Methods for the detection and characterization of exoplanets and their population

by

Daniel Foreman-Mackey

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Physics

New York University

May 2015

Professor David W. Hogg

Acknowledgements

First, a huge thanks is in order for my advisor, David Hogg, for these years of support, encouragement, and friendship. I couldn't have imagined a better advisor for me and I'm looking forward to many years of collaboration. Hogg's enthusiasm is infectious and our lively discussions over lunch, coffee, or beer have convinced me to attempt a career in academia and made me a better scientist and person.

I would also like to thank my other dissertation readers, Jonathan Goodman and Micheal Blanton, for taking the time to read my thesis and provide useful suggestions and clarifications that improved the presentation of this work.

One of my favorite parts of studying astronomy is the community and the opportunities to meet and get to know amazing people from around the world. Over the course of my Ph.D. research, I have had the opportunity to collaborate with more people than I can possibly list here. I am especially grateful for the summers that I spent in Germany with Bernhard Schölkopf and Hans-Walter Rix; the garden chats in Tübingen and Heidelberg were some of the highlights of my scientific career thus far. I have also enjoyed and benefited from conversations and collaborations with Brendon Brewer, Tom Barclay, Charlie Conroy, Bekki Dawson, Eric Ford, Morgan Fouesneau, Jonathan Goodman, Dustin Lang, Phil Marshall, Ben Montet, Tim Morton, Mike O'Neil, Adrian Price-Whelan, Elisa Quintana, Jonathan Sick, Dan Weisz, and many others.

My time spent living in NYC wouldn't have been the same without my eclectic (and perfectly eccentric) group of housemates and friends. In particular, I'd like to thank Mira, Stefan, Nate, Claire, Katherine, Jacob, Silas, and Stumm for all the good times on Dean Street. I would also like to thank Ruth for being there for me when I needed her and sharing in life and in science, even from the opposite side of the ocean.

Finally, I would like to thank my family, Annie, Clarke, and Elaine for their continued

love, encouragement, and support every step of the way. I would never have made it here without them!

Organizations The results in this dissertation are based, in large part, on observations make by the *Kepler* Mission and subsequent *K2* Mission with the same instrument. None of my work would have been possible without the tireless efforts of the scientists and engineers on the *Kepler* team. This work also relied heavily on the public data and literature resources provided by the Mikulski Archive for Space Telescopes, the NASA Exoplanet Archive, and the NASA Astrophysics Data System.

I would like to thank the organizers and attendees of the SAMSI workshop "Modern Statistical and Computational Methods for Analysis of *Kepler* Data". This workshop introduced me to the *Kepler* data and drastically lowered the barrier to entry into the field of exoplanets. The connections made at this workshop have led to many important and long-lasting friendships and collaborations.

Abstract

The study of exoplanets has been revolutionized in recent years thanks, in large part, to new data collected by NASA's *Kepler* Mission. The Mission has enabled the discovery of thousands of planets orbiting stars throughout the Galaxy. These discoveries span orders of magnitude in physical parameter space but many of the most physically interesting questions remain open. The deepest of these questions is: how common are planetary systems like our own Solar System? In this dissertation, I approach this question from several different angles and make inferences about the frequency and distribution of planets based on the large, publicly-available datasets from the *Kepler* and *K2* Missions.

I develop two powerful and practical methods for mining for planetary transit signals in the hundreds of thousands of stellar light curves measured by *Kepler*. The first method is designed to find planets using the data from the *K2* phase of the Mission where systematics introduced by the instrument dominate the measurements. Applying this method to the first publicly available dataset from *K2*, I published more than thirty new exoplanet candidates. The second transit search technique is designed to find transits of planets with orbital periods longer than the four year baseline of the *Kepler* Mission. These are interesting planets because they are expected to have the largest dynamical influence on the formation and evolution of their planetary systems but, to date, no systematic search for these signals has been published. I demonstrate that this method is robust and tractable and make predictions for the planet yields in the *Kepler* dataset.

I derive a general framework for making justified probabilistic inferences about the population of planets based on noisy and incomplete catalogs of exoplanet measurements. Applying this to a previously published catalog of exoplanets orbiting stars like our Sun, I measure the joint period–radius distribution of these planets taking into account survey selection effects and the large measurement uncertainties. Despite the fact that this catalog includes

no true Earth analogs, I use the detected systems and weak smoothness assumptions about the underlying distribution to make a probabilistic estimate of the frequency of Earth-like planets.

The main contributions of this dissertation are the development of methods for probabilistic and the release of open source implementations. One of these methods is *emcee*, a method for Markov Chain Monte Carlo (MCMC) sampling of probability distributions. MCMC has been a popular method for approximate inference in astronomy for well over a decade but most implementations require extensive hand tuning in order to achieve acceptable performance on all but the simplest problems. Thanks to its affine-invariant sampling algorithm the *emcee* method performs efficiently for many real problems in astronomy. The code has an active user base and online community of contributors.

Contents

List of Figures

List of Tables

Introduction

Over the past few years, the field of extrasolar planet (exoplanet) research has really taken off thanks, in large part, to the exquisite time series photometry measured by the *Kepler* Mission (Borucki et al., 2010). The Mission enabled the discovery of thousands of planets and planet candidates outside the Solar System (Rowe et al., 2015). The zoo of planetary systems is extremely diverse—with sizes, masses, and orbital periods spanning orders of magnitude—and the statistics are now sufficient to test theories of planet formation and evolution.

The *Kepler* Mission has changed the face of exoplanet research because of its photometric precision and the sheer volume of the dataset. In order to discover small planets that serendipitously transit their host stars, the *Kepler* spacecraft was designed to monitor the brightness of about 150,000 stars in one $10^\circ \times 10^\circ$ patch of the sky nearly continuously—at a half-hour cadence—for more than three years with a relative precision of a few parts-per-million for the brightest stars. The Mission surpassed its fiducial goals and took data for over 4 years before two of the reaction wheels used to stabilize the pointing failed in the Spring of 2013.

Despite the fact that most planets never transit their host star—based on geometric effects alone—and the fact that transit surveys are most sensitive to large planets on short orbits, the discoveries made in the *Kepler* dataset and careful characterization of the selection effects

and search completeness have enabled detailed studies of the true underlying distribution of planets over a wide range in parameter space (examples include Howard et al., 2012; Petigura et al., 2013a; Foreman-Mackey et al., 2014; Dressing & Charbonneau, 2015, and Chapter 2). These observational studies of the population of exoplanets are arguably the ultimate goal of the *Kepler* Mission because they open the door to direct comparison with theories of planet formation and evolution.

The different constraints on the intrinsic rate and distribution of planets differ in detail but several overarching results are solid. The evidence suggests that every cool M-star has at least one planet in orbit (Dressing & Charbonneau, 2013, 2015) and more than half of the other main sequence stars should have planetary systems (Howard et al., 2012; Fressin et al., 2013; Petigura et al., 2013a; Foreman-Mackey et al., 2014). Of these planets, the most intrinsically common are in the "super-Earth" or "mini-Neptune" range from about twice to four-times the radius of Earth. A combination of radial velocity follow-up and hierarchical inference methods indicate that most of these mini-Neptunes gaseous instead of rocky (Weiss & Marcy, 2014; Rogers, 2015) but since there are no planets like this in the Solar System, understanding them is crucial to our theories of planetary system formation.

One shortcoming of the *Kepler* Mission was that it only targeted one field and in that frame, the main focus was on relatively faint F, G, and K dwarf stars. This target selection was chosen to enable the study of long-period planets and the discovery of Earth-sized planets orbiting Sun-like stars. Unfortunately many of these stars and their planetary systems are not amenable to radial velocity follow-up because the star is too faint to achieve the required velocity precision or the expected velocity amplitude is too small to detect. In the Summer of 2014, the *Kepler* instrument was re-purposed and it began taking data in a mode called *K2* with substantial degraded pointing accuracy (Howell et al., 2014). Because of technical constraints, *K2* targets a different field in the ecliptic plane every three months.

This means that it can target stars in different environments and focus on gathering the census of planets orbiting bright, nearby stars. It has been demonstrated that the data from *K2* can reach precisions comparable to the original Mission and that it can be used to discover transiting exoplanets (Vanderburg & Johnson, 2014; Vanderburg et al., 2014; Crossfield et al., 2015; Foreman-Mackey et al., 2015, and Chapter 3). The discoveries made using *K2*—and the upcoming *TESS* Mission—improve our knowledge of the population of exoplanets, especially those planets that orbit the cool M-stars that were not prioritized by the *Kepler* Mission. These discoveries also present excellent targets for radial velocity follow-up and even spectroscopic observations of their atmospheres using the planned *James Webb Space Telescope*.

The technical problem of searching for transits in the massive datasets produced by a time series mission like *Kepler* is a hard one. The relative change in brightness caused by the transit of planet in front of its host star is given by the area ratio between the planet and star (Winn, 2010). Therefore, when an Earth-sized planet transits a Sun-like star, the amplitude of the signal is smaller than 100 parts-per-million. What's more, in the case of the Earth's orbit, this signal would only last for a little over half a day, once every year. Add to this the fact that most light curves are fraught with signals induced by stellar variability (Basri et al., 2013), spot activity (McQuillan et al., 2014), and instrumental effects (Stumpe et al., 2012; Smith et al., 2012) with amplitudes far exceeding most transits. In order to find transits, we must, therefore, develop methods for efficiently and robustly mining large sets of light curves for tiny, sparse signals. Nearly all transit search algorithms rely on some sort of matched filter that is made insensitive to noise by pre-processing the light curves to remove the trends or by designing an estimator that is insensitive to these effects (Kovács et al., 2002, 2005; Berta et al., 2012; Petigura et al., 2013a; Foreman-Mackey et al., 2015).

Despite the attempts made to develop search algorithms that are robust to systematics

and variability, all automated search results are completely dominated by false signals induced by poorly characterized noise in the light curves. In practice, automatic removal of these events has not been demonstrated to be sufficient—although the results are starting to look promising (Jenkins et al., 2014)—and all published catalogs of planet candidates are manually vetted. This means that the published list of candidates is *chosen by a person—or group of people—going through the data by hand*. This method is not efficient or scalable so a substantial set of heuristic filtering is applied to the candidate list even before anyone looks at the light curves. One of the standard filters is to only consider candidates with at least three observed transits (for example Petigura et al., 2013a; Burke et al., 2014; Rowe et al., 2015). This greatly restricts the range of parameter space that can be search. In particular, these methods will miss any planets on orbits longer than a fraction of the survey baseline.

While transit surveys present the most effective means for systematic exoplanet characterization, their use is limited by existing transit search methodologies for planets with long orbital periods. In many cases, massive long-period planets dominate the dynamics of the planetary systems—like Jupiter in the Solar System—but their existence is completely missed by *Kepler*. This shortcoming becomes even more severe for *K2* and *TESS*, transit surveys with shorter baselines. The final Chapter of this dissertation presents a novel method for transit search designed specifically to discover and quantify these important planets.

The study of exoplanets and their population has been driven by the public *Kepler* dataset and, in particular, by methods and software solutions developed by graduate students and young researchers around the world to squeeze all the available information out of the existing dataset. This dissertation presents methods developed with exactly these goals in mind. Each Chapter is accompanied by an open source implementation of the method and code to reproduce the results and figures. Of these projects, the most popular is the Markov Chain Monte Carlo implementation *emcee* (Foreman-Mackey et al., 2013, and Chapter 1).

With nearly 300 citations at the time of writing[1] and an active community on *GitHub*[2], *emcee* has enabled many modest and ambitious probabilistic inferences across astrophysics.

Chapters 1 and 2 have both been refereed and published in the astronomical literature. Chapter 3 has been submitted to *The Astrophysical Journal* and updated in response to the referee's comments. Chapter 4 is in preparation for submission. All of these Chapters were co-authored with collaborators but the majority of the work and writing in each Chapter is mine. Here, I describe my specific contributions to each Chapter:

1. For Chapter 1, I generalized the algorithm proposed by Goodman & Weare (2010) through discussions with Jonathan Goodman and David Hogg. I implemented the algorithm with contributions from Dustin Lang and wrote the paper with some additions by David Hogg.

