets. This task is absolutely

crucial to the field, as it generates a statistically significant sample of the exoplanet distribution that can in turn be used to constrain theories of planet formation, migration and so on. NASA's exoplanet archive lists 3550 confirmed exoplanets; a large portion of these have been detected only in the past five years. This sample has been built up with a wide variety of observational techniques, each sensitive to specific regions of the parameter space across which exoplanets are distributed. Some methods, like radial velocity or transit observations, are biased towards planets orbiting very close to their host star; others, like direct imaging, preferentially detect planets at large separations.

The presence of such biases invites the conclusion that the exoplanet distribution implied by the sample of observed exoplanets is not the same as the true or intrinsic distribution; the intrinsic distribution has been filtered through the biases specific to the various detection methods to produce the observed distribution. Indeed, we argue that for a given detection technique, we may think of the instrumental, observational, and statistical biases associated with that technique contributing to a mapping. This mapping transforms the intrinsic frequency distribution to an observed frequency distribution. It is dependent on a wide range of parameters (e.g. the planet eccentricity or the rotation rate of the host star) and may not have a functional form. Nevertheless, if the mapping for a given detection method can described (equivalent to the selection effects of the technique being well characterized), then it may be possible to recover information about the true exoplanet frequency distribution that is hidden in our observed distribution.

In the following pages, we apply this concept to the radial velocity detection technique. This involves: i) creating a synthetic 'intrinsic' exoplanet distribution, ii) simulating a twenty-year radial velocity survey of the distribution, iii) fitting the observed lightcurves with Bayesian statistical methods, and iv) compiling the fit information to create an observed distribution. We emphasize that this work falls somewhere between a proof-

of-concept and a reproduction of the biases inherent in the analytic equations discussed in Section 3 that govern radial velocity detectability. Because the biases that affect the intrinsic-to-observed mapping are myriad, it is not within the scope of this project to characterize them all accurately. Instead we hope to take a (synthetic) data-based approach to illustrating the effect of these observations. This leads us to state one of the guiding principles of this project: *eliminate or mitigate the effect of all biases other than those we are interested in characterizing.* This principle will be used throughout this work to justify various assumptions.

In Section 2 the distribution is described, and simplifying assumptions are justified. Section 3 discusses the survey, and Section 4 details the fitting process. In Section 5 we present the resulting observed exoplanet distribution, and compare it to the intrinsic distribution. In Section 6 we summarize our results, relate our work to the current state of exoplanetology, and propose future work.

2.   The Distribution

The distribution is comprised of planetary systems; each system is defined by an inclination and a stellar mass. Each system is identical, with a inclination of 90° and a stellar mass of $1\,\mathrm{M}_\odot$. Of course, fixing the inclination at any value is not realistic, and we hope to correct this in the future. Despite this we note that inclination is a purely observational parameter, unrelated to intrinsic exoplanet properties such as mass or semi-major axis, and so will not introduce systematic biases into the synthetic distribution. Fixing the stellar mass has two justifications. Stellar mass is correlated with a host of other properties (for example, luminosity or spectral line contrast) that bias the real-world observed exoplanet distribution in a variety of ways. As we are not able to account for these effects, fixing the stellar mass is the best way to prevent unknown biases into our sample.

Second, the majority of exoplanet systems detected by radial velocity lie around solar-mass stars (as seen in Figure 1), so fixing the stellar mass at $1\,\mathrm{M_\odot}$ allows better comparison between synthetic and real-world distributions.

Each system is assigned exactly one planet. Each planet is defined by a mass m, a semi-major axis a, an eccentricity e, an argument of periastron ω, and time since periastron t_0. Eccentricity is set to zero for all planets; as ω is meaningless for $e = 0$ it too is set to zero. The decision to fix eccentricity at zero was primarily motivated by difficulties in the fitting process (an expert in the field once once referred to eccentricity as a 'bugaboo'). Exoplanet mass ranges between $0.01\,\mathrm{M_\oplus}$ and $13\,\mathrm{M_{Jup}}$, the brown dwarf mass limit. Semi-major axis ranges between $0.1\,\mathrm{au}$, below this fitting gets very inconsistent, and $10\,\mathrm{au}$. This upper limit corresponds to a period of $\sim 30\,\mathrm{years}$, or 1.5 times the length of the survey discussed in the following section. As such, planets beyond $10\,\mathrm{au}$ will have limited phase coverage, significantly limiting the confidence of detections. Finally, t_0 is assigned a random number between 0 and the planet's period.[2] Planets are spaced logarithmically (in the final version..) between the boundaries given above, forming a two-dimensional logarithmic grid in $m - a$ parameter space (Figure 2). This allows easier comparison of the observed and intrinsic exoplanet distributions, as seen in Section 5.

