Deriving the abundance of Earth analogs in the presence of noisy, incomplete data

Joshua S. Speagle[1]

Received _____; accepted _____

[1]Harvard University Department of Astronomy, 60 Garden Street, Cambridge, MA, 02138, USA

ABSTRACT

The occurrence rates of Earth-analogs $\Gamma_\oplus$ is currently unknown. The most rigorous bounds that to date come from the *Kepler*spacecraft, which has identified hundreds of planet candidates in a variety of systems and orbits. While some of these are Earth-sized and some are on year-long orbits, no "Earth 2.0" has yet been discovered. Conducting proper population inference over this dataset to estimate $\Gamma_\oplus$, however, is hampered by noisy measurements and strong selection biases. We reproduce the work of Foreman-Mackey *et al.* (2014) to develop a general hierarchical probabilistic framework for making rigorous inference in the presence of survey (in)completeness and observational uncertainties with few underlying assumptions. We make extensive comparisons to previous work on estimating $\Gamma_\oplus$ including by Petigura *et al.* (2013b) and Dong & Zhu (2013). We also test the impact of several common approximations on synthetic data. We then re-derive Foreman-Mackey *et al.* (2014)'s estimate $\Gamma_\oplus$using a catalog of small planet candidates around G dwarf stars taken from Petigura *et al.* (2013a). We end by discussing further improvements to these techniques.

Subject headings: methods: data analysis — methods: statistical — catalogs — planetary systems — stars: statistics

1. Introduction

NASA's *Kepler*mission has ushered in a golden age for exoplanet science by enabling the discovery *thousands* of exoplanet candidates (Batalha *et al.* 2013; Burke *et al.* 2014). In particular with such large numbers of exoplanet candidates available (most of which are likely real; Morton & Johnson 2011; Fressin *et al.* 2013), we are now able to make robust conclusions about the overall population of exoplanets for the first time. In particular, many of these planets orbit "Sun-like" stars (Petigura *et al.* 2013b) in large enough numbers that it is now possible to infer changes in the population of exoplanets as a function of their physical parameters (mainly period and radius[1]). This is interesting not only from the perspective of estimating (crudely) how common our own circumstances are, but also can offer meaningful probabilistic constraints on planet formation theories.

One of the best exoplanet catalogs available in 2014 was a sample of $\sim 40,000$ Sun-like stars and an associated set of ~ 600 exoplanet candidates published by Petigura *et al.* (2013a). In addition to developing their own planet search pipeline (*TERRA* Petigura *et al.* 2013a), the authors estimated the detection efficiency of their analysis using synthetic planet injections into real *Kepler*light curves (see also Dong & Zhu 2013).

A landmark study in this area of research was (Foreman-Mackey *et al.* 2014), which proposed a flexible probabilistic model to account for both selection effects and measurement errors. Prior to this point, most studies resorted to three main approaches to trying to account for selection effects (none dealt with measurement errors):

1. making conservative cuts on the candidates and assuming that the resulting catalog

[1]Measuring masses is hard because spectroscopy is time-consuming and difficult for this particular population of exoplanets, but in the future we might have enough information to make meaningful inferences in other physical properties as well.

is complete enough to make good predictions (Catanzarite & Shao 2011; Traub 2012; Tremaine & Dong 2012),

2. assuming a particular analytic form for the detection efficiency as a function of approximate signal-to-noise (Youdin 2011; Howard *et al.* 2012; Dressing & Charbonneau 2013; Dong & Zhu 2013; Fressin *et al.* 2013; Morton & Swift 2013), and

3. determining the detection efficiency empirically by injecting synthetic signals into the raw data and testing recovery (Christiansen *et al.* 2013; Petigura *et al.* 2013a,b)

In addition, studies also utilized two different statistical analyses for turning these different approaches into quantitative occurence rate predictions. The first is what Foreman-Mackey *et al.* (2014) referred to as "inverse-detection-efficiency" (Howard *et al.* 2012; Dong & Zhu 2013; Dressing & Charbonneau 2013; Swift *et al.* 2013; Petigura *et al.* 2013b), which tries to estimate the true rate density by weighting the observed data using binned representation of the selection function. The second is parametric likelihood modeling, where the posterior distribution of observed exoplanets can be computed by comparing their likelihoods to specific parameterizations of the selection function/rate density (Tabachnik & Tremaine 2002; Youdin 2011; Dong & Zhu 2013).

The elegance of (Foreman-Mackey *et al.* 2014)'s approach was its ability to generate a data-driven model for the underlying rate density using a Gaussian process, and further incorporate measurement errors into inference for the first time. This latter portion is especially important because not only can measurement uncertainties correlate with stellar parameters, but these are often quite large compared to the relevant scales. Since the selection function can evolve strongly across an object's uncertain radius, these can have a significantly impact on any conclusions we draw from our inference.

Before we get into the details on population inference, it is important to first note what

assumptions Foreman-Mackey *et al.* (2014) *have* made when formulating their method:

1. the candidates present in the Petigura *et al.* (2013b) catalog are drawn independently from an inhomogeneous Poisson process controlled by the *censored* occurrence rate density (i.e. all planets are realized/observed randomly and independently of each other),

2. all candidates are *real exoplanets* (i.e. there are no false positives),

3. the provided uncertainties are accurate,

4. the detection efficiency of the pipeline *is exactly know* (i.e. errors on the empirical estimation are negligible), and

5. the *True*[2] occurrence rate density is *globally self-similar and smooth* (i.e. it can be modeled by a Gaussian Process with a stationary kernel).

We will return to these assumptions in Section 10.

In this paper, we define the occurence rate of Earth analogs $\Gamma_\oplus$ to be *the expected number of planets per star per natural logarithmic bin in period and radius, evaluated at the period and radius of Earth*

$$\Gamma_\oplus \quad = \quad \left.\frac{\mathrm{d}N}{\mathrm{d}\ln P\ \mathrm{d}\ln R}\right|_{R=R_\oplus,\, P=P_\oplus} \quad . \tag{1}$$

As no Earth analogs have been detected, this constraint requires an extrapolation in both period and radius.

---

[2]In this *Article*, we follow Foreman-Mackey *et al.* (2014) and use "*True*" to describe observables such as the exoplanet occurrence rate density that would be trivially measured in the limit of very high signal-to-noise data (i.e. as $N \to \infty$. By contrast, "true" describes a simulation quantity whose value we know exactly (but might not be exactly measureable).)

In Section 2, we provide a brief overview of the problem and re-derive several of the previous results in the literature. In Section 5, we outline the probabilistic model of Foreman-Mackey *et al.* (2014) and describe the computational methods they use to explore it. In Section 7, we examine how the model compares to previous methods on synthetic data. In Section 9, we use Petigura *et al.* (2013b)'s catalog of planet candidates and the associated set of detection efficiencies to infer $\Gamma_\oplus$. We conclude in Section 10.

2. Previous Work

TO BE CONTINUED...DFM's ORIGINAL PAPER CONTINUES BELOW.

3. The likelihood method

The first ingredient for any probabilistic inference is a likelihood function; a description of the probability of observing a specific dataset given a set of model parameters. In this particular project, the dataset is a catalog of exoplanet measurements and the model parameters are the values that set the shape and normalization of the occurrence rate density. Throughout this *Article*, we use the notation $\Gamma_{\boldsymbol{\theta}}(\boldsymbol{w})$ for the occurrence rate density Γ—parameterized by the parameters $\boldsymbol{\theta}$—as a function of the physical parameters $\boldsymbol{w}$ (orbital period, planetary radius, *etc.*). In this framework, the occurrence rate density can be "parametric"—for example, a power law—or a "non-parametric" function—such as a histogram where the bin heights are the parameters $\boldsymbol{\theta}$.

We'll model the catalog as a draw from the inhomogeneous Poisson process set by the *observable* rate density $\hat{\Gamma}_{\boldsymbol{\theta}}$. This leads to the previously known result (see Tabachnik &

Tremaine 2002; Youdin 2011 for some of the examples from the exoplanet literature)

$$p(\{\boldsymbol{w}_k\}\,|\,\boldsymbol{\theta}) \quad = \quad \exp\left(-\int \hat{\Gamma}_{\boldsymbol{\theta}}(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}\right) \prod_{k=1}^{K} \hat{\Gamma}_{\boldsymbol{\theta}}(\boldsymbol{w}_k) \quad . \tag{2}$$

In this equation, the integral in the normalization term is the expected number of observable exoplanets in the sample.

The main thing to note here is that $\hat{\Gamma}_{\boldsymbol{\theta}}$ is the rate density of exoplanets that you would expect to observe taking into account the geometric transit probability and any other detection efficiencies. In practice, we can model the observable rate density as

$$\hat{\Gamma}_{\boldsymbol{\theta}}(\boldsymbol{w}) \quad = \quad Q_\mathrm{c}(\boldsymbol{w})\,\Gamma_{\boldsymbol{\theta}}(\boldsymbol{w}) \tag{3}$$

where $Q_\mathrm{c}(\boldsymbol{w})$ is the detection efficiency (including transit probability) at $\boldsymbol{w}$ and $\Gamma_{\boldsymbol{\theta}}(\boldsymbol{w})$ is the object that we want to infer: the *True* occurrence rate density. We haven't yet discussed any specific functional form for $\Gamma_{\boldsymbol{\theta}}(\boldsymbol{w})$ and all of this derivation is equally applicable whether we model the rate density as, for example, a broken power law or a histogram.

The observed rate density $\hat{\Gamma}$ is a quantitative description of the rate density at which planets appear in the Petigura *et al.* (2013b) catalog; it is not a description of the *True* rate density of exoplanets. Inasmuch as the detection efficiency $Q_\mathrm{c}(\boldsymbol{w})$ is calculated correctly, the function $\Gamma_{\boldsymbol{\theta}}(\boldsymbol{w})$ will represent the *True* rate density of exoplanets, at least where there is support in the data. In practice, an estimate of the detection efficiency will not include every decision or effect in the pipeline and as this function becomes more accurate, our inferences about the *True* rate density $\Gamma_{\boldsymbol{\theta}}(\boldsymbol{w})$ will be less biased.

For the results in this *Article*, we will assume that the completeness function $Q_\mathrm{c}(\boldsymbol{w})$ is known empirically on a grid in period and radius but that is not a requirement for the validity of this method. Instead, we could use a functional form for the completeness and even infer its parameters along with the parameters of the rate density.

Finally, we model the rate density as a piecewise constant step function

$$\Gamma_{\boldsymbol{\theta}}(\boldsymbol{w}) = \begin{cases} \exp(\theta_1) & \boldsymbol{w} \in \Delta_1, \\ \exp(\theta_2) & \boldsymbol{w} \in \Delta_2, \\ \dots & \\ \exp(\theta_J) & \boldsymbol{w} \in \Delta_J, \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where the parameters θ_j are the log step heights and the bins Δ_j are fixed *a priori*. In Appendix A, we use this parameterization and derive the analytic maximum likelihood solution for the step heights. This result is similar to and just as simple as the inverse-detection-efficiency method and it is guaranteed to provide a lower variance estimate of the rate density than the standard procedure.

One major benefit of expressing the problem of occurrence rate inference probabilistically is that it can now be formally extended to include the effects of observational uncertainties.