2. For Chapter 2, I developed the project idea in collaboration with David Hogg and Timothy Morton. I then implemented the project and wrote the paper with contributions from David Hogg.

3. Of the published Chapters, Chapter 3 was the most collaborative. I developed the idea for the algorithm building on previous work with David Hogg, Dun Wang, and Bernhard Schölkopf. Using this algorithm, I wrote the code to search for transits in the *K2* Campagin 1 dataset and deployed it on the NYU HPC Butinah cluster[3]. I wrote the majority of the paper with Sections contributed by Ben Montet and Timothy Morton.

4. The fundamental ideas underlying Chapter 4 were developed through discussions with Bernhard Schölkopf and David Hogg. The implementation and text are mine.

[1] `http://adsabs.harvard.edu/cgi-bin/nph-ref_query?bibcode=2013PASP..125..306F&refs=CITATIONS&db_key=AST`

[2] `https://github.com/dfm/emcee`

[3] `http://nyuad.nyu.edu/en/research/infrastructure-and-support/nyuad-hpc.html`

Chapter 1

emcee: The MCMC Hammer

This Chapter is joint work with David W. Hogg (NYU), Jonathan Goodman (NYU), and Dustin Lang (Princeton/CMU) published in *Publications of the Astronomical Society of the Pacific* as Foreman-Mackey et al. (2013).

1.1 Chapter abstract

We introduce a stable, well tested Python implementation of the affine-invariant ensemble sampler for Markov chain Monte Carlo (MCMC) proposed by Goodman & Weare (2010). The code is open source and has already been used in several published projects in the astrophysics literature. The algorithm behind *emcee* has several advantages over traditional MCMC sampling methods and it has excellent performance as measured by the autocorrelation time (or function calls per independent sample). One major advantage of the algorithm is that it requires hand-tuning of only 1 or 2 parameters compared to $\sim N^2$ for a traditional algorithm in an N-dimensional parameter space. In this Chapter, we describe the algorithm and the details of our implementation. Exploiting the parallelism of the ensemble method, *emcee* permits *any* user to take advantage of multiple CPU cores without extra effort. The

code is available online at `http://dan.iel.fm/emcee` under the MIT License.

1.2 Introduction

Probabilistic data analysis—including Bayesian inference—has transformed scientific research in the past decade. Many of the most significant gains have come from numerical methods for approximate inference, especially Markov chain Monte Carlo (MCMC). For example, many problems in cosmology and astrophysics[1] have directly benefited from MCMC because the models are often expensive to compute, there are many free parameters, and the observations are usually low in signal-to-noise.

Probabilistic data analysis procedures involve computing and using either the posterior probability density function (PDF) for the parameters of the model or the likelihood function. In some cases it is sufficient to find the maximum of one of these, but it is often necessary to understand the posterior PDF in detail. MCMC methods are designed to sample from—and thereby provide sampling approximations to—the posterior PDF efficiently even in parameter spaces with large numbers of dimensions. This has proven useful in too many research applications to list here but the results from the NASA Wilkinson Microwave Anisotropy Probe (WMAP) cosmology mission provide a dramatic example (for example, Dunkley et al., 2005).

Arguably the most important advantage of Bayesian data analysis is that it is possible to *marginalize* over nuisance parameters. A nuisance parameter is one that is required in order to model the process that generates the data, but is otherwise of little interest. Marginalization is the process of integrating over all possible values of the parameter and hence

[1]The methods and discussion in this Chapter have general applicability, but we will mostly present examples from astrophysics and cosmology, the fields in which we have most experience

propagating the effects of uncertainty about its value into the final result. Often we wish to marginalize over all nuisance parameters in a model. The exact result of marginalization is the marginalized probability function $p(\Theta|D)$ of the set (list or vector) of model parameters Θ given the set of observations D

$$p(\Theta|D) = \int p(\Theta, \alpha|D)\ \mathrm{d}\alpha \quad , \tag{1.1}$$

where α is the set (list or vector) of nuisance parameters. Because the nuisance parameter set α can be very large, this integral is often extremely daunting. However, a MCMC-generated sampling of values (Θ_t, α_t) of the model and nuisance parameters from the joint distribution $p(\Theta, \alpha|D)$ automatically provides a sampling of values Θ_t from the marginalized PDF $p(\Theta|D)$.

In addition to the problem of marginalization, in many problems of interest the likelihood or the prior is the result of an expensive simulation or computation. In this regime, MCMC sampling is very valuable, but it is even *more* valuable if the MCMC algorithm is efficient, in the sense that it does not require many function evaluations to generate a statistically independent sample from the posterior PDF. The methods presented here are designed for efficiency.

Most uses of MCMC in the astrophysics literature are based on slight modifications to the Metropolis-Hastings (M–H) method (introduced below in Section 1.3). Each step in a M–H chain is proposed using a compact proposal distribution centered on the current position of the chain (normally a multivariate Gaussian or something similar). Since each term in the covariance matrix of this proposal distribution is an unspecified parameter, this method has $N\,[N+1]/2$ tuning parameters (where N is the dimension of the parameter space). To make matters worse, the performance of this sampler is very sensitive to these

tuning parameters and there is no fool-proof method for choosing the values correctly. As a result, many heuristic methods have been developed to attempt to determine the optimal parameters in a data-driven way (for example, Gregory, 2005; Dunkley et al., 2005; Widrow et al., 2008). Unfortunately, these methods all require a lengthy "burn-in" phase where shorter Markov chains are sampled and the results are used to tune the hyperparameters. This extra cost is unacceptable when the likelihood calls are computationally expensive.

The problem with traditional sampling methods can be visualized by looking at the simple but highly anisotropic density

$$p(\mathbf{x}) \propto f\left(-\frac{(x_1 - x_2)^2}{2\,\epsilon} - \frac{(x_1 + x_2)^2}{2}\right) \tag{1.2}$$

which would be considered difficult (in the small-ϵ regime) for standard MCMC algorithms. In principle, it is possible to tune the hyperparameters of a M–H sampler to make this sampling converge quickly, but if the dimension is large and calculating the density is computationally expensive the tuning procedure becomes intractable. Also, since the number of parameters scales as $\sim N^2$, this problem gets much worse in higher dimensions. Equation (1.2) can, however, be transformed into the much easier problem of sampling an isotropic density by an *affine transformation* of the form

$$y_1 = \frac{x_1 - x_2}{\sqrt{\epsilon}}\,, \qquad y_2 = x_1 + x_2 \quad . \tag{1.3}$$

This motivates affine invariance: an algorithm that is *affine invariant* performs equally well under all linear transformations; it will therefore be insensitive to covariances among parameters.

Goodman & Weare (2010, hereafter GW10) proposed an affine invariant sampling algorithm (Section 1.3) with only two hyperparameters to be tuned for performance. Hou et al.

(2012) were the first group to implement this algorithm in astrophysics. The implementation presented here is an independent effort that has already proved effective for many projects in the astronomical literature[2]. In what follows, we summarize the algorithm from GW10 and the implementation decisions made in *emcee*. We also describe the small changes that must be made to the algorithm to parallelize it.

1.3 The algorithm

A complete discussion of MCMC methods is beyond the scope of this Chapter. Instead, the interested reader is directed to a classic reference like MacKay (2003) and we will summarize some key concepts below.

The general goal of MCMC algorithms is to draw M samples $\{\Theta_i\}$ from the posterior probability density

$$p(\Theta, \alpha|D) = \frac{1}{Z}\, p(\Theta, \alpha)\, p(D|\Theta, \alpha) \quad , \tag{1.4}$$

where the prior distribution $p(\Theta, \alpha)$ and the likelihood function $p(D|\Theta, \alpha)$ can be relatively easily (but not necessarily quickly) computed for any particular value of (Θ_i, α_i). The normalization $Z = p(D)$ is independent of Θ and α once we have chosen the form of the generative model. This means that it is possible to sample from $p(\Theta, \alpha|D)$ without computing Z — unless one would like to compare the validity of two different generative models. This is important because Z is generally very expensive to compute.

Once the samples produced by MCMC are available, the marginalized constraints on Θ can be approximated by the histogram of the samples projected into the parameter subspace spanned by Θ. In particular, this implies that the expectation value of a function of the

[2] `http://adsabs.harvard.edu/cgi-bin/nph-ref_query?bibcode=2013PASP..125..306F`

model parameters $f(\Theta)$ is

$$\langle f(\Theta) \rangle = \int p(\Theta|D)\, f(\Theta)\, \mathrm{d}\Theta \approx \frac{1}{M} \sum_{i=1}^{M} f(\Theta_i) \quad . \tag{1.5}$$

Generating the samples Θ_i is a non-trivial process unless $p(\Theta, \alpha, D)$ is a very specific analytic distribution (for example, a Gaussian). MCMC is a procedure for generating a random walk in the parameter space that, over time, draws a representative set of samples from the distribution. Each point in a Markov chain $X(t_i) = [\Theta_i, \alpha_i]$ depends only on the position of the previous step $X(t_{i-1})$.

The Metropolis-Hastings (M–H) Algorithm The simplest and most commonly used MCMC algorithm is the M–H method (Algorithm 1; MacKay, 2003; Gregory, 2005; Press, 2007; Hogg et al., 2010a). The iterative procedure is as follows: (1) given a position $X(t)$ sample a proposal position Y from the transition distribution $Q(Y; X(t))$, (2) accept this proposal with probability

$$\min\left(1, \frac{p(Y|D)}{p(X(t)|D)} \frac{Q(X(t); Y)}{Q(Y; X(t))}\right) \quad . \tag{1.6}$$

The transition distribution $Q(Y; X(t))$ is an easy-to-sample probability distribution for the proposal Y given a position $X(t)$. A common parameterization of $Q(Y; X(t))$ is a multivariate Gaussian distribution centered on $X(t)$ with a general covariance tensor that has been tuned for performance. It is worth emphasizing that if this step is accepted $X(t+1) = Y$; Otherwise, the new position is set to the previous one $X(t+1) = X(t)$ (in other words, the position $X(t)$ is *repeated in the chain*).

The M–H algorithm converges (as $t \to \infty$) to a stationary set of samples from the distribution but there are many algorithms with faster convergence and varying levels of

implementation difficulty. Faster convergence is preferred because of the reduction of computational cost due to the smaller number of likelihood computations necessary to obtain the equivalent level of accuracy. The inverse convergence rate can be measured by the autocorrelation function and more specifically, the integrated autocorrelation time (see Section 1.4). This quantity is an estimate of the number of steps needed in the chain in order to draw independent samples from the target density. A more efficient chain has a shorter autocorrelation time.

Algorithm 1 The procedure for a single Metropolis-Hastings MCMC step.

Draw a proposal $Y \sim Q(Y; X(t))$
$q \leftarrow [p(Y)\, Q(X(t); Y)]/[p(X(t))\, Q(Y; X(t))]$ *// This line is generally expensive*
$r \leftarrow R \sim [0, 1]$
if $r \leq q$ **then**
    $X(t+1) \leftarrow Y$
else
    $X(t+1) \leftarrow X(t)$
end if

The stretch move GW10 proposed an affine-invariant ensemble sampling algorithm informally called the "stretch move." This algorithm significantly outperforms standard M–H methods producing independent samples with a much shorter autocorrelation time (see Section 1.4 for a discussion of the autocorrelation time). For completeness and for clarity of notation, we summarize the algorithm here and refer the interested reader to the original paper for more details. This method involves simultaneously evolving an ensemble of K *walkers* $S = \{X_k\}$ where the proposal distribution for one walker k is based on the current positions of the $K - 1$ walkers in the *complementary ensemble* $S_{[k]} = \{X_j, \forall j \neq k\}$. Here, "position" refers to a vector in the N-dimensional, real-valued parameter space.