3. Observations

Each exoplanet system is observed fifty times across the twenty year survey; the systemic radial velocity associated with a given date is calculated as follows. First, the mean

---

[2]I actually made a mistake in the simulations presented in this draft—all the planets have $t_0 = 0$. It doesn't really matter anyways, since there's only one planet in each system.
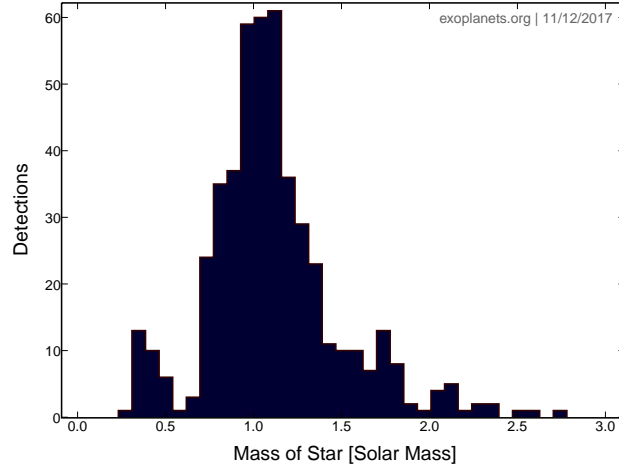
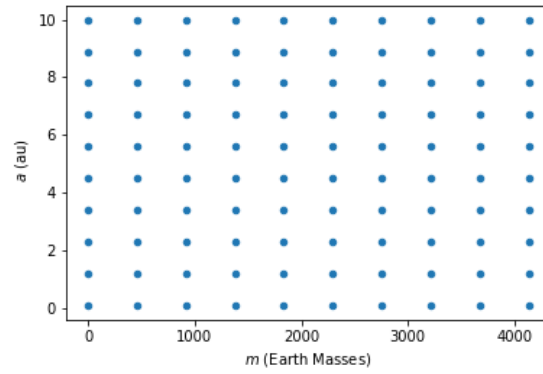Fig. 1.—: Number of exoplanet system detections as a function of stellar mass.

Fig. 2.—: Intrinsic Exoplanet Distribution

anomaly M is calculated from the planet's period P and time since periastron passage t_0.

$$M = 2\pi \frac{(t - t_0)}{P} \tag{1}$$

The eccentric anomaly E can be determined from M by solving Kepler's Equation with a numerical solver:

$$M = E - e \sin E \tag{2}$$

and the true anomaly f can be determined from E as follows:

$$f = 2 \arctan \left(\sqrt{\frac{1 + e}{1 - e}} \frac{\sin E}{\cos E} \right) \tag{3}$$

Finally, the radial velocity of the star $v_{r,1}$ can be calculated now that f is known.

$$v_{r,1} = \sqrt{\frac{G}{(m_1 + m_2)a(1 - e^2)}} m_2 \sin i \tag{4}$$

$$\cdot (\cos(\omega + f) + e \cos \omega) \tag{5}$$

This process is repeated for each date in the time series, generating a synthetic light curve. The resulting resulting curves agree well with theoretical predictions; for a $e = 0$ Jupiter analogue around a solar analogue, We derive a K_1 of $28.42 \, \mathrm{m/s}$, compared to the theoretical value of $28.43 \, \mathrm{m/s}$. To make observations more realistic, Gaussian $1 \, \mathrm{m/s}$ noise is applied to produce the 'observed' radial velocity curve. We ignore all stellar noise contributions due to the difficulty in accurately characterizing them, and because of correlations between stellar noise and exoplanet characteristics (by way of spectral type, for example).