4. A brief introduction to hierarchical inference

The general question that we are trying to answer in this *Article* is: *what constraints can we put on the occurrence rate density of exoplanets given all the light curves measured by* Kepler*?* In the case of negligible measurement uncertainties, this is equivalent to optimizing Equation (2) but when this approximation is no longer valid, we must instead compute the *marginalized likelihood*

$$p(\{\boldsymbol{x}_k\} \,|\, \boldsymbol{\theta}) = \int p(\{\boldsymbol{x}_k\} \,|\, \{\boldsymbol{w}_k\})\, p(\{\boldsymbol{w}_k\} \,|\, \boldsymbol{\theta})\, \mathrm{d}\{\boldsymbol{w}_k\} \tag{5}$$

where $\{\boldsymbol{x}_k\}$ is the set of all light curves, one light curve $\boldsymbol{x}_k$ per target k, $\boldsymbol{\theta}$ is the vector of parameters describing the population occurrence rate density $\Gamma_{\boldsymbol{\theta}}(\boldsymbol{w})$ and $\boldsymbol{w}_k$ is the vector of

physical parameters describing the planetary system (orbital periods, radius ratios, stellar radius, *etc.*) around target k. In this equation, our only assumption is that the datasets depend on the rate density of exoplanets only through the catalog $\{\boldsymbol{w}_k\}$. In our case, this assumption qualitatively means that the signals found in the light curves depend only on the actual properties of the planet and star, and not on the distributions from which they are drawn. It is worth emphasizing that—as we will discuss further below—the catalog only provides *probabilistic constraints* on $\{\boldsymbol{w}_k\}$; not perfect delta-function measurements.

In other words, we treat the catalog as being a dimensionality reduction of the raw data with all the relevant information retained. In the context of *Kepler*, the catalog reduces the set of downloaded time series (approximately 70,000 data points for the typical *Kepler* target) to probabilistic constraints on a handful of physical parameters—$\boldsymbol{w}$ from above—like the orbital period and planetary radius. If we take this set of parameters $\{\boldsymbol{w}_k\}$ as *sufficient statistics* of the data then we can, in theory, compute Equation (5)—up to an unimportant constant—without ever looking at the raw data again! This is important because the high-dimensional integral in Equation (5) won't generally have an analytic solution and each evaluation of the per-object likelihood $p(\boldsymbol{x}_k \,|\, \boldsymbol{w}_k)$ is expensive, making numerical methods intractable.

Instead, we will reuse the hard work that went into building the catalog. We must first notice that each entry in a catalog is a representation of the posterior probability

$$p(\boldsymbol{w}_k \,|\, \boldsymbol{x}_k,\, \boldsymbol{\alpha}) \quad = \quad \frac{p(\boldsymbol{x}_k \,|\, \boldsymbol{w}_k)\, p(\boldsymbol{w}_k \,|\, \boldsymbol{\alpha})}{p(\boldsymbol{x}_k \,|\, \boldsymbol{\alpha})} \tag{6}$$

of the parameters $\boldsymbol{w}_k$ conditioned on the observations of that object $\boldsymbol{x}_k$. The notation $\boldsymbol{\alpha}$ is a reminder that the catalog was produced under a specific choice of a—probably "uninformative"—*interim prior* $p(\boldsymbol{w}_k \,|\, \boldsymbol{\alpha})$. This prior was chosen by the author of the catalog and it is different from the likelihood $p(\boldsymbol{w}_k \,|\, \boldsymbol{\theta})$ from Equation (2).

Now, we can use these posterior measurements to simplify Equation (5) to a form

that can, in many common cases, be evaluated efficiently. To find this result, multiply the integrand in Equation (5) by

$$\frac{p(\{\boldsymbol{w}_k\}\,|\,\{\boldsymbol{x}_k\},\,\boldsymbol{\alpha})}{p(\{\boldsymbol{w}_k\}\,|\,\{\boldsymbol{x}_k\},\,\boldsymbol{\alpha})} \;=\; \prod_{k=1}^{K}\frac{p(\boldsymbol{w}_k\,|\,\boldsymbol{x}_k,\,\boldsymbol{\alpha})}{p(\boldsymbol{w}_k\,|\,\boldsymbol{x}_k,\,\boldsymbol{\alpha})} \tag{7}$$

and use Equation (6) to find

$$\frac{p(\{\boldsymbol{x}_k\}\,|\,\boldsymbol{\theta})}{p(\{\boldsymbol{x}_k\}\,|\,\boldsymbol{\alpha})} \;=\; \int \frac{p(\{\boldsymbol{w}_k\}\,|\,\boldsymbol{\theta})}{p(\{\boldsymbol{w}_k\}\,|\,\boldsymbol{\alpha})}\,p(\{\boldsymbol{w}_k\}\,|\,\{\boldsymbol{x}_k\},\,\boldsymbol{\alpha})\,\mathrm{d}\{\boldsymbol{w}_k\} \quad . \tag{8}$$

The data only enter this equation through the posterior constraints provided by the catalog $\{\boldsymbol{w}_k\}$! For our purposes, this is the *definition* of hierarchical inference.

The constraints in Equation (6) can always be—and often are—propagated as a list of N samples $\{\boldsymbol{w}_k\}^{(n)}$ from the posterior

$$\{\boldsymbol{w}_k\}^{(n)} \;\sim\; p(\{\boldsymbol{w}_k\}\,|\,\{\boldsymbol{x}_k\},\,\boldsymbol{\alpha}) \quad . \tag{9}$$

We can use these samples and the Monte Carlo integral approximation to estimate the marginalized likelihood from Equation (8)—up to an irrelevant constant—as

$$p(\{\boldsymbol{x}_k\}\,|\,\boldsymbol{\theta}) \;\approx\; \frac{Z_{\boldsymbol{\alpha}}}{N}\sum_{n=1}^{N}\frac{p(\{\boldsymbol{w}_k\}^{(n)}\,|\,\boldsymbol{\theta})}{p(\{\boldsymbol{w}_k\}^{(n)}\,|\,\boldsymbol{\alpha})} \tag{10}$$

where the constant $Z_{\boldsymbol{\alpha}} = p(\{\boldsymbol{x}_k\}\,|\,\boldsymbol{\alpha})$ is not a function of the parameters $\boldsymbol{\theta}$. This is very efficient to compute as long as an evaluation of $p(\{\boldsymbol{w}_k\}\,|\,\boldsymbol{\theta})$ is not expensive. That being said, Equation (10) could be a high variance estimator of Equation (8), depending on the number of independent samples N and the initial choice of $p(\{\boldsymbol{w}_k\}\,|\,\boldsymbol{\alpha})$. Additionally, the support of $p(\{\boldsymbol{w}_k\}\,|\,\boldsymbol{\theta})$ in $\{\boldsymbol{w}_k\}$ space is restricted to be narrower than that of $p(\{\boldsymbol{w}_k\}\,|\,\boldsymbol{\alpha})$. Besides this caveat, in the limit of infinite samples, the approximation in Equation (10) becomes exact. Equation (10) is the *importance sampling approximation* to the integral in Equation (8) where the trial density is the posterior probability for the catalog measurements.

A very simple example is the familiar procedure of making a histogram. If you model the function $p(\{\boldsymbol{w}_k\}\,|\,\boldsymbol{\theta})$ as a piecewise constant rate density—where the step heights are

the parameters—and if the uncertainties on the catalog are negligible compared to the bin widths then the maximum marginalized likelihood solution for $\boldsymbol{\theta}$ is a histogram of the catalog entries. The case of non-negligible uncertainties is described by Hogg *et al.* (2010b) using a method similar to the one discussed here.

5. Model generalities

Now, we can substitute Equation (2) into Equation (8) and apply the importance sampling approximation (Equation 10) to derive the following expression for the marginalized likelihood

$$\frac{p(\{\boldsymbol{x}_k\}\,|\,\boldsymbol{\theta})}{p(\{\boldsymbol{x}_k\}\,|\,\boldsymbol{\alpha})} \approx \exp\left(-\int \hat{\Gamma}_{\boldsymbol{\theta}}(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}\right) \prod_{k=1}^{K} \frac{1}{N_k} \sum_{n=1}^{N_k} \frac{\hat{\Gamma}_{\boldsymbol{\theta}}(\boldsymbol{w}_k^{(n)})}{p(\boldsymbol{w}_k^{(n)}\,|\,\boldsymbol{\alpha})} \tag{11}$$

where the values $\{\boldsymbol{w}_k{}^{(n)}\}$ are samples drawn from the posterior probability

$$\boldsymbol{w}_k{}^{(n)} \sim p(\boldsymbol{w}_k\,|\,\boldsymbol{x}_k,\,\boldsymbol{\alpha}) \tag{12}$$

as described in the previous section. Equation (11) is the *money equation* for our method. It lets us efficiently compute the *marginalized likelihood of the entire set of light curves for a particular occurrence rate density.*

In this equation, we're making the further assumption that the catalog treated the objects independently. This is a somewhat subtle point if we were to consider targets with more than one transiting planet—a point that we will return to below—but for the considerations of the dataset considered here, it is a justified simplification.

For the remainder of this *Article*, we model the rate density as a two-dimensional histogram with fixed logarithmic bins in period and radius. When we include observational uncertainties—using Equation (11)—the maximum likelihood result is no longer analytic.

Therefore, if we want to compute the "best-fit" rate density, we can use a standard non-linear optimization algorithm.

In the regions of parameter space that we tend to care about, the completeness is low and there are only a few observations with large uncertainties. In this case, we're especially interested in probabilistic constraints on the occurrence rate density; not just the best-fit model. To do this, we must apply a prior $p(\boldsymbol{\theta})$ on the rate density parameters and generate samples from the posterior probability

$$p(\boldsymbol{\theta} \,|\, \{\boldsymbol{x}_k\}) \quad \propto \quad p(\boldsymbol{\theta})\, p(\{\boldsymbol{x}_k\} \,|\, \boldsymbol{\theta}) \tag{13}$$

using Markov chain Monte Carlo (MCMC).

There is a lot of flexibility in the choice of functional form of $p(\boldsymbol{\theta})$. In the well-sampled parts of parameter space there are a lot of detected planets and the choice of prior makes little difference, but in the regions that we care about, the detection efficiency is low and applying a prior that captures our beliefs about the rate density is necessary. This will be especially important when we extrapolate the rate density function to the location of Earth—in Section 8—where no exoplanets have been found. Therefore, instead of using an uninformative prior, we want to use a prior that encourages the occurrence rate density to be "smooth" but it should be flexible enough to capture structure that is supported by the data. To achieve this, we model the logarithmic step heights as being drawn from a Gaussian process (Rasmussen & Williams 2006; Gibson *et al.* 2012; Ambikasaran *et al.* 2014). This model encodes our prior belief that, on the grid scale that we consider, the rate density should be smooth but it is otherwise very flexible about the form of the function.