To update the position of a walker at position X_k, a walker X_j is drawn randomly from

the remaining walkers $S_{[k]}$ and a new position is proposed:

$$X_k(t) \to Y = X_j + Z\left[X_k(t) - X_j\right] \tag{1.7}$$

where Z is a random variable drawn from a distribution $g(Z = z)$. It is clear that if g satisfies

$$g(z^{-1}) = z\, g(z), \tag{1.8}$$

the proposal of Equation (1.7) is symmetric. In this case, the chain will satisfy detailed balance if the proposal is accepted with probability

$$q = \min\left(1,\, Z^{N-1}\,\frac{p(Y)}{p(X_k(t))}\right) \quad , \tag{1.9}$$

where N is the dimension of the parameter space. This procedure is then repeated for each walker in the ensemble *in series* following the procedure shown in Algorithm 2.

GW10 advocate a particular form of $g(z)$, namely

$$g(z) \propto \begin{cases} \dfrac{1}{\sqrt{z}} & \text{if } z \in \left[\dfrac{1}{a}, a\right], \\ 0 & \text{otherwise} \end{cases} \tag{1.10}$$

where a is an adjustable scale parameter that GW10 set to 2.

The parallel stretch move It is tempting to parallelize the stretch move algorithm by simultaneously advancing each walker based on the state of the ensemble instead of evolving the walkers in series. Unfortunately, this subtly violates detailed balance. Instead, we must split the full ensemble into two subsets ($S^{(0)} = \{X_k,\, \forall k = 1, \ldots, K/2\}$ and $S^{(1)} = \{X_k,\, \forall k = K/2 + 1, \ldots, K\}$) and simultaneously update all the walkers in $S^{(0)}$ — using the

Algorithm 2 A single stretch move update step from GW10

for $k = 1, \ldots, K$ **do**
  Draw a walker X_j at random from the complementary ensemble $S_{[k]}(t)$
  $z \leftarrow Z \sim g(z)$, Equation (1.10)
  $Y \leftarrow X_j + z\,[X_k(t) - X_j]$
  $q \leftarrow z^{N-1}\,p(Y)/p(X_k(t))$ *// This line is generally expensive*
  $r \leftarrow R \sim [0, 1]$
  if $r \leq q$, Equation (1.9) **then**
    $X_k(t+1) \leftarrow Y$
  else
    $X_k(t+1) \leftarrow X_k(t)$
  end if
end for

stretch move procedure from Algorithm 2 — based *only* on the positions of the walkers in the other set ($S^{(1)}$). Then, using the new positions $S^{(0)}$, we can update $S^{(1)}$. In this case, the outcome is a valid step for all of the walkers. The pseudocode for this procedure is shown in Algorithm 3. This code is similar to Algorithm 2 but now the computationally expensive inner loop (starting at line 2 in Algorithm 3) can be run in parallel.

The performance of this method — quantified by the autocorrelation time — is comparable to the serial stretch move algorithm but the fact that one can now take advantage of generic parallelization makes it extremely powerful.

1.4 Tests

Judging the convergence and performance of an algorithm is a non-trival problem and there is a huge associated literature (see, for example, Cowles & Carlin, 1996, for a review). In astrophysics, spectral methods have been used extensively (for example Dunkley et al., 2005). Below, we advocate for one such method: the autocorrelation time. The autocorrelation time is especially applicable because it is an affine invariant measure of the performance.

First, however, we should take note of another extremely important measurement: the

Algorithm 3 The parallel stretch move update step

for $i \in \{0, 1\}$ **do**
  for $k = 1, \ldots, K/2$ **do**
    // This loop can now be done in parallel for all k
    Draw a walker X_j at random from the complementary ensemble $S^{(\sim i)}(t)$
    $X_k \leftarrow S_k^{(i)}$
    $z \leftarrow Z \sim g(z)$, Equation (1.10)
    $Y \leftarrow X_j + z\,[X_k(t) - X_j]$
    $q \leftarrow z^{n-1}\, p(Y)/p(X_k(t))$
    $r \leftarrow R \sim [0, 1]$
    if $r \leq q$, Equation (1.9) **then**
      $X_k(t + \frac{1}{2}) \leftarrow Y$
    else
      $X_k(t + \frac{1}{2}) \leftarrow X_k(t)$
    end if
  end for
  $t \leftarrow t + \frac{1}{2}$
end for

acceptance fraction a_f. This is the fraction of proposed steps that are accepted. There appears to be no agreement on the optimal acceptance rate but it is clear that both extrema are unacceptable. If $a_f \sim 0$, then nearly all proposed steps are rejected, so the chain will have very few independent samples and the sampling will not be representative of the target density. Conversely, if $a_f \sim 1$ then nearly all steps are accepted and the chain is performing a random walk with no regard for the target density so this will also not produce representative samples. As a rule of thumb, the acceptance fraction should be between 0.2 and 0.5 (for example, Gelman et al., 1996). For the M–H algorithm, these effects can generally be counterbalanced by decreasing (or increasing, respectively) the eigenvalues of the proposal distribution covariance. For the stretch move, the parameter a effectively controls the step size so it can be used to similar effect. In our tests, it has never been necessary to use a value of a other than 2, but we make no guarantee that this is the optimal value.

Autocorrelation time The autocorrelation time is a direct measure of the number of evaluations of the posterior PDF required to produce independent samples of the target density. GW10 show that the stretch-move algorithm has a significantly shorter autocorrelation time on several non-trivial densities. This means that fewer PDF computations are required—compared to a M–H sampler—to produce the same number of independent samples.

The autocovariance function of a time series $X(t)$ is

$$C_f(T) = \lim_{t\to\infty} \operatorname{cov}\left[f\left(X(t+T)\right), f\left(X(t)\right)\right]. \tag{1.11}$$

This measures the covariances between samples at a time lag T. The value of T where $C_f(T) \to 0$ measures the number of samples that must be taken in order to ensure independence. In particular, the relevant measure of sampler efficiency is the integrated autocorrelation time

$$\tau_f = \sum_{T=-\infty}^{\infty} \frac{C_f(T)}{C_f(0)} = 1 + 2\sum_{T=1}^{\infty} \frac{C_f(T)}{C_f(0)}. \tag{1.12}$$

In practice, one can estimate $C_f(T)$ for a Markov chain of M samples as

$$C_f(T) \approx \frac{1}{M-T} \sum_{m=1}^{M-T} \left[f(X(T+m)) - \langle f\rangle\right] \left[f(X(m)) - \langle f\rangle\right]. \tag{1.13}$$

We advocate for the autocorrelation time as a measure of sampler performance for two main reasons. First, it measures a quantity that *we are actually interested in* when sampling in practice. The longer the autocorrelation time, the more samples that we must generate to produce a representative sampling of the target density. Second, the autocorrelation time is affine invariant. Therefore, it is reasonable to measure the performance and diagnose the convergence of the sampler on densities with different levels of anisotropy.

emcee can optionally calculate the autocorrelation time using the Python module *acor*[3] to estimate the autocorrelation time. This module is a direct port of the original algorithm (described by GW10) and implemented by those authors in C++.[4]

1.5 Discussion & tips

The goal of this project has been to make a sampler that is a useful tool for a large class of data analysis problems—a "hammer" if you will. If development of statistical and data-analysis understanding is the key goal, a user who is new to MCMC benefits enormously by writing her or his own Metropolis–Hastings code (Algorithm 1) from scratch before downloading *emcee*. For typical problems, the *emcee* package will perform better than any home-built M–H code (for all the reasons given above), but the intuitions developed by writing and tuning a self-built MCMC code cannot be replaced by reading this document and running this pre-built package. That said, once those intuitions are developed, it makes sense to switch to *emcee* or a similarly well engineered piece of code for performance on large problems.

Ensemble samplers like *emcee* require some thought for initialization. One general approach is to start the walkers at a sampling of the prior or spread out over a reasonable range in parameter space. Another general approach is to start the walkers in a very tight N-dimensional ball in parameter space around one point that is expected to be close to the maximum probability point. The first is more objective but, in practice, we find that the latter is much more effective if there is any chance of walkers getting stuck in low probability modes of a multi-modal probability landscape. The walkers initialized in the small ball will

[3] `http://github.com/dfm/acor`
[4] `http://www.math.nyu.edu/faculty/goodman/software/acor`

expand out to fill the relevant parts of parameter space in just a few autocorrelation times. A third approach would be to start from a sampling of the prior, and go through a "burn-in" phase in which the prior is transformed continuously into the posterior by increasing the "temperature." Discussion of this kind of annealing is beyond the scope of this document.

It is our present view that autocorrelation time is the best indicator of MCMC performance (the shorter the better), but there are several proxies. The easiest and simplest indicator that things are going well is the acceptance fraction; it should be in the 0.2 to 0.5 range (there are theorems about this for specific problems; for example Gelman et al., 1996). In principle, if the acceptance fraction is too low, you can raise it by decreasing the a parameter; and if it is too high, you can reduce it by increasing the a parameter. However, in practice, we find that $a = 2$ is good in essentially all situations. That means that when using *emcee if the acceptance fraction is getting very low, something is going very wrong.* Typically a low acceptance fraction means that the posterior probability is multi-modal, with the modes separated by wide, low probability "valleys." In situations like these, the best idea (though expensive of human time) is to split the space into disjoint single-mode regions and sample each one independently, combining the independently sampled regions "properly" (also expensive, and beyond the scope of this document) at the end. In previous work, we have advocated clustering methods to remove multiple modes (Hou et al., 2012). These work well when the different modes have *very* different posterior probabilities.

Another proxy for short autocorrelation time is large expected or mean squared jump distance (ESJD; Pasarica & Gelman 2010). The higher the ESJD the better; if walkers move (in the mean) a large distance per chain step then the autocorrelation time will tend to be shorter. The ESJD is not an affine-invariant measure of performance, and it doesn't have a trivial interpretation in terms of independent samples, so we prefer the autocorrelation time in principle. In practice, however, because the ESJD is a simple expectation value it can be

more robustly evaluated on short chains.

With *emcee* you want (in general) to *run with a large number of walkers*, like hundreds. In principle, there is no reason not to go large when it comes to walker number, until you hit performance issues. Although each step takes twice as much compute time if you double the number of walkers, it also returns to you twice as many independent samples per autocorrelation time. So go large. In particular, we have found that—in almost all cases of low acceptance fraction—increasing the number of walkers improves the acceptance fraction. The one disadvantage of having large numbers of walkers is that the burn-in phase (from initial conditions to reasonable sampling) can be slow; burn-in time is a few autocorrelation times; the total run time for burn-in scales with the number of walkers. These considerations, all taken together, suggest using the smallest number of walkers for which the acceptance fraction during burn-in is good, or the number of samples you want out at the end (see below), whichever is *greater*. A more ambitious project would be to increase the number of walkers after burn-in; this requires thought beyond the scope of this document; it can be accomplished by burning in a set of small ensembles and then merging them into a big ensemble for the final run.

One mistake many users of MCMC methods make is to take *too many* samples! If all you want your MCMC to do is produce one- or two-dimensional error bars on two or three parameters, then you only need dozens of independent samples. With ensemble sampling, you get this from a *single snapshot* or single timestep, provided that you are using dozens of walkers (and we would recommend that you use hundreds in most applications). The key point is that *you should run the sampler for a few (say 10) autocorrelation times.* Once you have run that long, no matter how you initialized the walkers, the set of walkers you obtain at the end should be an independent set of samples from the distribution, of which you rarely need many.

Another common mistake, of course, is to run the sampler for *too few* steps. You can identify that you haven't run for enough steps in a couple of ways: If you plot the parameter values in the ensemble as a function of step number, you will see large-scale variations over the full run length if you have gone less than an autocorrelation time. You will also see that if you try to measure the autocorrelation time (with, say, *acor*), it will give you a time that is always a significant fraction of your run time; it is only when the correlation time is much shorter (say by a factor of 10) than your run time that you are sure to have run long enough. The danger of both of these methods—an unavoidable danger at present—is that you can have a huge dynamic range in contributions to the autocorrelation time; you might think it is 30 when in fact it is 30 000, but you don't "see" the 30 000 in a run that is only 300 steps long. There is not much you can do about this; it is generic when the posterior is multi-modal: The autocorrelation time within each mode can be short but the mode–mode migration time can be long. See above on low acceptance ratio; in general when your acceptance ratio gets low your autocorrelation time is very, very long.