To produce self-consistent data that reasonably approximate real-world radial velocity observations, we assume that all of the synthetic observations are taken as part of the same radial velocity survey. The survey is conducted over twenty years, with twenty observations taken each night. Imposing the requirement that each system is observed
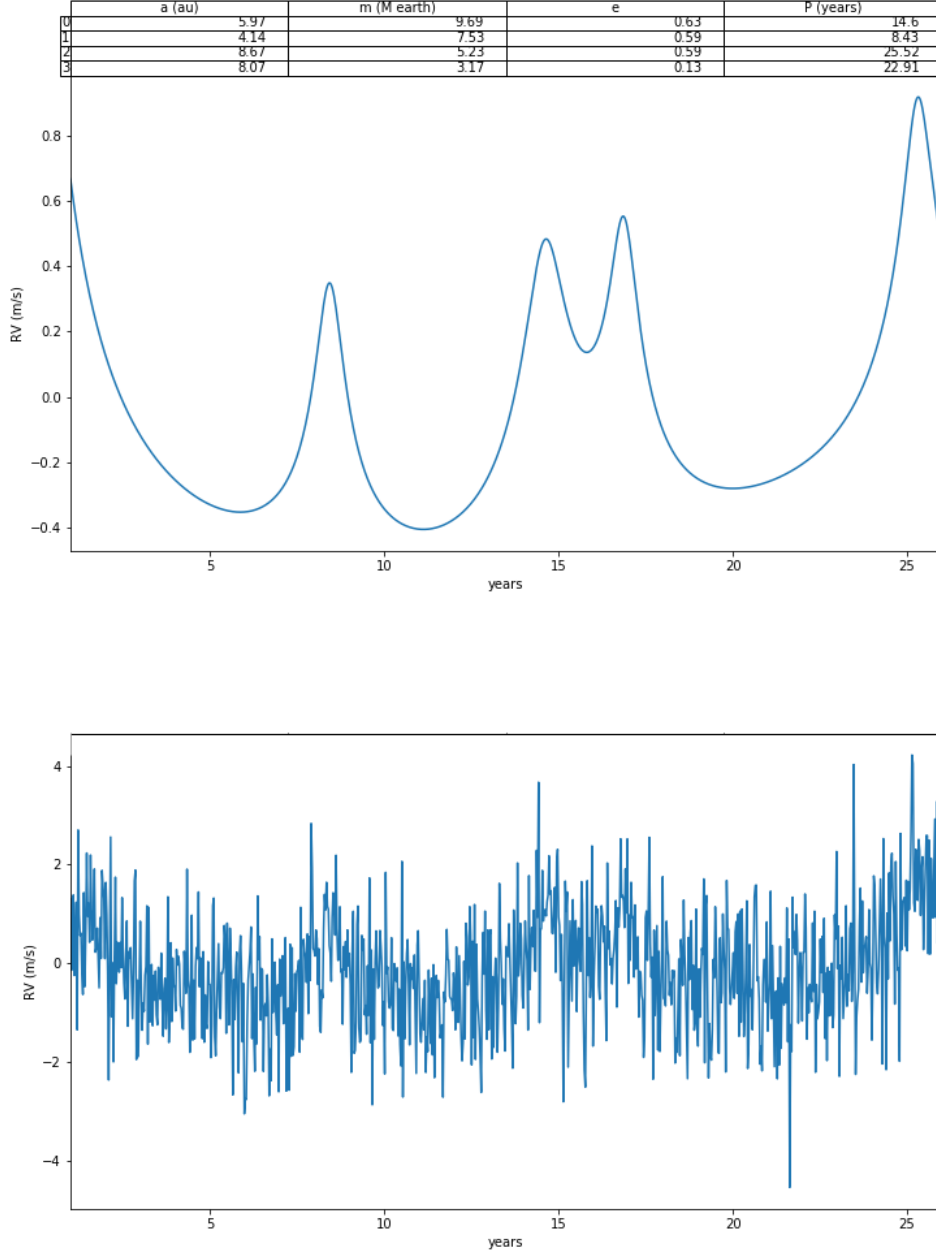
	a (au)	m (M earth)	e	P (years)
0	5.97	9.69	0.63	14.6
1	4.14	7.53	0.59	8.43
2	8.67	5.23	0.59	25.52
3	8.07	3.17	0.13	22.91