Mathematically, the Gaussian process density is

$$\begin{aligned} p(\boldsymbol{\theta}) &= p(\boldsymbol{\theta} \,|\, \mu,\, \lambda) \\ &= \mathcal{N}\left[\boldsymbol{\theta};\, \mu\, \mathbf{1},\, \mathrm{K}(\{\Delta_j\},\, \boldsymbol{\lambda})\right] \end{aligned} \tag{14}$$

where $\mathcal{N}(\cdot;\, \mu\,\mathbf{1},\, \mathrm{K})$ is a J-dimensional Gaussian[3] with a constant mean μ and covariance matrix K that depends on the bin centers $\{\Delta_j\}$ and a set of hyperparameters $\boldsymbol{\lambda} = (\lambda_0,\, \lambda_P,\, \lambda_R)$. The covariance function that we use is an anisotropic, axis-aligned exponential-squared kernel so elements of the matrix are

$$\mathrm{K}_{ij} \;=\; \lambda_0 \,\exp\left(-\frac{1}{2}\,[\Delta_i - \Delta_j]^{\mathrm{T}}\,\Sigma^{-1}\,[\Delta_i - \Delta_j]\right) \tag{15}$$

where Σ^{-1} is the diagonal matrix

$$\Sigma^{-1} \;=\; \begin{pmatrix} 1/\lambda_P^2 & 0 \\ 0 & 1/\lambda_R^2 \end{pmatrix} \quad . \tag{16}$$

The Gaussian process model for the step heights given in Equation (14) is very flexible but the results will depend on the values of the hyperparameters μ and $\boldsymbol{\lambda}$. Therefore, instead of fixing these parameters to specific values, we add another level to our hierarchical probabilistic model and marginalize over this choice. In other words, we apply priors—uniform in the logarithm—on μ and $\boldsymbol{\lambda}$, and sample from the joint posterior

$$p(\boldsymbol{\theta},\, \mu,\, \boldsymbol{\lambda}\,|\,\{\boldsymbol{x}_k\}) \;\propto\; p(\mu,\, \boldsymbol{\lambda})\, p(\boldsymbol{\theta}\,|\,\mu,\, \boldsymbol{\lambda})\, p(\{\boldsymbol{x}_k\}\,|\,\boldsymbol{\theta}) \quad . \tag{17}$$

Strictly speaking, in this model, $p(\boldsymbol{\theta}\,|\,\mu,\, \boldsymbol{\lambda})$ can't really be called a "prior" anymore and the constraints on the step heights are no longer independent.

There is an efficient algorithm called elliptical slice sampling (ESS; Murray *et al.* 2010; Murray & Prescott Adams 2010) for sampling the step heights $\boldsymbol{\theta}$ from the density in Equation (17). In practice, for problems with this specific structure, ESS outperforms more traditional MCMC methods commonly employed in astrophysics (e.g., Foreman-Mackey *et al.* 2012). Our implementation is adapted from Jo Bovy's BSD licensed ESS code[4]. To

[3] J is the total number of bins.

[4] https://github.com/jobovy/bovy_mcmc/blob/master/bovy_mcmc/elliptical_slice.py

simultaneously marginalize over the hyperparameter choice, we use the Metropolis–Hastings update from Algorithm 1 in Murray & Prescott Adams (2010). We tune the Metropolis–Hastings proposal by hand until we get an acceptance fraction of $\sim 0.2 - 0.4$ for the hyperparameters.

For all the results below, we run a Markov chain with 10^6 steps for the heights and update the hyperparameters every 10 steps. We only keep the final 2×10^5 steps and discard the earlier samples as burn-in. By estimating the empirical integrated autocorrelation time of the chain (Goodman & Weare 2010), we find that the resulting chain has $\gtrsim 4000$ independent posterior samples. These samples provide an approximation to the marginalized probability distribution for $\boldsymbol{\theta}$.

6. Data and completeness function

Using an independent exoplanet search and characterization pipeline, Petigura *et al.* (2013b) published a catalog of 603 planet candidates orbiting stars in their "Sun-like" sample of *Kepler* targets. For each candidate, Petigura *et al.* (2013b) used Markov chain Monte Carlo to sample the posterior probability density for the radius ratio, transit duration, and impact parameter assuming uninformative uniform priors. They then incorporated the uncertainties in the stellar radius and published constraints on the physical radii of their candidates. Given this data reduction and since we don't have access to the individual posterior constraints on radius ratio and stellar radius, we can't directly compute the importance weights $p(\{\boldsymbol{w}_k\} \,|\, \boldsymbol{\alpha})$ needed for Equation (10). For the rest of this *Article*, we'll make the simplifying assumption that these weights are constant in log-period and log-radius but the results don't seem to be sensitive to this specific choice.

Petigura *et al.* (2013b) did not publish or share posterior samples of their measurements of the physical parameter (Equation 9). They did publish a list of periods, radii and radius uncertainties based on their analysis. Assuming that there is no measurement uncertainty on the period measurement and that the radius posterior is Gaussian in linear radius (with a standard deviation given by the published uncertainty), we draw 512 samples for $\boldsymbol{w}_k$ and use these as an approximation to the posterior probability function.

A huge benefit of this dataset is that Erik Petigura and collaborators published a rigorous analysis of the empirical end-to-end completeness of their transit search pipeline. Instead of choosing a functional form for the detection efficiency of the pipeline as a function of the parameters of interest, Petigura *et al.* (2013b) injected synthetic signals of known period and radius into the raw aperture photometry and determined the empirical recovery after the full analysis.

We use all the injected samples from Petigura *et al.* (2013b) to compute the mean (marginalized) detection efficiency in bins of $\ln P$ and $\ln R$. In each bin, this efficiency is simply the fraction of recovered injections. For the purposes of this *Article*, we neglect the counting uncertainties introduced by the finite number of samples used to estimate the completeness. The largest injected signal had a radius of $16\,R_\oplus$ but, because of the measurement uncertainties on the radii, we need to model the distribution at larger radii. To do this, we approximate the survey completeness for $R > 16\,R_\oplus$ as 1.

Given our domain knowledge of how detection efficiency depends on the physical parameters, the intuitive choice would be to measure the survey completeness in radius ratio or signal-to-noise instead of period and radius. It is also likely that a change of coordinates would yield a higher precision result. That being said, it is still correct to measure the completeness in period and radius, and there are a few practical reasons for our choice. The main argument is that since the radius uncertainties are dominated by uncertainties

in the stellar parameters, it is not possible to use the published catalog (Petigura *et al.* 2013b) to compute constraints on radius ratios. In the future, this problem would be solved by publishing a representation of *the full posterior density function for each object in the catalog*. In this case, the most useful data product would be *posterior samples for each target's radius ratio and stellar radius*.

The detection efficiency also depends on the geometric transit probability $R_\star/a$. Since we are modeling the distribution in the period–radius plane, we need to compute the transit probability marginalized over stellar radius and mass. This marginalized distribution scales only with the period of the orbit as $\propto P^{-2/3}$. In theory, this marginalization should be over the *True* distribution of these parameters in the selected stellar catalog but we'll approximate it by the empirical distribution; a reasonable simplification given the size of the dataset. At a period of 10 days[5], the median transit probability in the selected sample of stars is 5.061% so we model the transit probability[6] as a function of period as

$$Q_\mathrm{t}(P) \quad = \quad 0.05061 \left[\frac{P}{10\,\mathrm{days}}\right]^{-2/3} \quad . \tag{18}$$

This expression is clearly only valid for $P \gtrsim 1.4\,\mathrm{days}$ but the dataset that we are using (Petigura *et al.* 2013b) explicitly only includes periods longer than five days so this is not a problem. We're using the *median* transit probability (instead of the mean) because it is a more robust estimator in the presence of outliers but in our experiments, the results do not seem to be very sensitive to this choice.

Implicit in the expression for the transit probability in Equation (18) is the assumption

[5]This period is chosen arbitrarily because the power law only needs to be normalized at one point.

[6]We are using the letter Q to indicate probabilities since we are already using P to mean period.

that all of the planets are on circular orbits. Recently, Kipping (2014) demonstrated that when eccentric orbits are included, our given value is an underestimate by about 10%. This effect will propagate directly to our inferred rate densities. Even though the degeneracy is not exact—due to our choice of priors on the rate density parameters—it is not a bad approximation to assume that it is and scale the results down by your preferred factor. The right thing to do would be to marginalize over this effect directly during inference but that exercise is beyond the scope of the current *Article*. To complicate matters, the detection probability of a transit is also a non-trivial function of the duration. To account for this effect, so non-circular orbits should also be injected when measuring the survey completeness.

7. Validation using synthetic catalogs

In order to get a feeling for the constraints provided by our method and to explore any biases introduced by ignoring the observational uncertainties, we start by "observing" two synthetic catalogs from qualitatively different known occurrence rate density functions. For each of these simulations, we take the completeness function computed by Petigura *et al.* (2013b) as given. In general, Equation (2) can be sampled using a procedure called thinning (Lewis & Shedler 1979) but for our purposes, we'll simply consider a piecewise constant rate density evaluated on a fine grid in log-period and log-radius. For this discrete function, the generative procedure is simple;

1. loop over each grid cell i,
2. draw Poisson random integer $K_i \sim \mathrm{Poissson}(\hat{\Gamma}_i)$ with the observable rate density in the cell, and
3. distribute K_i catalog entries in the cell randomly.

We then choose fractional observational uncertainties on the radii from the Petigura *et al.* (2013b) catalog and apply them to the true catalog as Gaussian noise.

We generate synthetic catalogs from two qualitatively different rate density functions. Both distributions are generated by a separable model

$$\Gamma_{\boldsymbol{\theta}}(\ln P,\, \ln R) \quad = \quad \Gamma^{(P)}_{\boldsymbol{\theta}}(\ln P)\, \Gamma^{(R)}_{\boldsymbol{\theta}}(\ln R) \tag{19}$$

but fit using the full general model. The first catalog—*Catalog A*—is generated assuming a smooth occurrence surface where both distributions are broken power laws. The second—*Catalog B*—is designed to be exactly the distribution inferred by Petigura *et al.* (2013b) in the range that they considered and then smoothly extrapolated outside that range. The catalogs generated from these two models are shown in Figure 1 and Figure 2, respectively and the data are available online[7].

For each catalog, we directly apply both the inverse-detection-efficiency procedure as implemented by Petigura *et al.* 2013b[8] and our probabilistic method, marginalizing over the hyperparameters of the Gaussian process regularization. Figure 1 and Figure 2 show the results of this analysis in both cases. In particular, the side panels compare the marginalized occurrence rate density in period and radius to the true functions that were used to generate the catalogs. Figure 1 shows that even if the *True* rate density is a smooth function, the density inferred by the inverse-detection-efficiency method can appear to have sharp features. In this first example—where the true distribution is well described by our Gaussian process model—the probabilistic inference of the occurrence rate density is both more precise and accurate.

---

[7] `http://dx.doi.org/10.5281/zenodo.11507`

[8] Our implementation reproduces their results when applied to the published catalog.

In the second example, the true rate density includes a sharp feature chosen to reproduce the result published by Petigura *et al.* (2013b). In this case, Figure 2 shows that the probabilistic constraints on the rate density are less precise but more accurate than results using the inverse-detection-efficiency method. This effect is most apparent in the parts of parameter space where the detection efficiency is low—long period and small radius.

When applied to either simulated catalog, the inverse-detection-efficiency method gives a high-variance estimate of the true occurrence rate density. One effect of this variance is that the inferred distribution will appear to have more small-scale structure than the true underlying distribution.

8. Extrapolation to Earth

As well as inferring the occurrence distribution of exoplanets, this dataset can also be used to constrain the rate density of Earth analogs. Explicitly, we constrain the occurrence rate density of exoplanets orbiting "Sun-like" stars[9], evaluated at the location of Earth:

$$\Gamma_\oplus = \Gamma(\ln P_\oplus, \ln R_\oplus) \tag{20}$$

$$= \left.\frac{\mathrm{d}N}{\mathrm{d}\ln P\ \mathrm{d}\ln R}\right|_{R=R_\oplus,\, P=P_\oplus} . \tag{21}$$

That is, $\Gamma_\oplus$ is the rate density of exoplanets around a Sun-like star (expected number of planets per star per natural logarithm of period per natural logarithm of radius), evaluated at the period and radius of Earth.

In Equation (20), we use the symbol Γ instead of the more commonly used η since we

[9]In this *Article*, we adopt the Petigura *et al.* (2013b) sample of G-stars as our definition of "Sun-like".

define "Earth analog" in terms of measurable quantities with no mention of habitability or composition. This might seem unsatisfying but the composition of an exoplanet is notoriously difficult to measure even with large uncertainty and any definition of habitability is still extremely subjective. With this in mind, we stick to the observable definition for this *Article*.