There are some cases where *emcee* won't perform as well as some more specialized sampling techniques. In particular, when the target density is multi-modal, walkers can become "stuck" in different modes. When this happens, the vector between walkers is no longer a good proposal direction. In these cases, the acceptance fraction and autocorrelation time can deteriorate quickly. While this is a fairly general problem, we find that in many applications the effect isn't actually very important. That being said, there are some problems where higher-end machinery (such as *DNest* Brewer et al., 2011) is necessary.

Another limitation to the stretch move and moves like it is that they implicitly assume that the parameters can be assembled into a vector-like object on which linear operations can be performed. This is not (trivially) true for parameters that have non-trivial constraints, like parameters that must be integer-valued or equivalent, or parameters that are subject to

deterministic non-linear constraints. Sometimes these issues can be avoided by reparameterization, but in some cases, samplers like *emcee* will not be useful, or might require clever or interesting improvements. The *emcee* package is open-source software; please push us changes!

1.6 Chapter acknowledgements

It is a pleasure to thank Eric Agol (UW), Jo Bovy (IAS), Brendon Brewer (Auckland), Jacqueline Chen (MIT), Alex Conley (Colorado), Will Meierjurgen Farr (Northwestern), Andrew Gelman (Columbia), John Gizis (Delaware), Fengji Hou (NYU), Jennifer Piscionere (Vanderbilt), Adrian Price-Whelan (Columbia), Hans-Walter Rix (MPIA), Jeremy Sanders (Cambridge), Larry Widrow (Queen's), and Joe Zuntz (Oxford) for helpful contributions to the ideas and code presented here.

Chapter 2

Exoplanet population inference and the abundance of Earth analogs from noisy, incomplete catalogs

This Chapter is joint work with David W. Hogg (NYU) and Timothy D. Morton (Princeton) published in *The Astrophysical Journal* as Foreman-Mackey et al. (2014).

2.1 Chapter abstract

No true extrasolar Earth analog is known. Hundreds of planets have been found around Sun-like stars that are either Earth-sized but on shorter periods, or else on year-long orbits but somewhat larger. Under strong assumptions, exoplanet catalogs have been used to make an extrapolated estimate of the rate at which Sun-like stars host Earth analogs. These studies are complicated by the fact that every catalog is censored by non-trivial selection effects and detection efficiencies, and every property (period, radius, *etc.*) is measured noisily. Here we

present a general hierarchical probabilistic framework for making justified inferences about the population of exoplanets, taking into account survey completeness and, for the first time, *observational uncertainties*. We are able to make fewer assumptions about the distribution than previous studies; we only require that the occurrence rate density be a smooth function of period and radius (employing a Gaussian process). By applying our method to synthetic catalogs, we demonstrate that it produces more accurate estimates of the whole population than standard procedures based on weighting by inverse detection efficiency. We apply the method to an existing catalog of small planet candidates around G dwarf stars (Petigura *et al.* 2013). We confirm a previous result that the radius distribution changes slope near Earth's radius. We find that the rate density of Earth analogs is about 0.02 (per star per natural logarithmic bin in period and radius) with large uncertainty. This number is much smaller than previous estimates made with the same data but stronger assumptions.

2.2 Introduction

NASA's *Kepler* mission has enabled the discovery of thousands of exoplanet candidates (Batalha et al. 2013; Burke et al. 2014). While many of these candidates have not been confirmed as bona fide planets, there is evidence that the false positive rate is low (Morton 2012; Fressin et al. 2013), enabling conclusions about the population of planets based on the catalog of candidates. Many of these planets orbit Sun-like stars (Petigura et al. 2013a), where the definition of Sun-like is given in terms of the star's temperature and surface gravity. Given these catalogs, it is interesting to ask what we can say about the population of exoplanets as a function of their physical parameters (period, radius, *etc.*). Observational constraints on the population can inform theories of planet formation and place probabilistic

bounds on the abundance of Earth analogs[1].

Petigura et al. (2013a) recently published an exoplanet population analysis based on an independent study of the *Kepler* light curves for 42,557 Sun-like stars. This study was especially novel because the authors developed their own planet search pipeline (*TERRA*; Petigura et al. 2013b) and determined the detection efficiency of their analysis empirically by injecting synthetic signals into real light curves measured by *Kepler*. The occurrence rate function determined by Petigura et al. (2013a) agrees qualitatively with previous studies of small planets orbiting Sun-like stars (Dong & Zhu 2013). In particular, both papers describe a "flattening" rate function (in logarithmic radius) for planets around Earth's radius. Even though no Earth analogs were discovered in their search, Petigura et al. (2013a) used the small candidates that they did find to place an extrapolated constraint on the frequency of Earth-like exoplanets, assuming a flat occurrence rate density in logarithmic period.

A very important component of any study of exoplanet populations is the treatment of detection efficiency. Speaking qualitatively, in a transit survey, small planets with long periods are much harder to detect than large planets orbiting close to their star. This effect is degenerate with any inferences about the rate density and it can be hard to constrain quantitatively. In practice, there are three methods for taking this effect into account: *(a)* making conservative cuts on the candidates and assuming that the resulting catalog is complete (Catanzarite & Shao 2011; Traub 2012; Tremaine & Dong 2012), *(b)* asserting an 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 2014), and *(c)* determining the detection efficiency empirically by injecting synthetic signals into the raw data and testing recovery (Christiansen et al. 2013;

[1]For our purposes, an "Earth analog" is an Earth-sized exoplanet orbiting a Sun-like star with a year-long period.

Petigura et al. 2013b,a).

There are two qualitatively different methods that are commonly used to infer the occurrence rate density from a catalog and a detection efficiency specification. The first is an intuitive method that we will refer to as "inverse-detection-efficiency" and the second is based on the likelihood function of the catalog given a parametric rate density. The inverse-detection-efficiency method involves making a histogram of the objects in the catalog where each point is weighted by its inverse detection probability. This method is very popular in the literature (Howard et al. 2012; Dong & Zhu 2013; Dressing & Charbonneau 2013; Swift et al. 2013; Petigura et al. 2013a) even though it is not motivated probabilistically. The alternative likelihood method models the catalog as a Poisson realization of the *observable* rate density of exoplanets taking the survey detection efficiencies and transit probabilities into account. This technique has been used to constrain parametric models—a broken power law, for example—for the occurrence rate density (Tabachnik & Tremaine 2002; Youdin 2011; Dong & Zhu 2013). In this Chapter, we start from the likelihood method but model the rate density non-parametrically as a piecewise-constant step function. Using this formulation of the problem, we derive a generalization that takes observational uncertainties into account. In Section 2.12, we show that the inverse-detection-efficiency method can be derived as a special case of the likelihood method in the limit of a smoothly varying completeness function.

In every previous study of exoplanet occurrence rates, the authors have assumed that the measurement uncertainties are negligible. This assumption is not justified because these uncertainties—especially on measurements (like exoplanet radius) that depend on the stellar parameters—can be large compared to the scales of interest. In this Chapter, we develop a flexible framework for probabilistic inference of exoplanet occurrence rate density that can be applied to incomplete catalogs with *non-negligible observational uncertainties*. Our

method takes the form of a hierarchical probabilistic (Bayesian) inference. We generalize the method introduced by Hogg et al. (2010b) to account for survey detection efficiencies. We run tests on simulated datasets—comparing results with the standard techniques that neglect observational uncertainties—and apply our method to a real catalog of small planets transiting Sun-like stars (Petigura et al. 2013a).

For the purposes of this Chapter, we make some strong assumptions, although we argue that they are weaker than the implicit assumptions in previous studies. None of these assumptions is necessary for the validity of our general method but they do simplify the specific procedures we employ. We assume that

- the candidates in the catalog are independent draws from an inhomogeneous Poisson process set by the censored occurrence rate density,
- every candidate is a real exoplanet (there are no false positives),
- the observational uncertainties on the physical parameters are non-negligible but known (the catalog provides probabilistic constraints on the parameters),
- the detection efficiency of the pipeline is known, and
- the *True*[2] occurrence rate density is *smooth*[3].

The first assumption—conditional independence of the candidates—is reasonable since the dataset that we consider explicitly includes only single transiting systems (Petigura et al. 2013a). The second assumption—neglecting false positives—is also strong and only weakly

[2]In this Chapter, we use "*True*" to describe an observable (for example, the exoplanet occurrence rate density) that would be trivially measured in the limit of very high signal-to-noise data. We use "true" to describe a simulation quantity with a value exactly known to us.

[3]We give our definition of "smooth" in more detail below but our model is very flexible so this is not a strong restriction.

justified by estimates of low false positive rates in the *Kepler* data (Fressin et al. 2013; Morton 2012). For this Chapter, we will neglect this issue and only comment on the effects but the prior distributions published by Fressin et al. (2013) could be directly applied in a generalization of our method.

We must emphasize one very important consequence of our assumptions. We assume that the catalog of exoplanet candidates is only missing planets with probabilities given by the empirical detection efficiency. In detail this must be false; the detection efficiency we use doesn't take into account the fact that the catalog doesn't include multiple transiting systems. A large fraction of the transiting planets discovered by the *Kepler* transit search pipeline are members of multiple transiting systems (see Lissauer et al. 2011, for example). Since Petigura et al. (2013a) only detected at most one planet per system, their catalog is actually a list of planet candidates *without a more detectable companion*. The global effects of this selection are not trivial and an in-depth discussion is beyond the scope of this Chapter but all of the results should be interpreted with this caveat in mind.

Conditioned on our assumptions and the choices made in the planet detection, vetting and characterization pipeline (Petigura et al. 2013b,a), we constrain the rate density of small exoplanets orbiting Sun-like stars. As part of this analysis we also place probabilistic constraints on the rate density[4] of Earth analogs $\Gamma_\oplus$, which we define as *the expected number of planets per star per natural logarithmic bin in period and radius, evaluated at the period and radius of Earth*

$$\Gamma_\oplus \quad = \quad \left.\frac{\mathrm{d}N}{\mathrm{d}\ln P\ \mathrm{d}\ln R}\right|_{R=R_\oplus,\, P=P_\oplus} \quad . \tag{2.1}$$

[4]In this Chapter, we use the word "rate" to indicate the dimensionless expectation value of a Poisson process and the words "rate density" to indicate a quantity that must be integrated over a finite bin in period and radius to deliver a rate.

Since no Earth analogs have been detected, this constraint requires an extrapolation in both period and radius. Petigura et al. (2013a) performed this extrapolation by assuming that the period distribution of planets in a small bin in radius is flat, obtaining $\Gamma_\oplus \approx 0.12$. We relax this assumption and extrapolate only by assuming that the occurrence rate density is a smooth function of period and radius; we find lower values for $\Gamma_\oplus$. We enforce the smoothness constraint by applying a flexible Gaussian process regularization to the bin heights.

In the next Section, we summarize the likelihood method for exoplanet population inference and in Section 2.4, we describe how to include the effects of observational uncertainties. The technical term for this procedure is *hierarchical inference* and while a general discussion of this field is beyond the scope of this Chapter, in Section 2.4, we present the basic probabilistic question and derive a computationally tractable inference procedure. In Section 2.5, we summarize the technique and derive the key equation for our method: Equation (2.11). We test our method on synthetic catalogs in Sections 2.7 and 2.8. In Section 2.9, we use the catalog of planet candidates and the empirically determined detection efficiency from Petigura et al. (2013a) to measure the occurrence rate density of small planets with long orbital periods.

Sections 2.3 and 2.4 provide a general pedagogical introduction to the methods used in this Chapter. Readers looking to *implement* a population inference are directed to Section 2.12 if measurement uncertainties are negligible or Section 2.5 (especially Equation 2.11) for problems with non-negligible uncertainties. Readers interested in our results—the inferred population of exoplanets and Earth-analogs—can safely skip to Section 2.9 and continue to the discussion in Section 2.10.