Fig. 3.—: Top: Radial velocity curve of a four-planet system around a solar analogue, with planet parameters listed at the top of the figure. The star is a solar analogue. Bottom: The same system, but with $1\,\mathrm{m/s}$ Gaussian noise added.

fifty times yields a exoplanet system size of ~ 3000 exoplanets. Observations are spaced logarithmically between 1×10^5 days and $1 \times 10^5 + (20 \times 365.25 \, \text{days})$ and then shifted back by 1×10^5 days, so that the first observation occurs at $t = 0$. Because logarithmic spacing converges to linear spacing for very large numbers, this method results in almost—but not quite—linearly spaced observations. In fact, the observations are spaced in such a way that breadth of relevant frequency space is well sampled. Figure 4 shows both the intrinsic and observed lightcurves of two planets—one short-period and one long-period —as observed by the survey. In both cases, coverage of the entirety of phase space is recovered.

4. Fitting the Data

With twenty years of observational data in hand, analysis can be performed. Bayesian statistical techniques are used to 'sample' parameter space and recover exoplanet properties. Parameter space is sampled by generating model radial velocity curves with different combinations of m, a and t_0 (e and ω are fixed to zero). The creation of these curves exactly follows the procedure described in Section 3, except that noise is not added. Model curves are then compared to the observed radial velocity curve and the log likelihood $L = -\chi^2/2$ is used to asses model quality of fit. Thus log likelihood values are assigned to different combinations of m, a and t_0, mapping out the log likelihood function throughout the three-dimensional parameter space. In an ideal situation without noise, the log likelihood function in the neighborhood of the solution takes the form of a trivariate Gaussian centered at the solution. Even with noise added, inspection reveals the log likelihood to be generally Gaussian. This can be more intuitively understood if a two-dimensional parameter space is considered; here the parameter space is analogous to a topographic landscape, with regions of higher and lower likelihoods manifesting as peaks and troughs respectively.

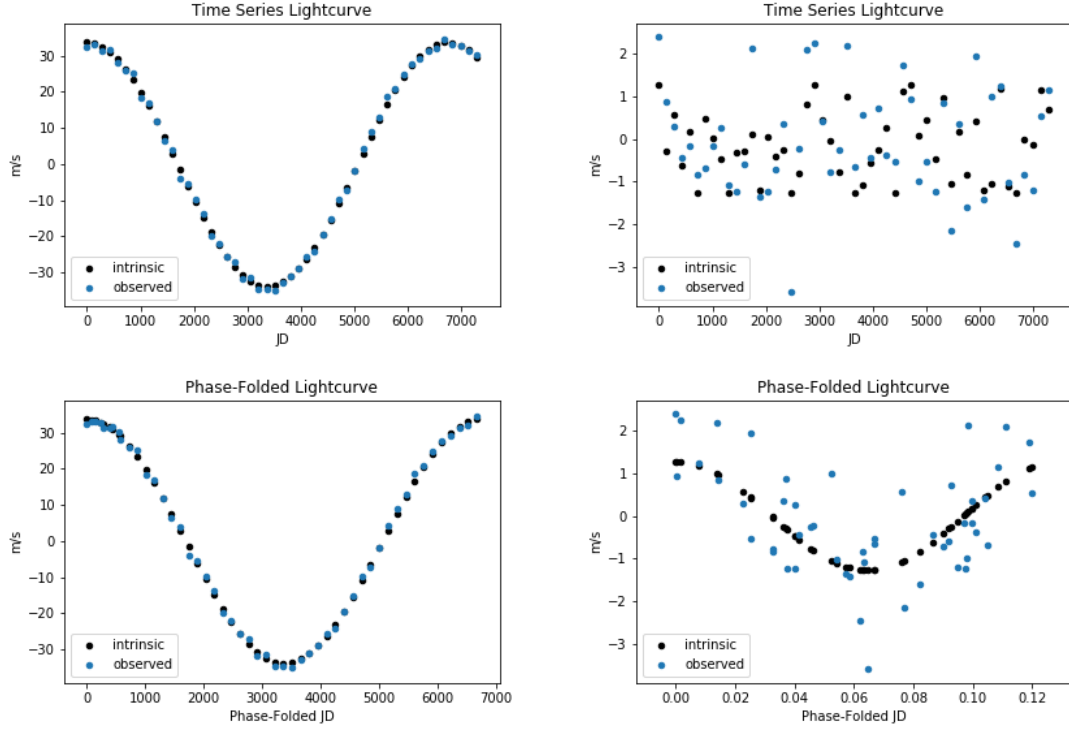The nested sampling fitting code `PyMultiNest`, a Python implementation of the