Since no Earth analogs have been found, any constraints on this density must be extrapolated from the existing observations. This is generally done by assuming a functional form for the occurrence rate density, constraining it using the observed candidates and extrapolating. All published extrapolations are based on rigid models of the occurrence rate density (for example, a power law) fit to the catalog and evaluated at the location of Earth (Catanzarite & Shao 2011; Traub 2012). Petigura *et al.* (2013b) used their catalog of planet candidates to constrain the rate of Earth analogs in a specific period–radius bin assuming an extremely rigid model: *flat in logarithmic period*. These results are all sensitive to the choice of extrapolation function and the specific definition of "Earth analog".

We weaken the assumptions necessary for extrapolation by only assuming that the distribution is smooth using the Gaussian process regularization described in Section 5. Under this model, the occurrence rate density at periods and radii where no objects have been detected will be constrained—with large uncertainty—by the heights of nearby bins. Therefore, even though there are no candidates that qualify as Earth analogs, we simply fit our model of the occurrence rate density in a large enough region of parameter space (including Earth) and compute the posterior constraints on $\Gamma_\oplus$. This works because the Gaussian process regularization actually captures our prior beliefs about the shape of the rate density function. This model—and any other extrapolation—will, of course, break down if there is an unmeasured sharp feature in the occurrence rate density near the location of Earth but our method is the most conservative extrapolation technique

published to date.

For comparison, we also implemented and applied the extrapolation technique applied by Petigura *et al.* (2013b). Their method assumes that, for small planets ($1 \leq R/R_\oplus < 2$) on long periods ($P > 50\,\mathrm{days}$), the occurrence rate density is a flat function of logarithmic period or, equivalently, the cumulative rate is linear. Petigura *et al.* (2013b) used the candidates in their catalog to estimate the slope of the empirical cumulative period distribution and used that function to extrapolate. Instead of defining $\Gamma_\oplus$ differentially, as we did in Equation (20), Petigura *et al.* (2013b) constrained the integral of the rate density over a box in period and radius ($1 \leq R/R_\oplus < 2$ and $200 \leq P/\mathrm{day} < 400$). Since their model implicitly assumes a constant rate density across the bin, the differential rate is just their number divided by the bin volume. This rate density (rate divided by bin volume) is what is shown as a comparison to our results in the figures.

Figures 3 and 4 compare our results and the results of the Petigura *et al.* (2013b) extrapolation procedure when applied to the synthetic catalogs. Since these catalogs were simulated from a known population model, we know the true value of $\Gamma_\oplus$ and it is indicated in the figures with a vertical gray line. In both cases, our method returns a less precise but more accurate result for the rate density and the error bars given by the functional extrapolation are overly optimistic. One major effect that leads to this bias is that the period distribution is not flat. Restricting the result to only include uniform models is equivalent to applying an extremely informative prior that doesn't have enough freedom to capture the complexity of the problem. As a result, the posterior constraints on $\Gamma_\oplus$ are dominated by this prior choice and the resulting uncertainties are much smaller than they should be.

9. Results from real data

Having developed this probabilistic framework for exoplanet population inferences and demonstrating that it produces reasonable results when applied to simulated datasets, we now turn to real data. As described in Section 6, we will use the catalog of small exoplanet candidates orbiting Sun-like stars published by Petigura *et al.* (2013b). This is a great test case because those authors empirically measured the detection efficiency of their pipeline as a function of the parameters of interest.

We directly applied our method to the Petigura *et al.* (2013b) sample and generated MCMC samples from the posterior probability for the occurrence rate density step heights, marginalizing over the hyperparameters of the Gaussian process model. The resulting MCMC chain is available online[10].

Figure 5 shows posterior samples from the inferred occurrence rate density as a function of period and radius conditioned on the catalog. The marginalized distributions are qualitatively consistent with the occurrence rate density measured using the inverse-detection-efficiency method with larger uncertainties.

The period distribution integrated over various radius ranges is shown in Figure 6. In agreement with Dong & Zhu (2013), we find that the period distribution of large planets ($R > 8\,R_\oplus$) is inconsistent with the distribution of smaller planets. The rate density of large planets appears to monotonically increase as a function of log period while the distribution for small planets seems to turn over at a relatively short period (around 50 days) and decrease for longer periods.

The equivalent results for the radius distribution are shown in Figures 7 and 8. Figure 7 shows the log-radius occurrence rate density integrated over various logarithmic

[10] http://dx.doi.org/10.5281/zenodo.11507

bins in period. The distributions in each period bin are qualitatively consistent; the rate density is dominated by small planets (around two Earth radii) with potential "features" near $R \sim 3R_\oplus$ and $R \sim 10R_\oplus$. These features appear in every period bin. They were also detected—using a completely different dataset and technique—by Dong & Zhu (2013) and a similar result is visible in the occurrence rate determined by Fressin *et al.* (2013, their Figure 7) at low signal-to-noise. Figure 8 shows the same result but presented as a function of linear radius. In these coordinates, the rate density in a single bin is no longer uniform; instead, scales as inverse radius.

Our constraint on the rate density of Earth analogs (as defined in Section 8) is in tension—even though our result has large fractional uncertainty—with the result from Petigura *et al.* (2013b). This is shown in Figure 9 where we compare the marginalized posterior probability function for $\Gamma_\oplus$ to the published value and uncertainty. Quantitatively, we find that the rate density of Earth analogs is

$$\Gamma_\oplus \quad = \quad 0.019^{+0.019}_{-0.010} \text{ nat}^{-2} \tag{22}$$

where the "nat^{-2}" indicates that this quantity is a rate density, per natural logarithmic period per natural logarithmic radius. Converted to these units, Petigura *et al.* (2013b) measured $0.119^{+0.046}_{-0.035}$ nat^{-2} for the same quantity (indicated as the vertical lines in Figure 9). This rate density is *exactly* what Petigura's extrapolation model predicts but, for comparison, we can also integrate our inferred rate density over their choice of "Earth-like" bin ($200 \le P/\text{day} < 400$ and $1 \le R/R_\oplus < 2$) to find a *rate* of Earth analogs. The published rate is $0.057^{+0.022}_{-0.017}$ (Petigura *et al.* 2013b) and our posterior constraint is

$$\int_{P=200\,\text{day}}^{400\,\text{day}} \int_{R=1\,R_\oplus}^{2\,R_\oplus} \Gamma_{\boldsymbol{\theta}}(\ln P,\, \ln R)\, \mathrm{d}[\ln R]\, \mathrm{d}[\ln P] \quad = \quad 0.019^{+0.010}_{-0.008} \quad . \tag{23}$$

Although they are mainly nuisance parameters, we also obtain posterior constraints on the hyperparameters μ and $\boldsymbol{\lambda}$. In particular, the constraints on the length scales in $\ln P$

and $\ln R$ are $\lambda_P = 3.65 \pm 1.03$ and $\lambda_R = 0.65 \pm 0.12$ respectively. Both of these scales are larger than a bin in their respective dimension. For completeness we also find the following constraints on the other hyperparameters

$$\mu = 5.44 \pm 1.56 \qquad \text{and} \qquad \ln \lambda_0 = 1.68 \pm 0.72 \quad . \tag{24}$$

The MCMC chains used to compute these values is available online[11].

10. Comparison with previous work

Our inferred rate density of Earth analogs (Equation 22) is not consistent with previously published results. In particular, our result is completely inconsistent with the earlier result based on *exactly the same dataset* (Petigura *et al.* 2013b). This inconsistency is due to the different assumptions made and the detailed cause merits some investigation. The two key differences between our analysis and previous work are *(a)* the form of the extrapolation function, and *(b)* the presence of measurement uncertainties on the planet radii.

To make their estimate of $\Gamma_\oplus$, Petigura *et al.* (2013b) asserted a flat distribution in logarithmic period for small planets. Our results suggest that the data *do not support* this assumption (see Figure 6). We find that the data require a *decreasing* period distribution in the relevant range. A similar result was also found by Dong & Zhu (2013) and it is apparent in Figure 2 of Petigura *et al.* (2013b).

To test the significance of the choice of extrapolation function, we relax the assumption of a uniform period distribution and allow the distribution to be linear in the same range ($R = 1 - 2\,R_\oplus$ and $P = 50 - 400$d). Under this model, the likelihood of the catalog of

[11] http://dx.doi.org/10.5281/zenodo.11507

planets in this range can be calculated using Equation (2). We apply uniform priors in the physically allowed range of slopes and intercepts for this distribution and estimate the posterior probability for the extrapolated rate using MCMC (Foreman-Mackey *et al.* 2012). This results give a much more uncertain and substantially lower estimate for the rate of Earth analogs

$$\Gamma_\oplus \quad = \quad 0.072^{+0.088}_{-0.047} \quad . \tag{25}$$

With the large error bars, this result is consistent with both results (see Figure 10 where this value is labeled "linear extrapolation") but it does not fully account for the discrepancy.

To examine the effects of measurement uncertainties, we repeat our analysis with the error bars on the radii artificially set to zero, keeping everything else the same. This analysis (labeled "uncertainties ignored" in Figure 10) gives the result

$$\Gamma_\oplus \quad = \quad 0.040^{+0.031}_{-0.019} \quad . \tag{26}$$

This result is relatively more precise and higher than our final result and consistent with the value obtained with linear extrapolation. This confirms the hypothesis that the discrepancy between our result and the previously published values is the combined result of both of our key generalizations.

For comparison, we have also included the value of $\Gamma_\oplus$ implied by Dong & Zhu (2013, their Table 2). This result is based on a power law fit to the period distribution of small planets ($R = 1 - 2\,R_\oplus$) on long periods ($P = 10 - 250\,\mathrm{d}$) in a different catalog (Batalha *et al.* 2013) with a parametric completeness model. There are a few factors to consider when comparing to this to our analysis. Firstly, while Dong & Zhu (2013) fit a power law in log period, this is still a very restrictive model when considering this large range of periods. A broken power law might be more applicable. Furthermore, their analysis did not incorporate the effects of measurement uncertainties. Finally, unlike the Petigura *et al.*

(2013b), the Batalha *et al.* (2013) catalog used by Dong & Zhu (2013) includes multiple transiting systems. As mentioned previously, the effect of this selection is hard to determine without further investigation but it should, intuitively, cause any inference based on the Petigura *et al.* (2013b) sample to be an underestimate of the *True* rate.

11. Discussion

We have developed a hierarchical probabilistic framework for inferring the population of exoplanets based on noisy incomplete catalogs. This method incorporates systematic treatment of observational uncertainties and detection efficiency. One major benefit of this framework is that it provides the best possible probabilistic measurements of the population under the assumptions listed in Section 1 and repeated below. After demonstrating the validity of our method on two qualitatively different synthetic exoplanet catalogs, we run our inference on a published catalog of small exoplanet candidates orbiting Sun-like stars (Petigura *et al.* 2013b) to determine the occurrence rate density these planets as a function of period and radius. We extrapolate this measurement to the location of Earth and constrain the rate density of Earth analogs with large error bars. In order to perform this extrapolation, we don't assume a specific functional form for the rate density. Instead, we only assume that it is a smooth function of logarithmic period and radius.

The occurrence rate density function that we infer is qualitatively consistent with previously published results using different inference techniques (Dong & Zhu 2013; Fressin *et al.* 2013; Petigura *et al.* 2013b). In particular, we find (see Figure 7) previously recorded features in the radius distribution around $R \sim 3\,R_\oplus$ and $R \sim 10\,R_\oplus$, although not at high signal-to-noise. We find that the period distributions for planets in different radius bins are different, in qualitative agreement with previous results (Dong & Zhu 2013). Figure 6 shows that larger planets tend to be on longer periods than smaller planets.