2.3 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 Chapter, 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)

$$p(\{\boldsymbol{w}_k\} \,|\, \boldsymbol{\theta}) \quad \propto \quad \exp\left(-\int \hat{\Gamma}_{\boldsymbol{\theta}}(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}\right) \prod_{k=1}^{K} \hat{\Gamma}_{\boldsymbol{\theta}}(\boldsymbol{w}_k) \quad . \tag{2.2}$$

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

$$\hat{\Gamma}_{\boldsymbol{\theta}}(\boldsymbol{w}) \quad = \quad Q_{\mathrm{c}}(\boldsymbol{w})\,\Gamma_{\boldsymbol{\theta}}(\boldsymbol{w}) \tag{2.3}$$

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. (2013a) catalog; it is not a description of the *True* rate density of exoplanets. Inasmuch as the detection efficiency $Q_{\rm c}(\boldsymbol{w})$ 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 Chapter, we will assume that the completeness function $Q_{\rm c}(\boldsymbol{w})$ 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

$$\Gamma_{\boldsymbol{\theta}}(\boldsymbol{w}) = \begin{cases} \exp(\theta_1) & \boldsymbol{w} \in \Delta_1, \\ \exp(\theta_2) & \boldsymbol{w} \in \Delta_2, \\ \dots & \\ \exp(\theta_J) & \boldsymbol{w} \in \Delta_J, \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

where the parameters θ_j are the log step heights and the bins Δ_j are fixed *a priori*. In Section 2.12, 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.

2.4 A brief introduction to hierarchical inference

The general question that we are trying to answer in this Chapter 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 (2.2) but when this approximation is no longer valid, we must instead compute the *marginalized likelihood*

$$p(\{\boldsymbol{x}_k\}\,|\,\boldsymbol{\theta}) \quad = \quad \int p(\{\boldsymbol{x}_k\}\,|\,\{\boldsymbol{w}_k\})\,p(\{\boldsymbol{w}_k\}\,|\,\boldsymbol{\theta})\,\mathrm{d}\{\boldsymbol{w}_k\} \tag{2.5}$$

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 (2.5)—up to an unimportant constant—without ever looking at the raw data again! This is important because the high-dimensional integral in Equation (2.5) 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

$$p(\boldsymbol{w}_k \,|\, \boldsymbol{x}_k,\, \boldsymbol{\alpha}) \quad = \quad \frac{p(\boldsymbol{x}_k \,|\, \boldsymbol{w}_k)\, p(\boldsymbol{w}_k \,|\, \boldsymbol{\alpha})}{p(\boldsymbol{x}_k \,|\, \boldsymbol{\alpha})} \tag{2.6}$$

of the parameters $\boldsymbol{w}_k$ conditioned on the observations of that object $\boldsymbol{x}_k$. The argument here is that the catalog was generated by running MCMC to draw posterior samples for the transit parameters—under an assumed prior distribution $p(\boldsymbol{w}_k \,|\, \boldsymbol{\alpha})$ and likelihood function $p(\boldsymbol{x}_k \,|\, \boldsymbol{w}_k)$—then the entry in the catalog is *a description of this probability distribution.* 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 (2.2).

Now, we can use these posterior measurements to simplify Equation (2.5) to a form that can, in many common cases, be evaluated efficiently. To find this result, multiply the integrand in Equation (2.5) by the trivial

$$1 \quad = \quad \frac{p(\{\boldsymbol{w}_k\} \,|\, \{\boldsymbol{x}_k\},\, \boldsymbol{\alpha})}{p(\{\boldsymbol{w}_k\} \,|\, \{\boldsymbol{x}_k\},\, \boldsymbol{\alpha})} \tag{2.7}$$

and use Equation (2.6) to factorize the denominator and find

$$\frac{p(\{\boldsymbol{x}_k\}\,|\,\boldsymbol{\theta})}{p(\{\boldsymbol{x}_k\}\,|\,\boldsymbol{\alpha})} \;=\; \int \frac{p(\{\boldsymbol{w}_k\}\,|\,\boldsymbol{\theta})}{p(\{\boldsymbol{w}_k\}\,|\,\boldsymbol{\alpha})}\, p(\{\boldsymbol{w}_k\}\,|\,\{\boldsymbol{x}_k\},\,\boldsymbol{\alpha})\,\mathrm{d}\{\boldsymbol{w}_k\} \tag{2.8}$$

with a few lines of algebra. 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 (2.6) can always be—and often are—propagated as a list of N samples $\{\boldsymbol{w}_k\}^{(n)}$ from the posterior

$$\{\boldsymbol{w}_k\}^{(n)} \;\sim\; p(\{\boldsymbol{w}_k\}\,|\,\{\boldsymbol{x}_k\},\,\boldsymbol{\alpha}) \quad . \tag{2.9}$$

We can use these samples and the Monte Carlo integral approximation to estimate the marginalized likelihood from Equation (2.8)—up to an irrelevant constant—as

$$p(\{\boldsymbol{x}_k\}\,|\,\boldsymbol{\theta}) \;\approx\; \frac{Z_{\boldsymbol{\alpha}}}{N}\sum_{n=1}^{N}\frac{p(\{\boldsymbol{w}_k\}^{(n)}\,|\,\boldsymbol{\theta})}{p(\{\boldsymbol{w}_k\}^{(n)}\,|\,\boldsymbol{\alpha})} \tag{2.10}$$

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 (2.10) could be a high variance estimator of Equation (2.8), 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 (2.10) becomes exact. Equation (2.10) is the *importance sampling approximation* to the integral in Equation (2.8) 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.

2.5 Model generalities

Now, we can substitute Equation (2.2) into Equation (2.8) and apply the importance sampling approximation (Equation 2.10) to derive the following expression for the marginalized likelihood

$$\frac{p(\{\boldsymbol{x}_k\}\,|\,\boldsymbol{\theta})}{p(\{\boldsymbol{x}_k\}\,|\,\boldsymbol{\alpha})} \approx \exp\left(-\int \hat{\Gamma}_{\boldsymbol{\theta}}(\boldsymbol{w})\,\mathrm{d}\boldsymbol{w}\right) \prod_{k=1}^{K} \frac{1}{N_k} \sum_{n=1}^{N_k} \frac{\hat{\Gamma}_{\boldsymbol{\theta}}(\boldsymbol{w}_k^{(n)})}{p(\boldsymbol{w}_k^{(n)}\,|\,\boldsymbol{\alpha})} \tag{2.11}$$

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

$$\boldsymbol{w}_k{}^{(n)} \sim p(\boldsymbol{w}_k\,|\,\boldsymbol{x}_k,\,\boldsymbol{\alpha}) \tag{2.12}$$

as described in the previous section. Equation (2.11) 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 Chapter, 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 (2.11)—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

$$p(\boldsymbol{\theta} \,|\, \{\boldsymbol{x}_k\}) \quad \propto \quad p(\boldsymbol{\theta})\, p(\{\boldsymbol{x}_k\} \,|\, \boldsymbol{\theta}) \tag{2.13}$$

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 2.8—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

$$\begin{aligned} p(\boldsymbol{\theta}) &= p(\boldsymbol{\theta} \,|\, \mu,\, \lambda) \\ &= \mathcal{N}\left[\boldsymbol{\theta};\, \mu\,\mathbf{1},\, \mathrm{K}(\{\Delta_j\},\, \boldsymbol{\lambda})\right] \end{aligned} \tag{2.14}$$

where $\mathcal{N}(\cdot;\, \mu\,\mathbf{1},\, \mathrm{K})$ is a J-dimensional Gaussian[5] 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

$$\mathrm{K}_{ij} = \lambda_0 \exp\left(-\frac{1}{2}\left[\Delta_i - \Delta_j\right]^{\mathrm{T}} \Sigma^{-1} \left[\Delta_i - \Delta_j\right]\right) \tag{2.15}$$

where Σ^{-1} is the diagonal matrix

$$\Sigma^{-1} = \begin{pmatrix} 1/\lambda_P^2 & 0 \\ 0 & 1/\lambda_R^2 \end{pmatrix} \quad . \tag{2.16}$$

The Gaussian process model for the step heights given in Equation (2.14) 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

$$p(\boldsymbol{\theta},\, \mu,\, \boldsymbol{\lambda} \,|\, \{\boldsymbol{x}_k\}) \propto p(\mu,\, \boldsymbol{\lambda})\, p(\boldsymbol{\theta} \,|\, \mu,\, \boldsymbol{\lambda})\, p(\{\boldsymbol{x}_k\} \,|\, \boldsymbol{\theta}) \quad . \tag{2.17}$$

[5] J is the total number of bins.

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 & Adams 2010) for sampling the step heights $\boldsymbol{\theta}$ from the density in Equation (2.17). In practice, for problems with this specific structure, ESS outperforms more traditional MCMC methods commonly employed in astrophysics (e.g., Foreman-Mackey et al. 2013). Our implementation is adapted from Jo Bovy's BSD licensed ESS code[6]. To simultaneously marginalize over the hyperparameter choice, we use the Metropolis–Hastings update from Algorithm 1 in Murray & 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 $\boldsymbol{\theta}$.

2.6 Data and completeness function

Using an independent exoplanet search and characterization pipeline, Petigura et al. (2013a) published a catalog of 603 planet candidates orbiting stars in their "Sun-like" sample of *Kepler* targets. For each candidate, Petigura et al. (2013a) 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 un-

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

certainties 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 (2.10). For the rest of this Chapter, 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. (2013a) did not publish or share posterior samples of their measurements of the physical parameter (Equation 2.9). 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. (2013a) 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. (2013a) 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 Chapter, 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. 2013a) 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[7], the median transit probability in the selected sample of stars is 5.061% so we model the transit probability[8] as a function of period as

$$Q_\mathrm{t}(P) \quad = \quad 0.05061 \left[\frac{P}{10\,\mathrm{days}}\right]^{-2/3} \quad . \tag{2.18}$$

This expression is clearly only valid for $P \gtrsim 1.4\,\mathrm{days}$ but the dataset that we are using

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

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

(Petigura et al. 2013a) 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 (2.18) is the assumption 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 Chapter. 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.

2.7 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. (2013a) as given. In general, Equation (2.2) 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. (2013a) 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

$$\Gamma_{\boldsymbol{\theta}}(\ln P, \ln R) \quad = \quad \Gamma_{\boldsymbol{\theta}}^{(P)}(\ln P)\, \Gamma_{\boldsymbol{\theta}}^{(R)}(\ln R) \tag{2.19}$$

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. (2013a) in the range that they considered and then smoothly extrapolated outside that range. The catalogs generated from these two models are shown in Figure 2.1 and Figure 2.2, respectively and the data are available online[9].

For each catalog, we directly apply both the inverse-detection-efficiency procedure as implemented by Petigura et al. 2013a[10] and our probabilistic method, marginalizing over the hyperparameters of the Gaussian process regularization. Figure 2.1 and Figure 2.2 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

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

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

used to generate the catalogs. Figure 2.1 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.

In the second example, the true rate density includes a sharp feature chosen to reproduce the result published by Petigura et al. (2013a). In this case, Figure 2.2 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.

2.8 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[11], evaluated at the location of Earth:

$$\Gamma_\oplus = \Gamma(\ln P_\oplus, \ln R_\oplus) \tag{2.20}$$

$$= \left.\frac{\mathrm{d}N}{\mathrm{d}\ln P\ \mathrm{d}\ln R}\right|_{R=R_\oplus,\, P=P_\oplus} . \tag{2.21}$$

[11] In this Chapter, we adopt the Petigura et al. (2013a) sample of G-stars as our definition of "Sun-like".