Fig. 4.—: The effect of logarithmic-almost-linear spacings between observations. On the left, the time series lightcurve (above) and phase-folded lightcurve (below) for a long-period (~ 7000 day) exoplanet. On the right, the same two plots for a short-period (~ 0.1 day) exoplanet. At both extremes, the entirety of phase space is well sampled.

Bayesian inference algorithm `MULTINEST`, is used to perform the sampling described above (Buchner et al. 2014; Feroz et al. 2009). While an in-depth discussion of the `MULTINEST` algorithm falls outside the scope of this paper, a brief description is provded it essentially calculates the model evidence Z (the denominator of Bayes' Theorem) by transforming a n-dimensional integral over all of parameter space (i.e. the prior volume) to a one-dimensional sum across all likelihood values. This accomplished by distributing a set of 'live' points across parameter space. The point with the lowest-likelihood L_0 is iteratively removed from the set of live points and replaced with a new live point with likelihood $L > L_0$. To increase efficiency, the new point is drawn from the volume enclosed by an n-dimensional ellipsoid generated from the covariance matrix of the live points that provides an estimate of the iso-likelihood contour given by $L = L_0$. Thus the algorithm climbs through iso-likelihood contours that enclose decreasing volumes of parameter space until the remaining regions of parameter space are estimated to contribute an evidence ΔZ that is less than some predefined cutoff value. In this work, the iteration stops when $\log \Delta Z < 0.5$.

As a final step, the evidence Z is calculated by computing the 'area under the curve' from the iso-likelihood contours via the trapezium rule. The primary product of the nested sampling algorithm is, as described above, the evidence. As the evidence is the average likelihood over the prior parameter space, nested sampling allows parameter spaces defined by different hyperparameters (i.e. one planet vs. two planets) to be directly compared. While different sets of model hyperparameters are not considered in this work, the current fitting structure could be expanded to include comparisons between models with different numbers of planets with relative ease. Posterior probability distributions and the best-fit model can also be determined from the set of live and discarded points produced by nested sampling; we determine the 'observed' exoplanet properties from the best-fit model parameters for each system.

Mass and semi-major axis are sampled logarithmically with uniform priors. Logarithmic sampling was chosen to better cover the four and three orders of magnitude traversed by m and a, respectively and the prior volume is defined so that model planet properties remain physically justifiable. Mass is confined to the range $\log(0.01\,\mathrm{M}_\oplus) < \log m < \log(13\,\mathrm{M}_{\mathrm{Jup}}) + 0.5$. The overall bounds for semi-major axis are $0.05\,\mathrm{au} < a < 11\,\mathrm{au}$; however setting a single prior for the semi-major axis proved difficult. Because of the periodicity implicit in a from Kepler's Third Law, for a given radial velocity curve there exists a family of 'harmonics' of the correct solution. For example, if a planet has a semi-major axis that corresponds to a period of $2\,\mathrm{yr}$, other solutions (more precisely, Gaussian local likelihood maxima) exist at separations corresponding to periods of $0.5\,\mathrm{yr}$, $1\,\mathrm{yr}$, $4\,\mathrm{yr}$, $6\,\mathrm{yr}$ and so on. This is further complicated by the fact that due to the nonlinear relationship between separation and period ($a \propto P^{2/3}$), both the spacing and standard deviation of the Gaussian solutions diminish as a_{true} decreases. From this we may conclude that fitting for a becomes progressively harder as $a_{true} \to 0$ because the likelihood spike associated with the correct solution becomes more and more localized and thus harder for the algorithm to identify.