Our extrapolation of the rate density to the location of Earth is more general and conservative than any previously published method. We find a rate density of Earth analogs that is inconsistent with the result published by Petigura *et al.* (2013b). This discrepancy can be attributed to both the rigidity of the assumptions about the period distribution and the effects of non-negligible measurement uncertainties. Our extrapolation is also less confident than previous measurements. Again, this difference is due to the fact that we allow a much more flexible extrapolation function. This is another illustration that, against the standard data analysis folklore, the correct use of flexible models is *conservative*.

In contrast to previous work, we don't define "Earth analog" in terms of habitability or composition. Instead, we advocate for a definition in terms of more directly observable quantities (in this case, period and radius). Furthermore, we define $\Gamma_\oplus$ as a rate density (per star per logarithmic period per logarithmic radius) so that its value doesn't depend on choices about the "Earth-like" bin.

In our analysis we make a few simplifying assumptions. Every assumption has an effect on the results and could be relaxed as an extension of this project. For completeness, we list and discuss the effects of our assumptions below.

- **Conditional independence** We assume that every object in the catalog is a conditionally independent draw from the observable occurrence rate density. This is a bad assumption when applying this method to a different catalog where multiple transiting systems are included. In practice, the best first step towards relaxing this assumption is probably to follow Tremaine & Dong (2012) and assume that the mutual inclination distribution is the only source of conditional dependence between planets. For this *Article*, the assumption of conditional independence is justified because the dataset explicitly includes only systems with a single transiting exoplanet.

- **False positives** In our inferences, we assume that all of the candidates in the

catalog are *True* exoplanets. The rate of false positives in the *Kepler* catalog has been shown to be low but not negligible (Morton & Johnson 2011; Fressin *et al.* 2013). Since some of the objects in the catalog are probably false positives, our inferences about the occurrence rate density are biased high but without explicitly including a model of false positives, it's hard to say in detail what effect this would have on the distributions. In an extension of this work, we could incorporate the effects of false positives by switching to a mixture model (see Hogg *et al.* 2010a, for example) where each object is modeled as a mixture of *True* exoplanet and false positive. In this mixture model, the false positives would be represented using prior distributions similar to those used by Morton (2012) or Fressin *et al.* (2013).

- **Known observational uncertainties** To apply the importance sampling approximation to the published catalog, we assume that the measurement uncertainties are known and, in this case, Gaussian. The assumption of normally distributed uncertainties could be relaxed given a sampling representation of the posterior probability function for the physical parameters (period, radius, *etc.*). There is recent evidence that the stellar radii of *Kepler* targets might, on average, be underestimated (Bastien *et al.* 2014), introducing another source of noise. It is possible to relax the noise model and include effects like this but inference would be substantially more computationally expensive.

- **Given empirical detection efficiency** Petigura *et al.* (2013b) determined the end-to-end detection efficiency of their planet detection pipeline as a function of *True* period and radius by injecting synthetic signals into real light curves and testing recovery. We used these simulations as an exact representation of the detection efficiency of the catalog but there are several missing components. The biggest effect is probably the fact that this formulation doesn't include the selection of only the

most detectable signal in each light curve. This bias will be largest in the parts of parameter space where the baseline detection efficiency is lowest: at long periods and small radius. As a result, our inferences (and the results from Petigura *et al.* 2013b) about the occurrence rate of small planets on long periods is probably *underestimated* relative to *Truth*. In detail there is another limitation due to the fact that the stellar parameters are only known noisily and the transit light curve only constrains the radius ratio. This means that the marginalized detection efficiency should be measured as a function of radius ratio and the interpretation in terms of *True* radius is only approximately correct. Given the size of the dataset and the number of injection simulations, this effect should be small.

- **Smooth rate function** Throughout our analysis, we make the prior assumption that the occurrence rate density is a smooth function of logarithmic period and radius. This model is useful because it allows us to make probabilistically justified inferences about the exoplanet population in regions of parameter space with low detection efficiency. The assumption that the rate density should be smooth is intuitive but there is no theoretical indication that it must be true at all scales. That being said, the Gaussian process regularization that we use to enforce smoothness is flexible enough to capture substantial departures from smooth if they were supported by the data.

Our assumptions are severe but we believe that this is the most conservative population inference method currently on the market.

Under the assumptions that we have made here, our inference of the occurrence rate density of exoplanets places a probabilistic constraint on the number of transiting Earth analogs in the existing *Kepler* dataset. If we adopt the definition of "Earth-like" from Petigura *et al.* (2013b, $200 \le P/\mathrm{day} < 400$ and $1 \le R/R_\oplus < 2$), and integrate the product

inferred rate density function and the geometric transit probability (Equation 18) over this bin, we find that the expected number of Earth-like exoplanets transiting the stars in the sample of Sun-like stars chosen by Petigura *et al.* (2013b) is

$$N_{\oplus,\,\mathrm{transiting}} = 10.6^{+5.9}_{-4.5} \tag{27}$$

where the uncertainties are only on the expectation value and don't include the Poisson sampling variance. This is an exciting result because it means that, if we can improve the sensitivity of exoplanet search pipelines to small planets orbiting on long periods, then we should find some Earth analogs in the existing data. Furthermore, because of the treatment of multiple transiting systems in the catalog, the *True* expected number of transiting Earth-like exoplanets orbiting Sun-like stars is almost certainly larger than the values in Equation (27)!

Some of the caveats on the results in this paper are due to assumptions made for computational simplicity but a much more robust study would be possible given a complete representation of the posterior probability function for the physical parameters in the catalog. The use of MCMC to fit models to observations is becoming standard practice in astronomy and the results in many catalogs (including Petigura *et al.* 2013b) are given as statistics computed on posterior samplings. For the sake of hierarchical inferences like the method presented here, it would be very useful if the authors of upcoming catalogs also published samples from these distributions *along with the value of their prior function evaluated at each sample*. In this spirit, we have released the results of this paper as posterior samplings[12] for the occurrence rate density function.

All of the code used in this project is available from `http://github.com/dfm/exopop` under the MIT open-source software license. This code (plus some dependencies) can be

[12] `http://dx.doi.org/10.5281/zenodo.11507`

run to re-generate all of the figures and results in this *Article*; this version of the paper was generated with git commit `d56324d` (2014-08-28).

We would like to thank Erik Petigura (Berkeley) for freely sharing his data and code. It is a pleasure to thank Ruth Angus (Oxford), Tom Barclay (NASA Ames), Jo Bovy (IAS), Eric Ford (PSU), David Kipping (CfA), Ben Montet (Caltech/Harvard), and Scott Tremaine (IAS) for helpful contributions to the ideas and code presented here. We would also like to acknowledge the anonymous referee and the Scientific Editor, Eric Feigelson, for suggestions that substantially improved the paper. This project was partially supported by the NSF (grant AST-0908357), NASA (grant NNX08AJ48G), and the Moore–Sloan Data Science Environment at NYU. This research builds on ideas generated at a three-week workshop supported by NSF Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. This research made use of the NASA *Astrophysics Data System*.

APPENDIX

A. Inverse-detection-efficiency

One huge benefit of the inverse-detection-efficiency procedure is its simplicity. Therefore, it's worth noting that there is a probabilistically justified procedure that will always provide less biased results while being only marginally more complicated.

The standard procedure involves making a weighted histogram of the catalog entries where the weight for object $\boldsymbol{w}_k$ is $1/Q_\mathrm{c}(\boldsymbol{w}_k)$. This makes intuitive sense but it does not have a clear probabilistic justification or interpretation. As we will show below, the maximum likelihood result involves weighting the points by the inverse of the *integral* of the completeness function over the bin area.

To motivate this derivation, let's start by considering the following pathological example: a single bin where the completeness sharply drops from one to zero halfway across the bin. If we observe K objects in this bin, we would have observed about $2K$ objects in a complete sample. If we apply the inverse-detection-efficiency procedure to this dataset, each sample will get unit weight because they are all found in the part of the bin where the completeness is one. Therefore, we would *underestimate* the true rate in the bin by half. It's clear in this specific case that giving the points a weight of two would give a better solution and we'll derive the general result below.

If we model the occurrence rate density as a histogram with J fixed bin volumes Δ_j (Equation 4) then Equation (2) becomes

$$\ln p(\{\boldsymbol{w}_k\} \,|\, \boldsymbol{\theta}) \quad = \quad \sum_{k=1}^{K} \sum_{j=1}^{J} \mathbf{1}[\boldsymbol{w}_k \in \Delta_j] \left[\ln Q_\mathrm{c}(\boldsymbol{w}_k) + \theta_j\right] - \sum_{j=1}^{J} \exp(\theta_j) \int_{\Delta_j} Q_\mathrm{c}(\boldsymbol{w}) \, \mathrm{d}\boldsymbol{w} \tag{A1}$$

where the indicator function $\mathbf{1}[\cdot]$ is one if $\cdot$ is true and zero otherwise. Taking the gradient of this function with respect to $\boldsymbol{\theta}$ and setting it equal to zero, we find the maximum likelihood result

$$\exp({\theta_j}^*) \quad = \quad \frac{K_j}{\int_{\Delta_j} Q_\mathrm{c}(\boldsymbol{w}) \, \mathrm{d}\boldsymbol{w}} \tag{A2}$$

where K_j is the number of objects that fall within the bin j. We estimate the uncertainty $\delta\theta_j$ on this value by examining the curvature of the log-likelihood function near the maximum and find

$$\frac{\delta \exp({\theta_j}^*)}{\exp({\theta_j}^*)} \quad = \quad \frac{1}{\sqrt{K_j}} \quad . \tag{A3}$$

In our pathological example from above, the integral of the completeness function over the bin is 1/2, giving each sample the expected weight of 2. In more realistic cases, where the completeness function varies smoothly, the inverse-detection-efficiency result will begin to agree with Equation (A2) but the severity of this bias will be very problem dependent.

Therefore, if you have a dataset with negligible observational uncertainties, we recommend that you always apply Equation (A2) instead of the standard inverse-detection-efficiency procedure. As the uncertainties become more significant, there is no longer an analytic result and the method derived in this *Article* is necessary.