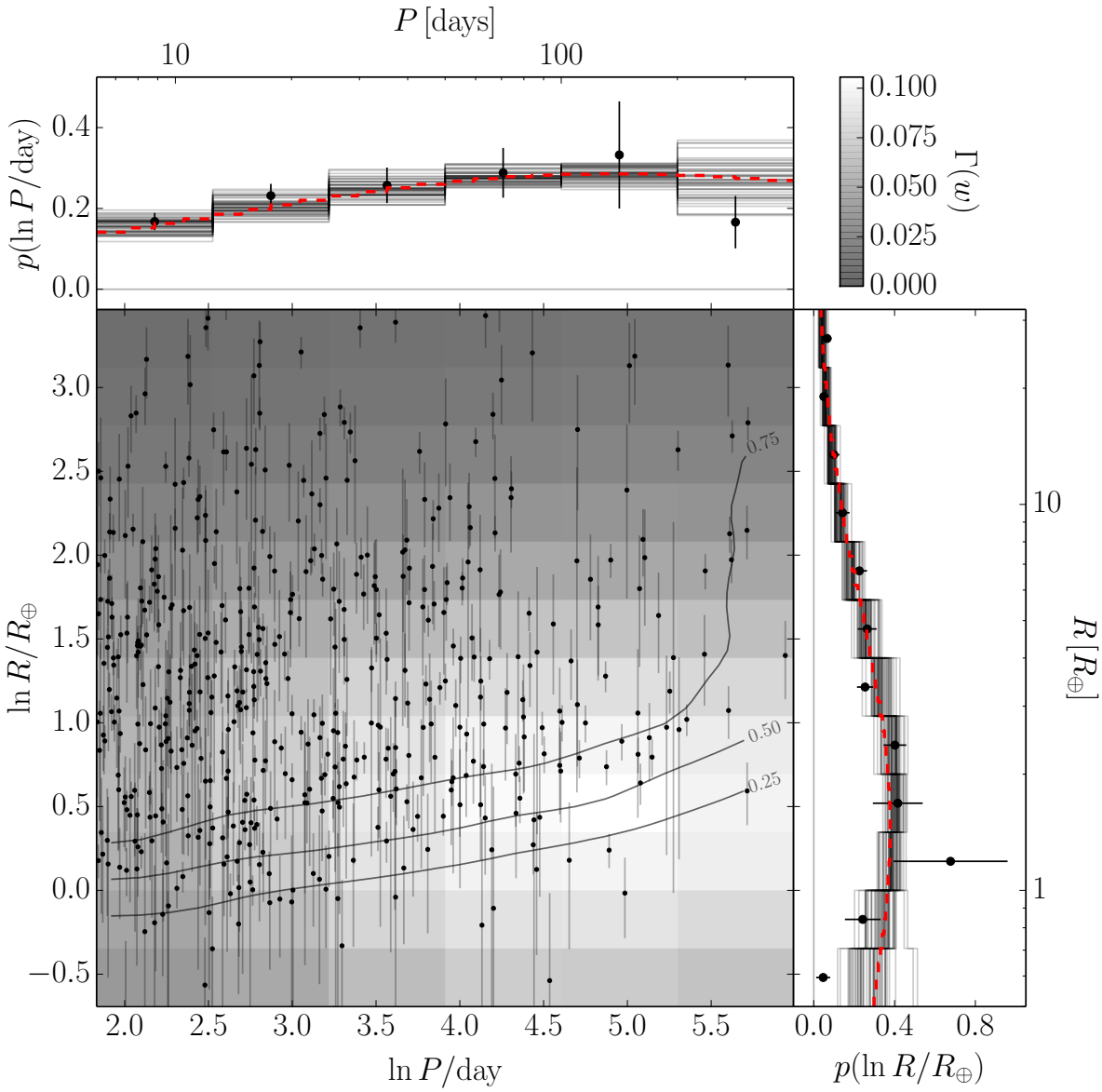

Figure 2.1: **Simulated data**. Inferences about the rate density based on the simulated catalog *Catalog A*. *Center:* the points with error bars show the exoplanet candidates in the simulated incomplete catalog, the contours show the survey completeness function (Petigura et al. 2013a), and the grayscale 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.

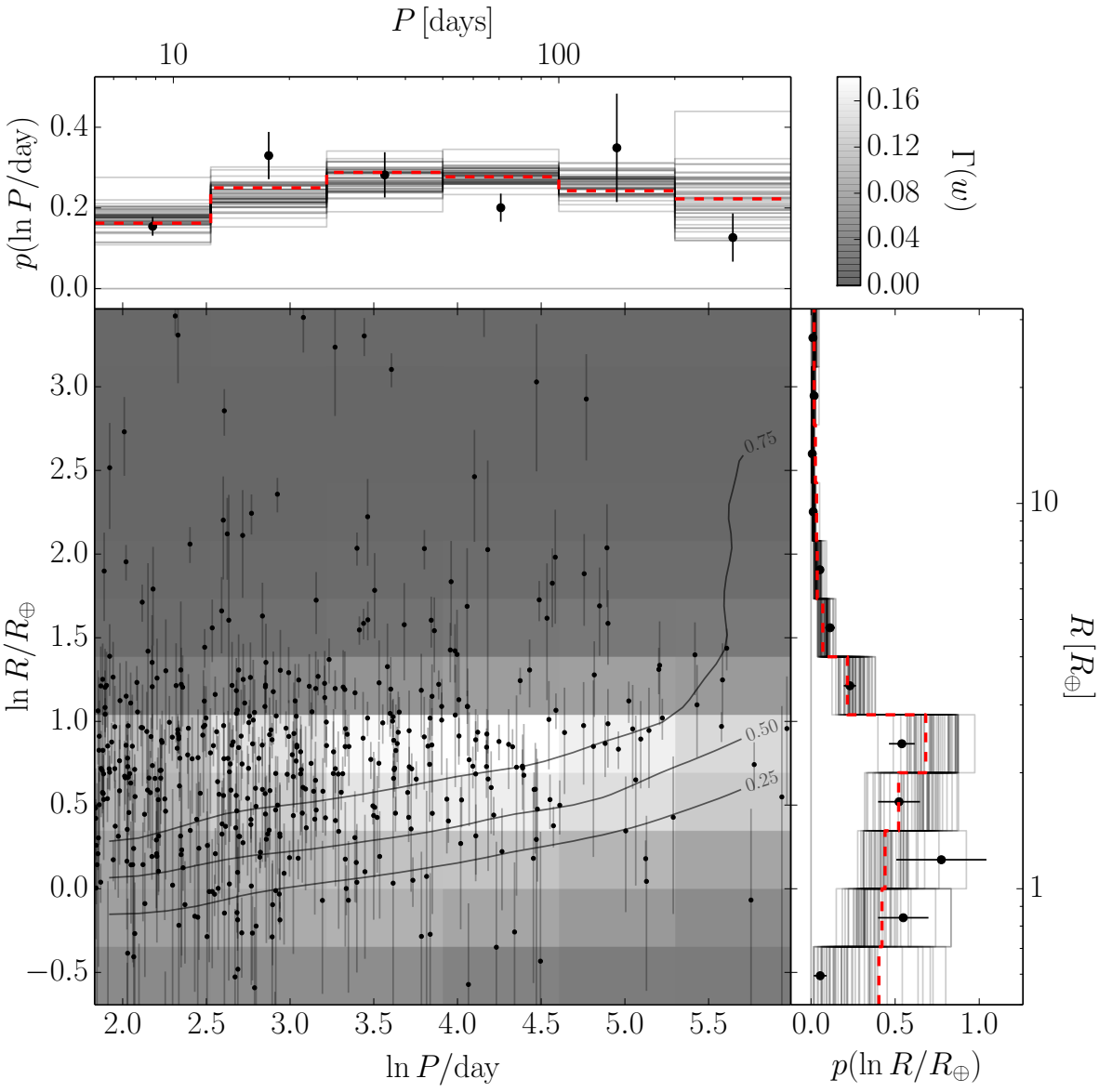

Figure 2.2: **Simulated data**. The same as Figure 2.1 for *Catalog B*.

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 (2.20), we use the symbol Γ instead of the more commonly used η since we 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 Chapter.

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. (2013a) 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 2.5. 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. (2013a). 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. (2013a) 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 (2.20), Petigura et al. (2013a) 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 2.3 and 2.4 compare our results and the results of the Petigura et al. (2013a) 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.

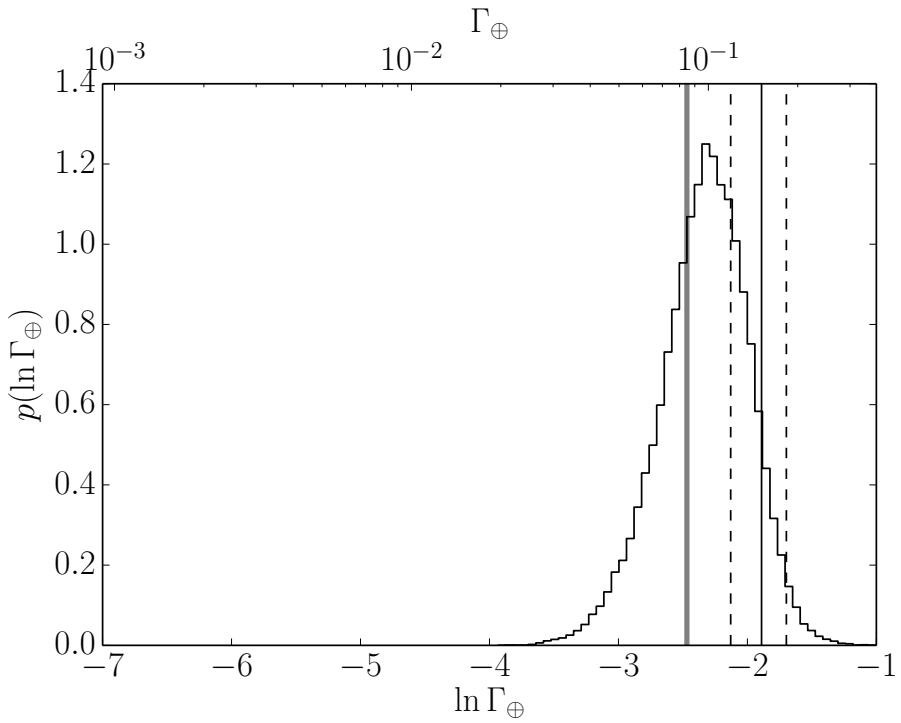

Figure 2.3: **Simulated data**. The extrapolated rate density of Earth analogs $\Gamma_\oplus$ as inferred by the different techniques applied to the *Catalog A* simulation. Applying the method used by Petigura et al. (2013a) gives a constraint indicated by the vertical black line with error bars shown 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.

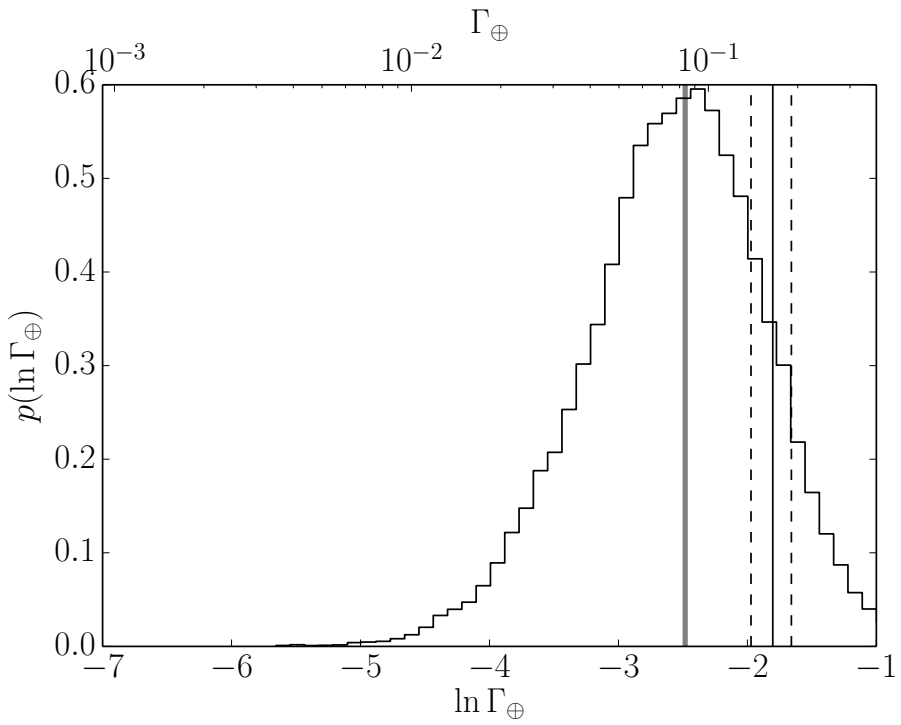

Figure 2.4: **Simulated data**. The same as Figure 2.3 for *Catalog B*.