This effect is exacerbated by the logarithmic time spacing of observations described in Section 3; as seen in Figure 4, for small a_{true} consecutive observations do not trace consecutive points in the planet's orbital phase space but rather take snapshots of the radial velocity at ostensibly random, unordered points throughout phase space. These effects combine to make the nested sampling algorithm systematically prefer higher values of a in the regime $a_{true} \lesssim 1\,\mathrm{au}$ as the harmonic solutions associated with larger a tend to be broader and thus easier to find, even if they have lower log likelihoods. As such, a decision tree was used to determine the prior in order to mitigate this effect. If $a_{true} < 2\,\mathrm{au}$, then the prior volume is defined such that $0.05\,\mathrm{au} < a < 2.5\,\mathrm{au}$. On the other hand, if $a_{true} > 2\,\mathrm{au}$ the prior is defined such that $1\,\mathrm{au} < a < 11\,\mathrm{au}$. While this certainly not realistic—exoplanetologists will rarely know the correct regime for a *a priori*—it is mostly a

time-saving convenience rather than a necessary sacrifice. As such, I offer that this decision does not significantly affect the integrity of the project.

Exoplanet detection is determined after the data has been fit. If the best-fit model peak signal-to-noise (SNR) ratio is greater than one, an exoplanet with said best-fit parameters is 'detected;' the observed distribution is then generated from all such detections. This cutoff was determined by trial and error; below this value, the best-fit model unreliably matches the data, while above the best fit model parameters are consistently in agreement with the true values. Figure 5 shows examples of fits with model peak SNR above and below the detection limit. While it may seem bizarre to use the *model* peak SNR rather than the data peak SNR, two factors motivate this decision. First, the peak SNR of the data is strongly affected by the Gaussian noise at the lower end of the detectability regime. If detection were determined directly from the data SNR, detection would be largely determined by the magnitude of largest of Gaussian noise added, introducing a random component into the logic of detection. Making use of the *model* peak SNR mitigates this random component, as the fitting does an excellent job of tracing the true, noiseless peak SNR. As can be seen in Figure 6a, the model SNR = 1 cutoff is remarkably consistent and also clearly illustrates the effect of mass and semi-major axis on the model SNR. Second, detection is determined before fitting even takes place if the data peak SNR is used. This seems contrary to the philosophy of this project, which is driven by an interest in tracing the regime of exoplanet detectability as determined (in part) by fitting techniques.
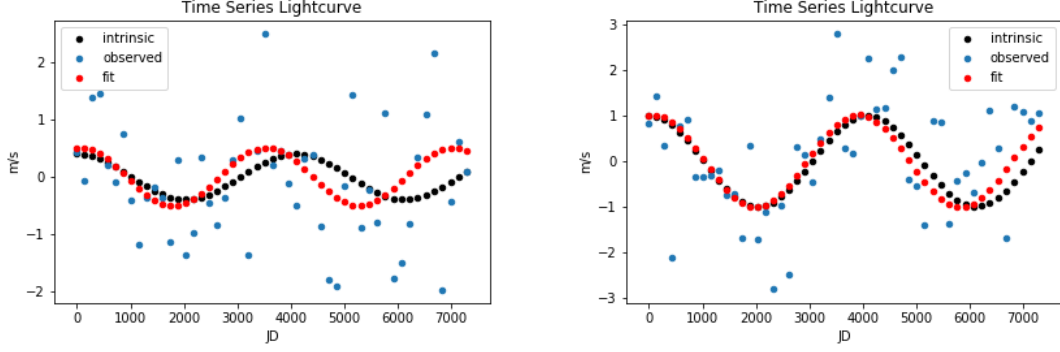
5.   Results