REFERENCES

Adams, R. P., Murray, I., & MacKay, D. J. C. 2009, ICML, 2009, 9 (online)

Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., & O'Neil, M. 2014, arXiv:1403.6015

Bastien, F. A., Stassun, K. G., & Pepper, J. 2014, ApJ, 788, L9 (arXiv:1405.0940)

Batalha, N. M., Rowe, J. F., Bryson, S. T., *et al.* 2013, ApJS, 204, 24 (arXiv:1202.5852)

Brewer, B. J., & Stello, D. 2009, MNRAS, 395, 2226 (arXiv:0902.3907)

Burke, C. J., Bryson, S. T., Mullally, F., *et al.* 2014, ApJS, 210, 19 (arXiv:1312.5358)

Carter, J. A., & Winn, J. N. 2009, ApJ, 704, 51 (arXiv:0909.0747)

Catanzarite, J., & Shao, M. 2011, ApJ, 738, 151 (arXiv:1103.1443)

Christiansen, J. L., Clarke, B. D., Burke, C. J., *et al.* 2013, ApJS, 207, 35 (arXiv:1303.0255)

Dong, S., & Zhu, Z. 2013, ApJ, 778, 53 (arXiv:1212.4853)

Dressing, C. D., & Charbonneau, D. 2013, ApJ, 767, 95 (arXiv:1302.1647)

Fang, J., & Margot, J.-L. 2012, ApJ, 761, 92 (arXiv:1207.5250)

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306 (arXiv:1202.3665)

Foreman-Mackey, D., Hogg, D. W., & Morton, T. D. 2014, ApJ, 795, 64F (arXiv:1406.3020)

Fressin, F., Torres, G., Charbonneau, D., *et al.* 2013, ApJ, 766, 81 (arXiv:1301.0842)

Gibson, N. P., Aigrain, S., Roberts, S., *et al.* 2012, MNRAS, 419, 2683 (arXiv:1109.3251)

Goodman, J. & Weare, J., 2010, Comm. App. Math. Comp. Sci., 5, 65

Hogg, D. W., Angus, R., Barclay, T., *et al.* 2013, arXiv:1309.0653

Hogg, D. W., Bovy, J., & Lang, D. 2010a, arXiv:1008.4686

Hogg, D. W., Myers, A. D., & Bovy, J. 2010b, ApJ, 725, 2166 (arXiv:1008.4146)

Howard, A. W., Marcy, G. W., Bryson, S. T., *et al.* 2012, ApJS, 201, 15 (arXiv:1103.2541)

Kipping, D. M. 2014, arXiv:1408.1393

Lewis, P. A. W., & Shedler, G. S. 1979, Naval Research Logistics Quarterly, 26, 403

Lissauer, J. J., Ragozzine, D., Fabrycky, D. C., *et al.* 2011, ApJS, 197, 8 (arXiv:1102.0543)

Morton, T. D. 2012, ApJ, 761, 6 (arXiv:1206.1568)

Morton, T. D., & Johnson, J. A. 2011, ApJ, 738, 170 (arXiv:1101.5630)

Morton, T. D., & Swift, J. J. 2013, arXiv:1303.3013

Murray, I., Prescott Adams, R., & MacKay, D. J. C. 2010, JMLR: W&CP, 9, 541 (arXiv:1001.0175)

Murray, I., & Prescott Adams, R. 2010, Advances in Neural Information Processing Systems, 23, 1723 (arXiv:1006.0868)

Petigura, E. A., Marcy, G. W., & Howard, A. W. 2013a, ApJ, 770, 69 (arXiv:1304.0460)

Petigura, E. A., Howard, A. W., & Marcy, G. W. 2013b, Proceedings of the National Academy of Science, 110, 19273 (arXiv:1311.6806)

Rasmussen, C. E. & Williams, C. K. I. 2006 Gaussian Processes for Machine Learning, MIT Press (online)

Roberts, S., McQuillan, A., Reece, S., & Aigrain, S. 2013, MNRAS, 435, 3639 (arXiv:1308.3644)

Swift, J. J., Johnson, J. A., Morton, T. D., *et al.* 2013, ApJ, 764, 105 (arXiv:1301.0023)

Tabachnik, S., & Tremaine, S. 2002, MNRAS, 335, 151 (arXiv:astro-ph/0107482)

Traub, W. A. 2012, ApJ, 745, 20 (arXiv:1109.4682)

Tremaine, S., & Dong, S. 2012, AJ, 143, 94 (arXiv:1106.5403)

Youdin, A. N. 2011, ApJ, 742, 38 (arXiv:1105.1782)

This manuscript was prepared with the AAS LaTeX macros v5.2.

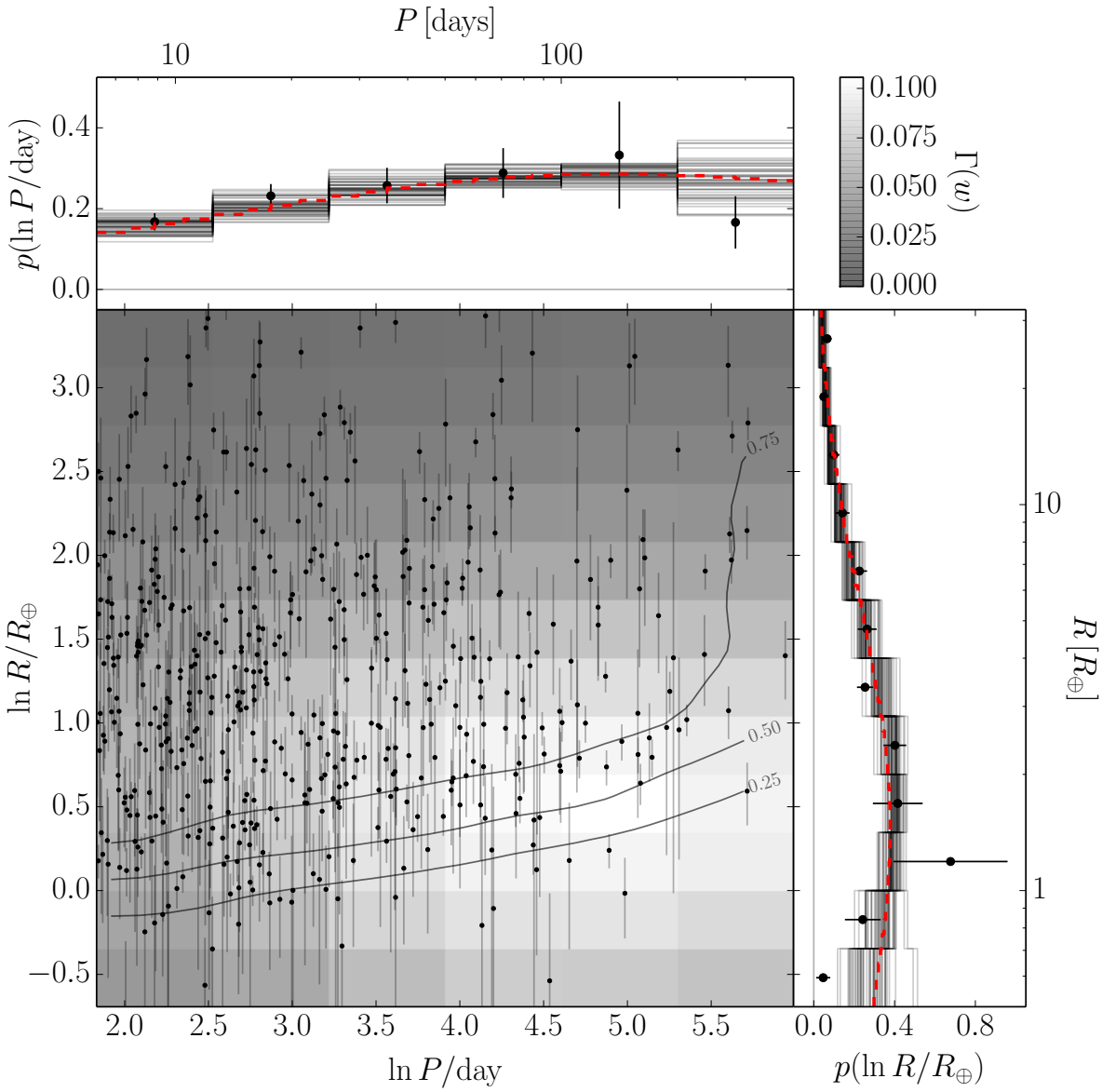

Fig. 1.— **Simulated data**. Inferences about the rate density based on the simulated catalog *Catalog A*. *Center:* the points with error bars show the exoplanet candidates in the simulated incomplete catalog, the contours show the survey completeness function (Petigura *et al.* 2013b), and the grayscale shows the median posterior occurrence surface. *Top and left:* the red dashed line shows the true distribution that was used to generate the catalog, the points with error bars show the results of the inverse-detection-efficiency procedure, and the histograms are posterior samples from the marginalized rate density as inferred by our method.

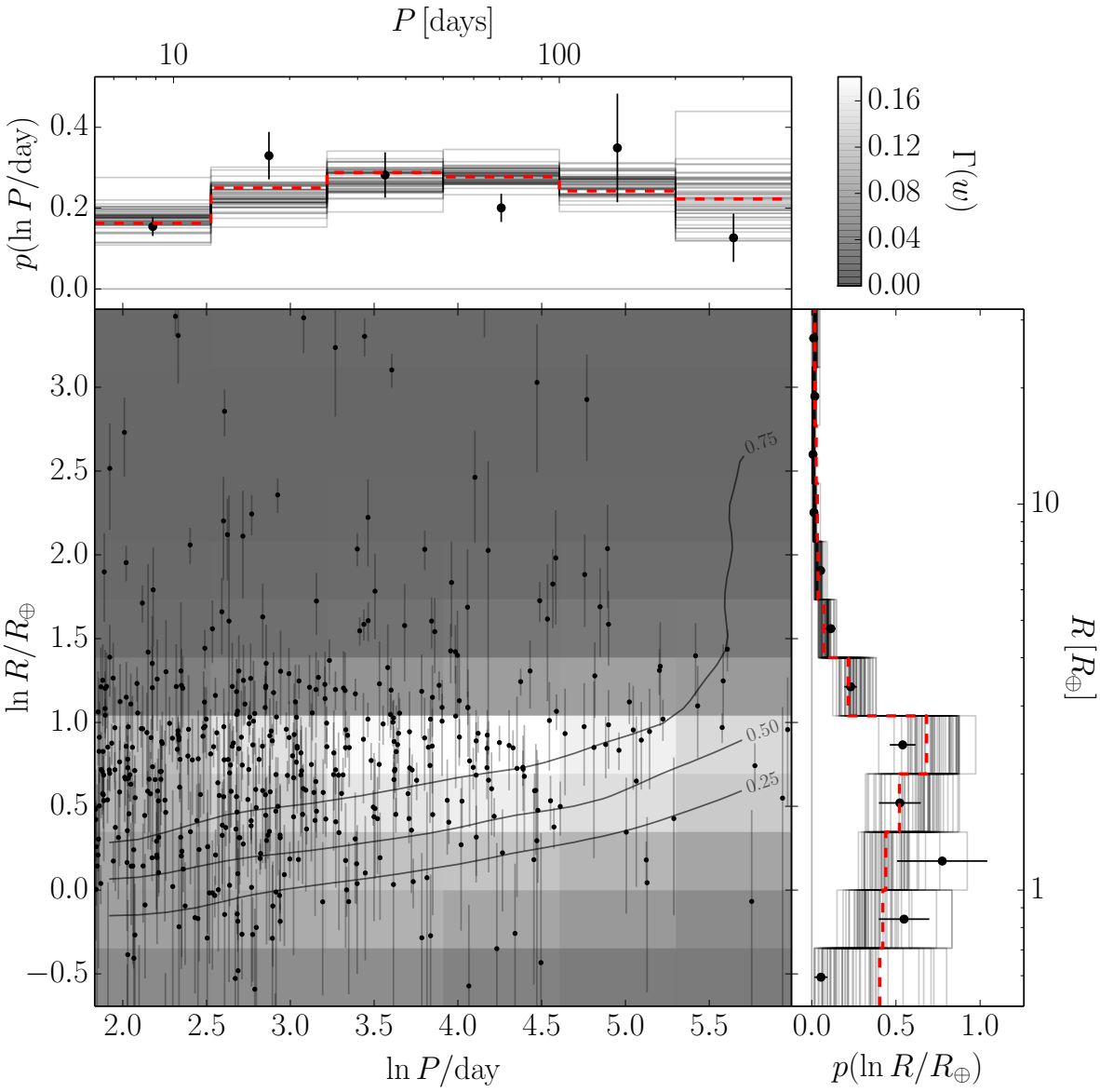

Fig. 2.— **Simulated data**. The same as Figure 1 for *Catalog B*.