2.9 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 2.6, we will use the catalog of small exoplanet candidates orbiting Sun-like stars published by Petigura et al. (2013a). 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. (2013a) 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[12].

Figure 2.5 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 2.6. 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 2.7 and 2.8. Figure 2.7 shows the log-radius occurrence rate density integrated over various logarithmic bins

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

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 2.8 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 2.8) is in tension—even though our result has large fractional uncertainty—with the result from Petigura et al. (2013a). This is shown in Figure 2.9 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

$$\Gamma_\oplus \quad = \quad 0.019^{+0.019}_{-0.010} \text{ nat}^{-2} \tag{2.22}$$

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. (2013a) measured $0.119^{+0.046}_{-0.035}$ nat^{-2} for the same quantity (indicated as the vertical lines in Figure 2.9). 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 ($200 \leq P/\text{day} < 400$ and $1 \leq R/R_\oplus < 2$) to find a *rate* of Earth analogs. The published rate is $0.057^{+0.022}_{-0.017}$ (Petigura et al. 2013a) and our posterior constraint is

$$\int_{P=200\,\text{day}}^{400\,\text{day}} \int_{R=1\,R_\oplus}^{2\,R_\oplus} \Gamma_{\boldsymbol{\theta}}(\ln P,\, \ln R)\, \mathrm{d}[\ln R]\, \mathrm{d}[\ln P] \quad = \quad 0.019^{+0.010}_{-0.008} \quad . \tag{2.23}$$

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

$$\mu = 5.44 \pm 1.56 \qquad \text{and} \qquad \ln \lambda_0 = 1.68 \pm 0.72 \quad . \tag{2.24}$$

The MCMC chains used to compute these values is available online[13].

2.10 Comparison with previous work

Our inferred rate density of Earth analogs (Equation 2.22) 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. 2013a). 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. (2013a) asserted a flat distribution in logarithmic period for small planets. Our results suggest that the data *do not support* this assumption (see Figure 2.6). 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. (2013a).

To test the significance of the choice of extrapolation function, we relax the assumption

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

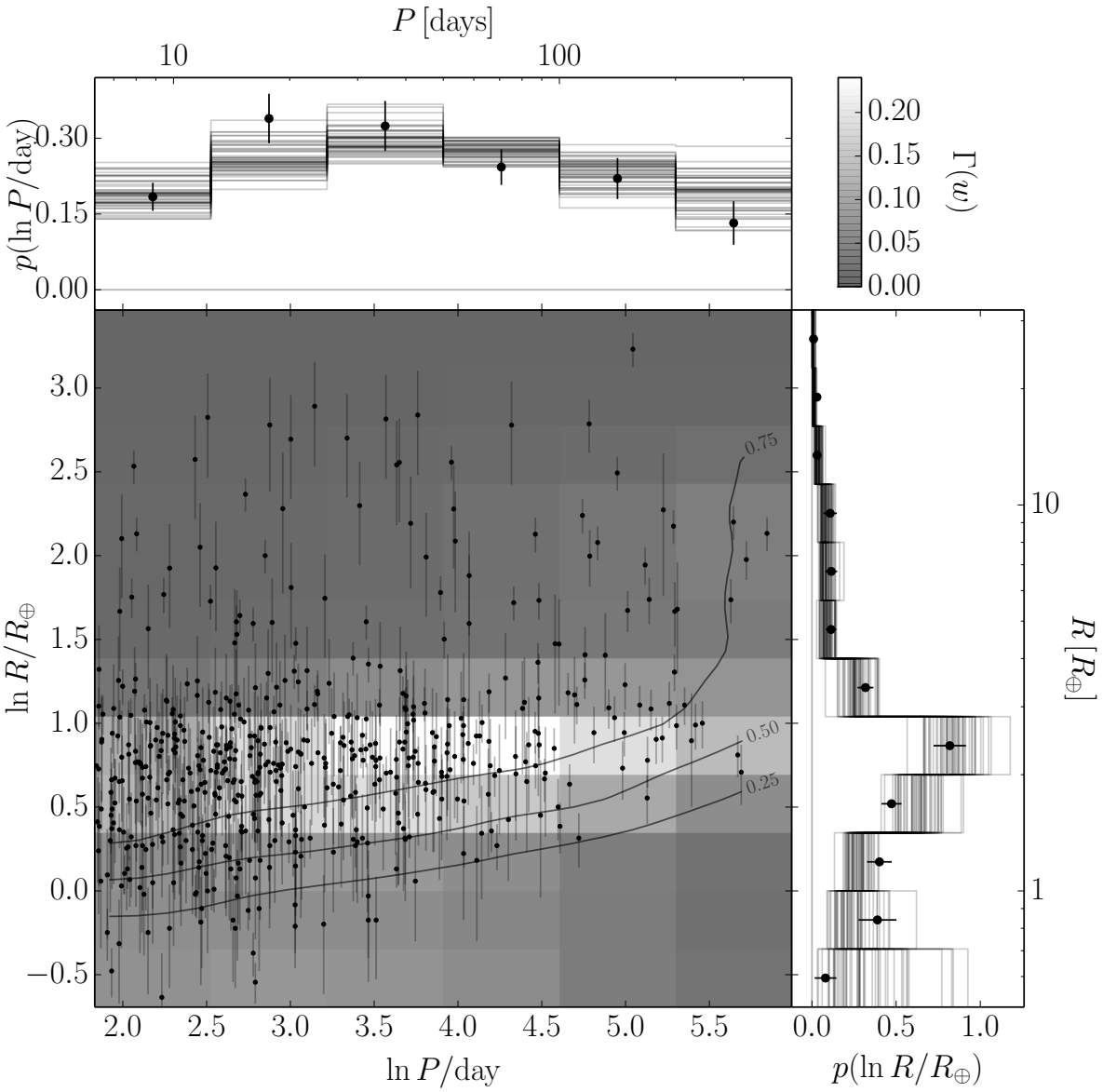

Figure 2.5: **Real data**. The same as Figure 2.1 when applied to the observed data from Petigura et al. (2013a). *Center:* the points with error bars show the catalog measurements, the contours show the survey completeness function, and the grayscale shows the median posterior occurrence surface. *Top 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.

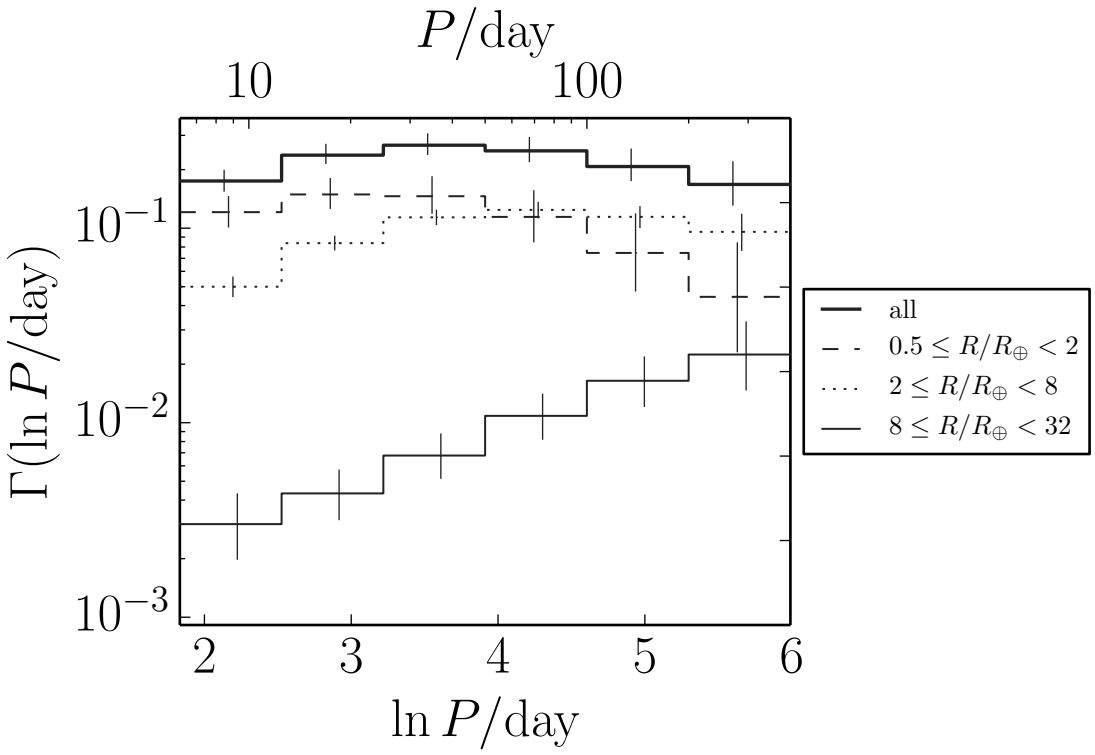

Figure 2.6: **Real data**. The occurrence rate density as a function of logarithmic period integrated over bins in logarithmic radius. The lines with error bars show the posterior sample median and 68th percentile and the line style specifies the radius bin. The period distribution for the largest planets 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.

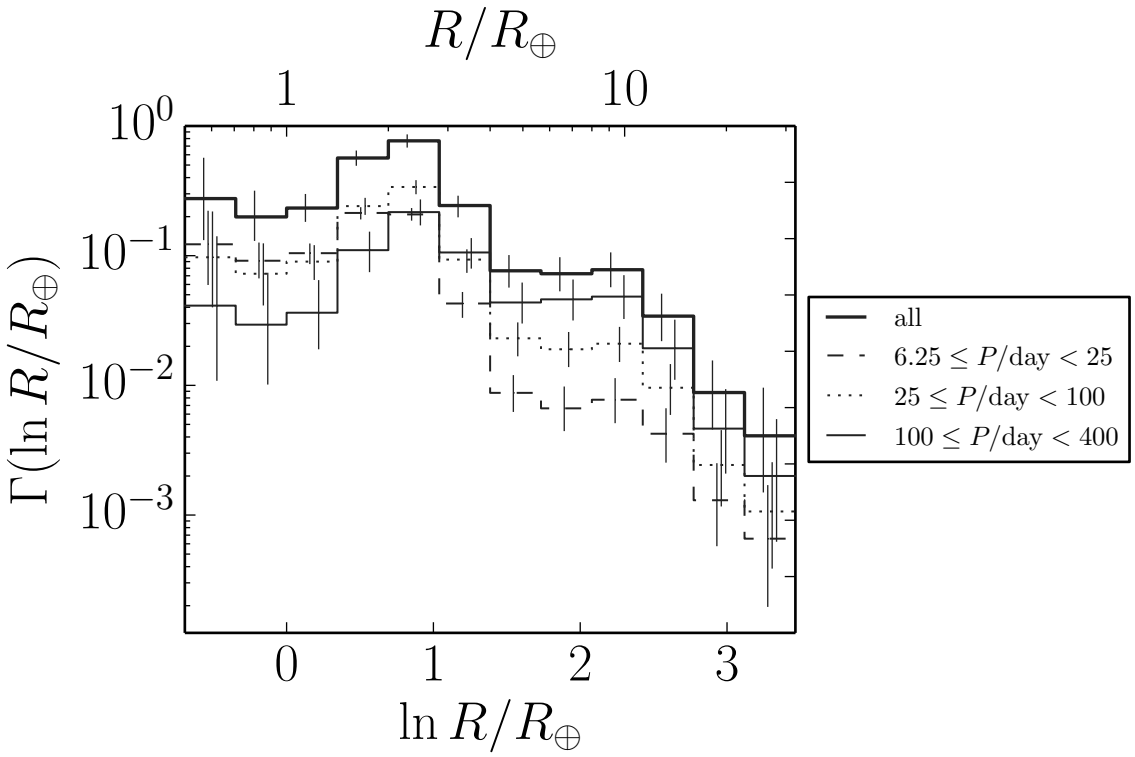

Figure 2.7: **Real data**. The occurrence rate density as a function of logarithmic radius integrated over bins in logarithmic period. The lines with error bars show the posterior sample median and 68th percentile and the line style specifies the period bin. The distributions in all the period bins are qualitatively consistent and there are plausibly features near $R \sim 3\,R_\oplus$ and $R \sim 10\,R_\oplus$.