Figure

Fig. 5.—: Systems with model peak SNR of 0.5 (left) and 1 (right). While these examples are not necessarily representative of the entire sample, they do provide a sense of the limits of detection. The fit on the left is a non-detection, while the fit on the right lies at the lower end of the detection regime.
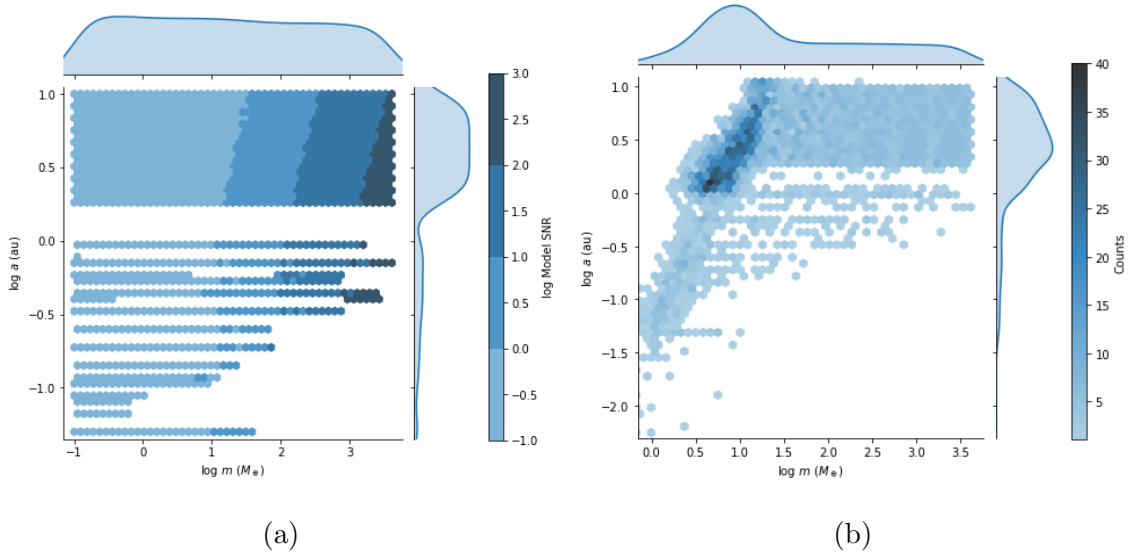

(a)   (b)

Fig. 6.—: A comparison of the intrinsic (left) and observed (right) distributions, color-coded by the best-fit model SNR. Planets with SNR < 1 are counted as nondetections.
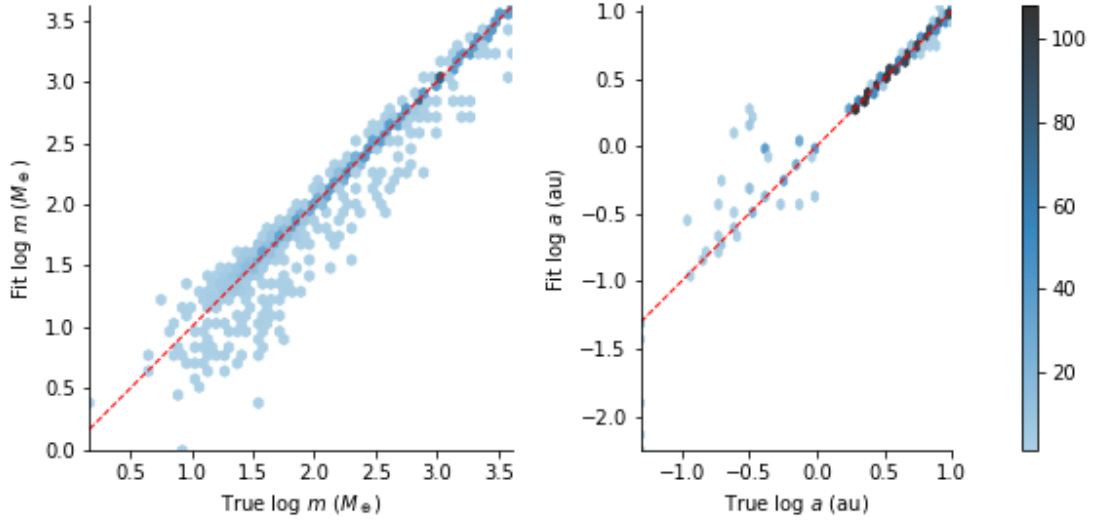
Fig. 7.—: A corner-type plot showing the correlations between the intrinsic mass distribution, the intrinsic semi-major axis distribution, the fit mass distribution, and the fit semi-major axis distribution. Again, the each planet is color-coded by the best-ft model SNR. Planets with SNR < 1 are counted as nondetections.

6. Conclusion

REFERENCES

Buchner, J., Georgakakis, A., Nandra, K., et al. 2014, A&A, 564, A125

Dumusque, X., Borsa, F., Damasso, M., et al. 2017, A&A, 598, A133

Feroz, F., Hobson, M. P., & Bridges, M. 2009, MNRAS, 398, 1601

Traub, W. A. 2016, arXiv:1605.02255

Winn, J. N., & Fabrycky, D. C. 2015, ARA&A, 53, 409