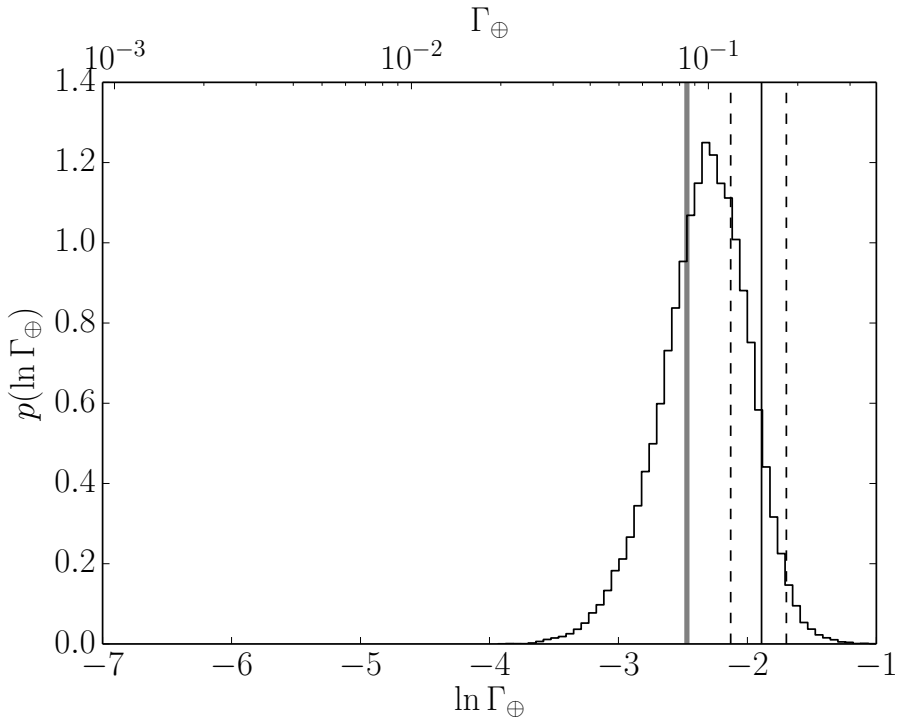

Fig. 3.— **Simulated data**. The extrapolated rate density of Earth analogs $\Gamma_\oplus$ as inferred by the different techniques applied to the *Catalog A* simulation. Applying the method used by Petigura *et al.* (2013b) gives a constraint indicated by the vertical black line with error bars shown as dashed lines. The histogram is the MCMC estimate of our posterior constraint on this rate density and the true value is indicated as the thick gray vertical line.

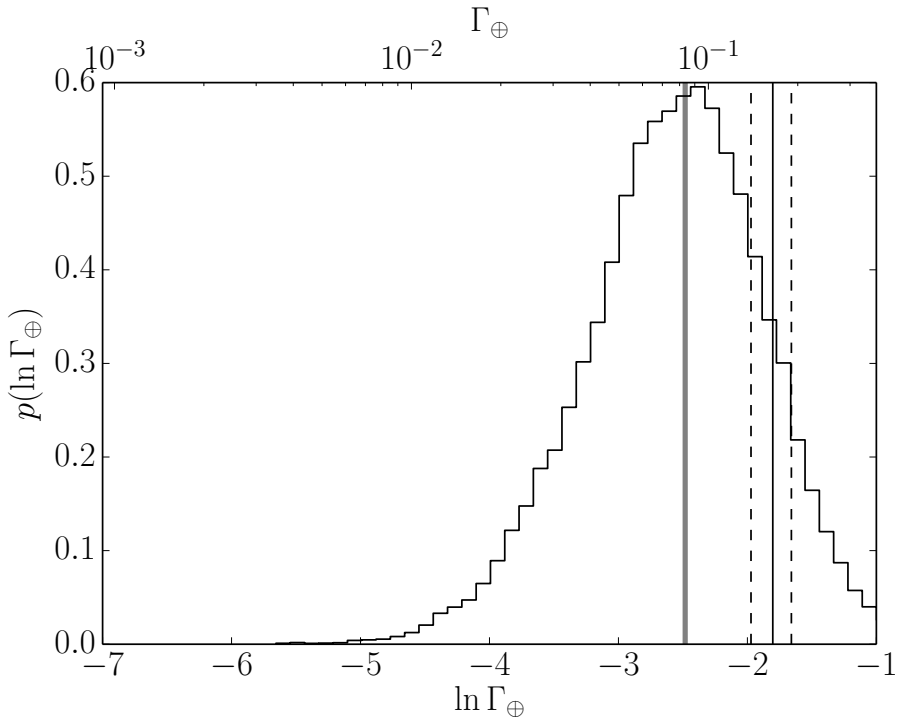

Fig. 4.— **Simulated data**. The same as Figure 3 for *Catalog B*.

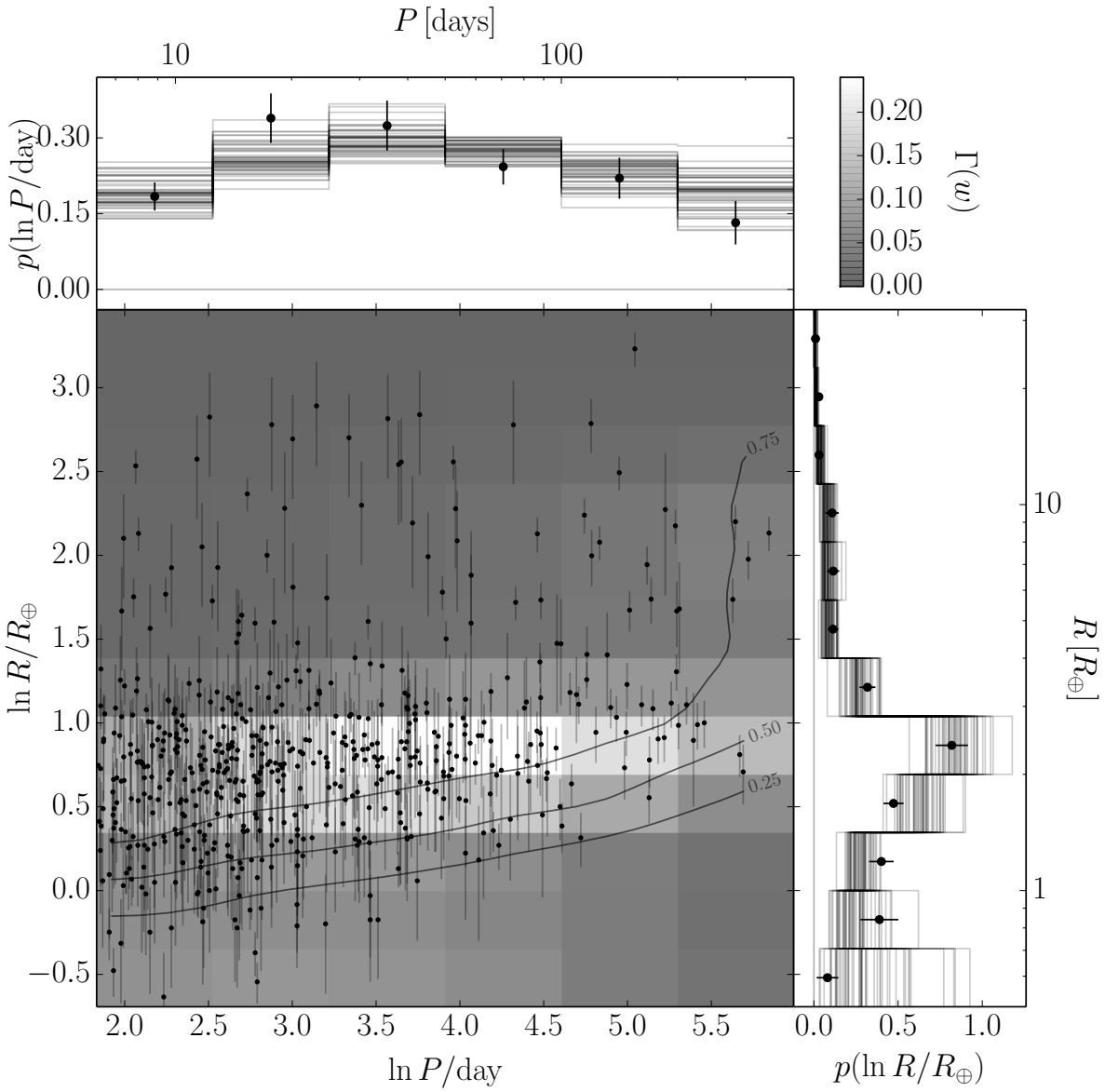

Fig. 5.— **Real data**. The same as Figure 1 when applied to the observed data from Petigura *et al.* (2013b). *Center:* the points with error bars show the catalog measurements, the contours show the survey completeness function, and the grayscale shows the median posterior occurrence surface. *Top and left:* the points with error bars show the results of the inverse-detection-efficiency procedure, and the histograms are posterior samples from the marginalized rate density as inferred by our method.

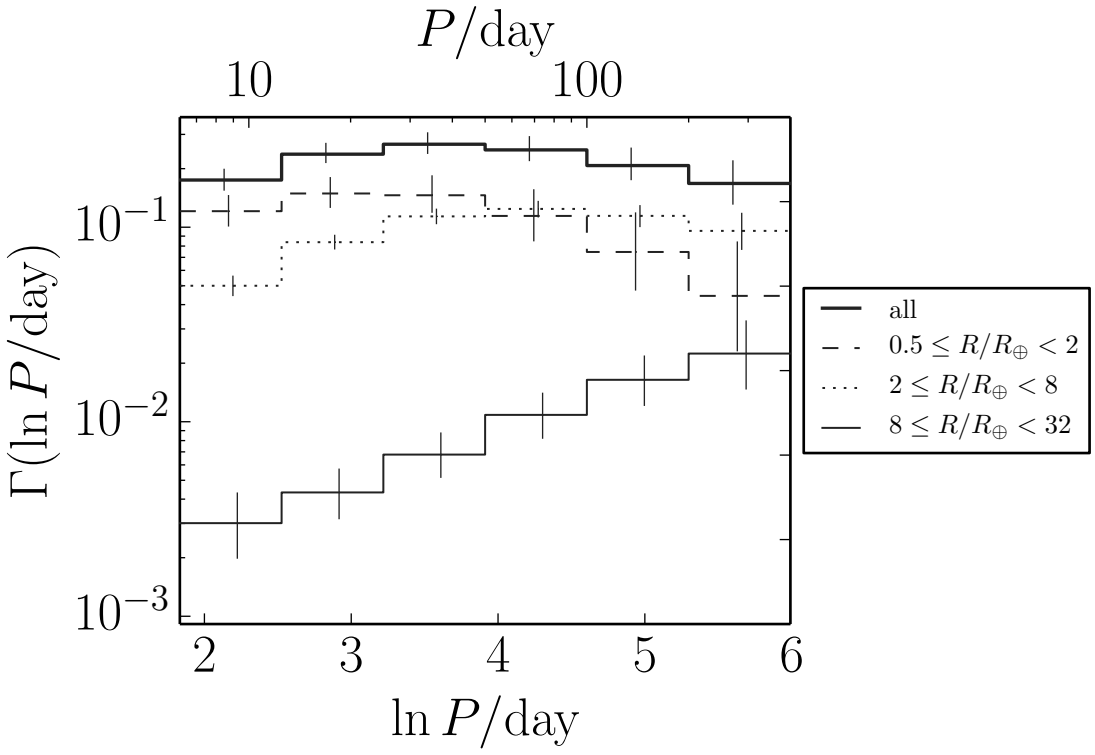

Fig. 6.— **Real data**. The occurrence rate density as a function of logarithmic period integrated over bins in logarithmic radius. The lines with error bars show the posterior sample median and 68th percentile and the line style specifies the radius bin. The period distribution for the largest planets in the sample ($8 \leq R/R_\oplus < 32$) continues to increase (as a function of $\ln P$) for all periods while the distribution seems to flatten and turn over at periods around 50 days.

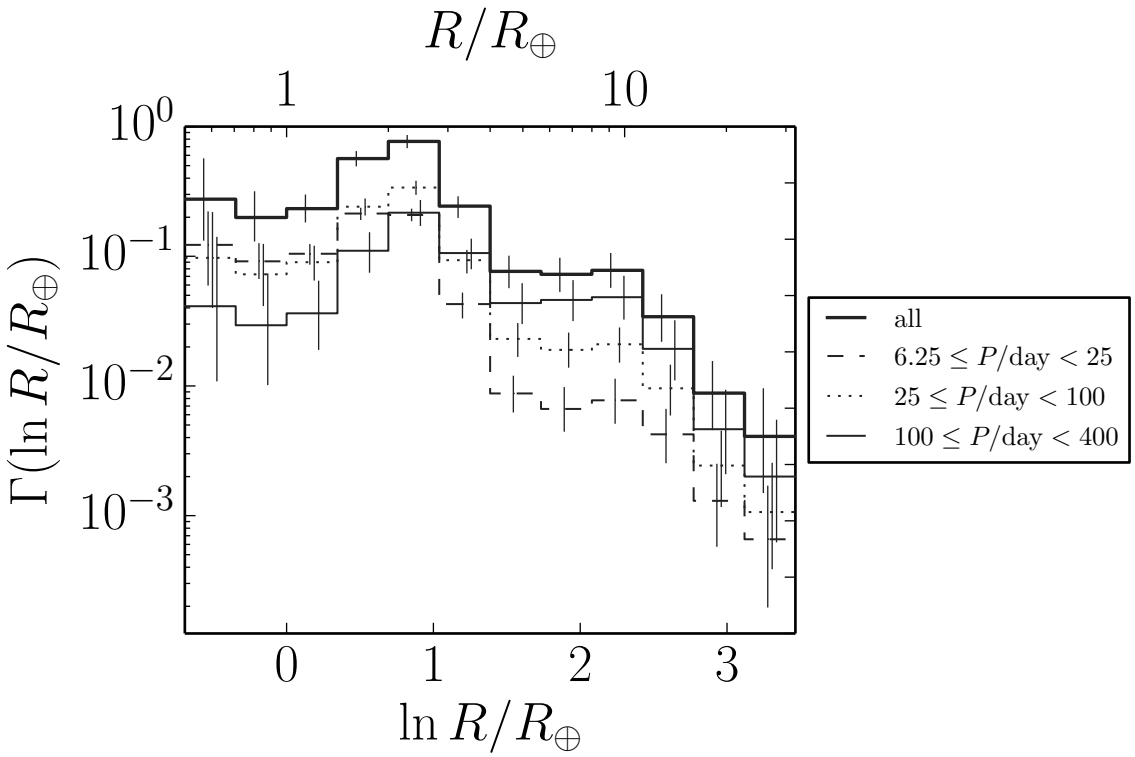

Fig. 7.— **Real data**. The occurrence rate density as a function of logarithmic radius integrated over bins in logarithmic period. The lines with error bars show the posterior sample median and 68th percentile and the line style specifies the period bin. The distributions in all the period bins are qualitatively consistent and there are plausibly features near $R \sim 3\,R_\oplus$ and $R \sim 10\,R_\oplus$.

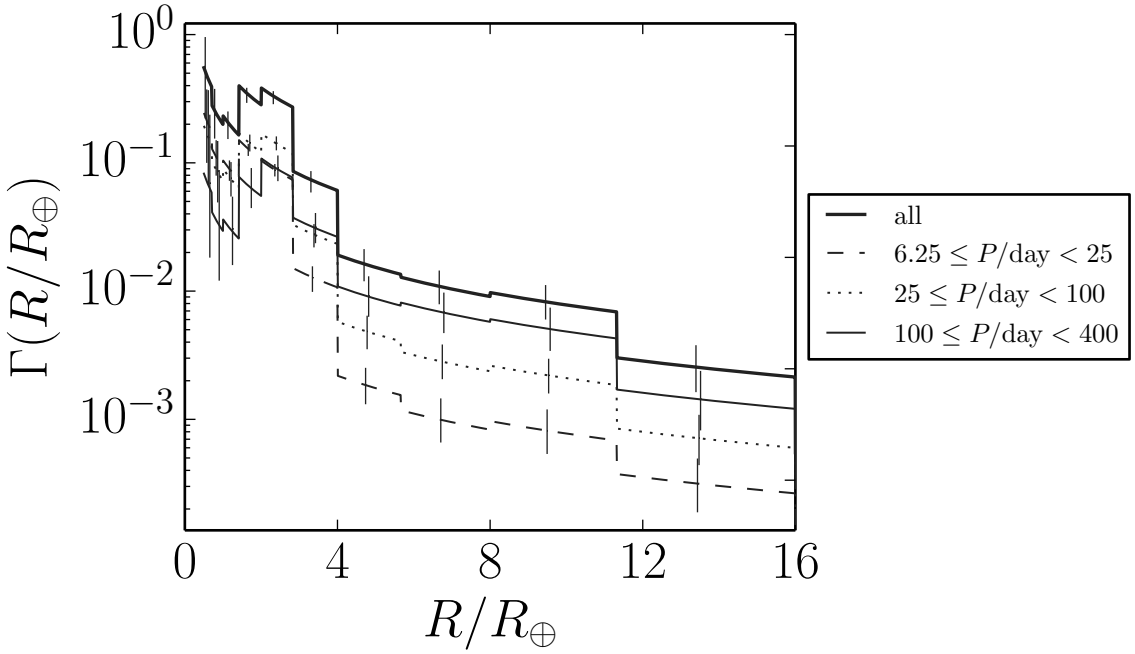

Fig. 8.— **Real data**. The same as Figure 7 but presented as a density in radius instead of logarithmic radius.

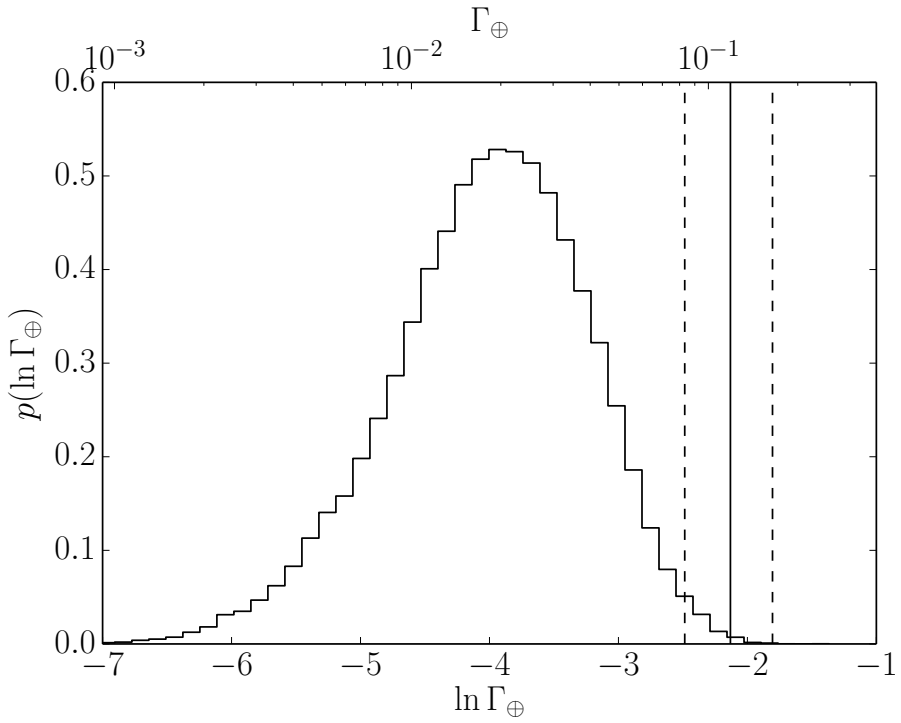

Fig. 9.— The extrapolated rate density of Earth analogs $\Gamma_\oplus$ (the same as Figure 3 but applied to the catalog from Petigura *et al.* 2013b). The histogram is the MCMC estimate of our posterior constraint on this rate density. The vertical black line with error bars shown as dashed lines is the result from Petigura *et al.* (2013b) converted to a rate density by dividing by their bin volume.

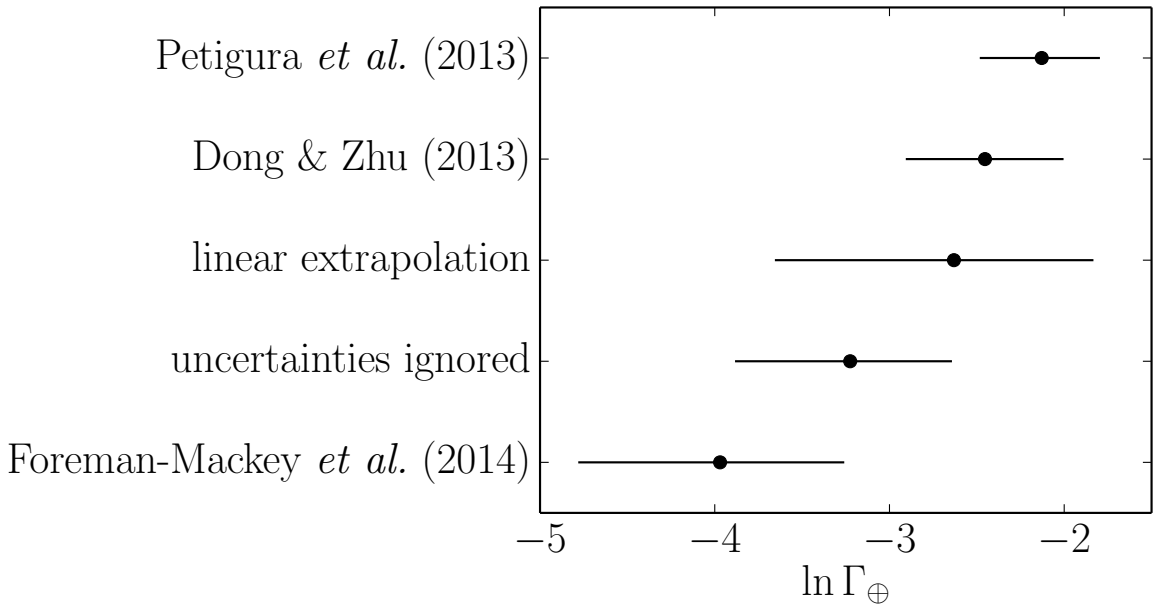

Fig. 10.— Comparison of various estimates of $\Gamma_\oplus$. From the top, the first value is the number published by Petigura *et al.* (2013b) and converted to consistent units. The second point shows the value implied by the power law model for the occurrence rate of $1 - 2\,R_\oplus$ planets from Dong & Zhu (2013). The point labeled "linear extrapolation" is the result of modeling the distribution of small planets ($1 - 2\,R_\oplus$) on long periods ($50 - 400$ days) but allowing the period distribution to be *linear* instead of *uniform*. The "uncertainties ignored" value is given by applying the model developed in this *Article* but with the error bars on radius artificially set to zero. Finally, the bottom point is the result of our full analysis.