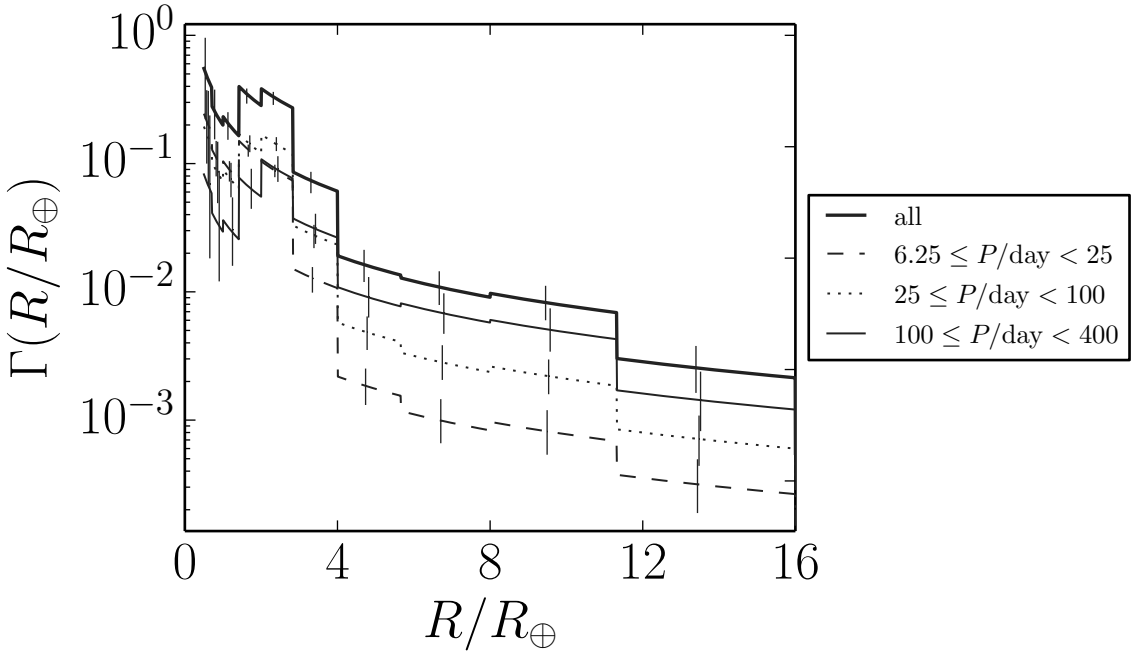

Figure 2.8: **Real data**. The same as Figure 2.7 but presented as a density in radius instead of logarithmic radius.

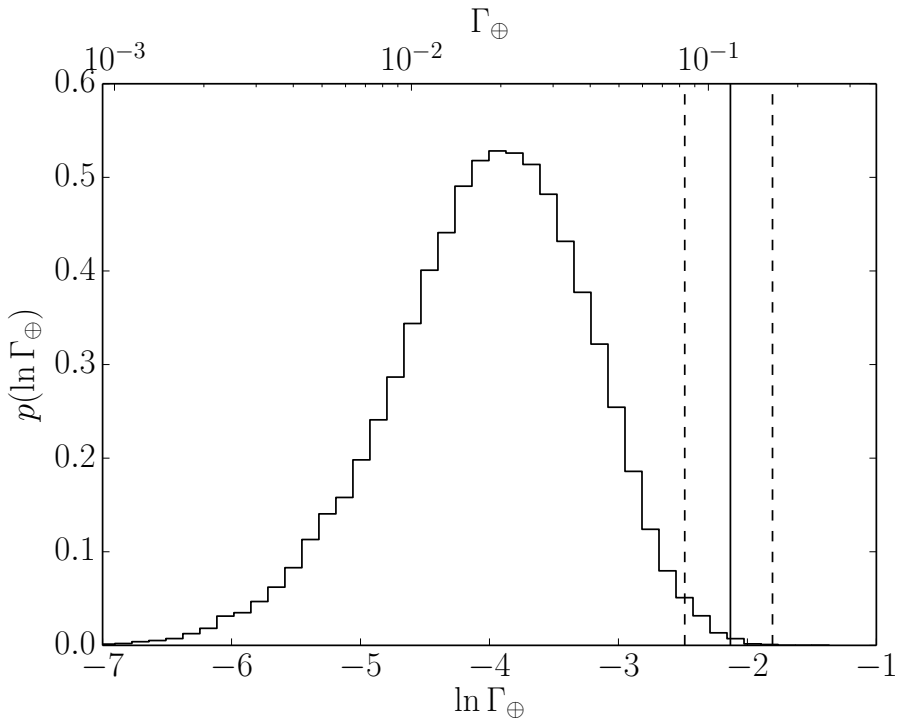

Figure 2.9: The extrapolated rate density of Earth analogs $\Gamma_\oplus$ (the same as Figure 2.3 but applied to the catalog from Petigura et al. 2013a). The histogram is the MCMC estimate of our posterior constraint on this rate density. The vertical black line with error bars shown as dashed lines is the result from Petigura et al. (2013a) converted to a rate density by dividing by their bin volume.

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 planets in this range can be calculated using Equation (2.2). 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. 2013). This results give a much more uncertain and substantially lower estimate for the rate of Earth analogs

$$\Gamma_\oplus \quad = \quad 0.072^{+0.088}_{-0.047} \quad . \tag{2.25}$$

With the large error bars, this result is consistent with both results (see Figure 2.10 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 2.10) gives the result

$$\Gamma_\oplus \quad = \quad 0.040^{+0.031}_{-0.019} \quad . \tag{2.26}$$

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 ($P = 10 - 250$ 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. (2013a), 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. (2013a) sample to be an underestimate of the *True* rate.

2.11 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 2.2 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. 2013a) 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.

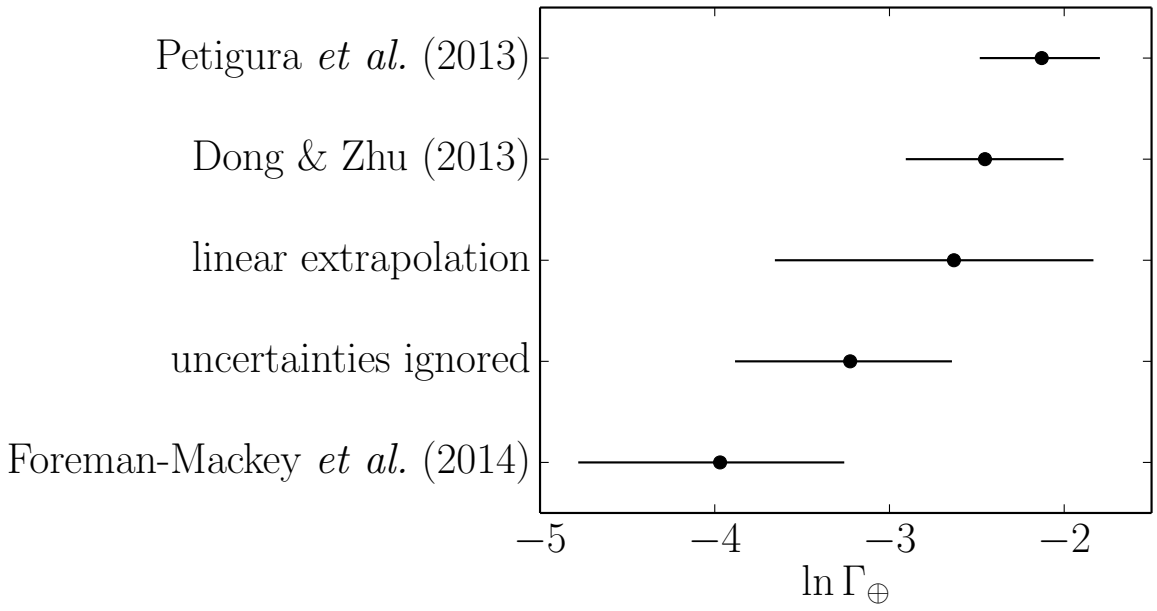

Figure 2.10: Comparison of various estimates of $\Gamma_\oplus$. From the top, the first value is the number published by Petigura et al. (2013a) and converted to consistent units. The second point shows the value implied by the power law model for the occurrence rate of $1 - 2\,R_\oplus$ planets from 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 ($50 - 400\,\mathrm{days}$) but allowing the period distribution to be *linear* instead of *uniform*. The "uncertainties ignored" value is given by applying the model developed in this Chapter but with the error bars on radius artificially set to zero. Finally, the bottom point is the result of our full analysis.

2013; Petigura et al. 2013a). In particular, we find (see Figure 2.7) 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 2.6 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. (2013a). 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 Chapter, 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 2012; 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. (2013a) 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. 2013a) 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. (2013a, $200 \le P/\mathrm{day} < 400$ and $1 \le R/R_\oplus < 2$), and integrate the product inferred rate density function and the geometric transit probability (Equation 2.18) 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. (2013a) is

$$N_{\oplus,\,\mathrm{transiting}} = 10.6^{+5.9}_{-4.5} \tag{2.27}$$

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 (2.27)!

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. 2013a) 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[14] 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 run to re-generate all of the figures and results in this Chapter.

2.12 Appendix: 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

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

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 2.4) then Equation (2.2) becomes

$$\ln p(\{\boldsymbol{w}_k\} \,|\, \boldsymbol{\theta}) \quad = \quad \sum_{k=1}^{K} \sum_{j=1}^{J} \mathbf{1}[\boldsymbol{w}_k \in \Delta_j] \, [\ln Q_\mathrm{c}(\boldsymbol{w}_k) + \theta_j] - \sum_{j=1}^{J} \exp(\theta_j) \int_{\Delta_j} Q_\mathrm{c}(\boldsymbol{w}) \, \mathrm{d}\boldsymbol{w} \tag{2.28}$$

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

$$\exp({\theta_j}^*) \quad = \quad \frac{K_j}{\int_{\Delta_j} Q_\mathrm{c}(\boldsymbol{w}) \, \mathrm{d}\boldsymbol{w}} \tag{2.29}$$

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

$$\frac{\delta \exp({\theta_j}^*)}{\exp({\theta_j}^*)} \quad = \quad \frac{1}{\sqrt{K_j}} \quad . \tag{2.30}$$

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 (2.29) 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 (2.29) 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 Chapter is necessary.

2.13 Chapter acknowledgements

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 research made use of the NASA *Astrophysics Data System*.

Chapter 3

A systematic search for transiting planets in the *K2* data

This Chapter is joint work with Benjamin T. Montet (Caltech, Harvard), David W. Hogg (NYU), Timothy D. Morton (Princeton), Dun Wang (NYU), and Bernhard Schölkopf (MPIS) submitted to *The Astrophysical Journal* as Foreman-Mackey et al. (2015).

3.1 Chapter abstract

Photometry of stars from the *K2* extension of NASA's *Kepler* mission is afflicted by systematic effects caused by small (few-pixel) drifts in the telescope pointing and other spacecraft issues. We present a method for searching *K2* light curves for evidence of exoplanets by simultaneously fitting for these systematics and the transit signals of interest. This method is more computationally expensive than standard search algorithms but we demonstrate that it can be efficiently implemented and used to discover transit signals. We apply this method to the full Campaign 1 dataset and report a list of 36 planet candidates transiting 31 stars,

along with an analysis of the pipeline performance and detection efficiency based on artificial signal injections and recoveries. For all planet candidates, we present posterior distributions on the properties of each system based strictly on the transit observables.

3.2 Introduction

The *Kepler* Mission was incredibly successful at finding transiting exoplanets in the light curves of stars. The Mission has demonstrated that it is possible to routinely measure signals in stellar light curves at the part-in-10^5 level. Results from the primary mission include the detection of planet transits with depths as small as 12 parts per million (Barclay et al., 2013).

The noise floor for *Kepler* data is often quoted as 15 parts per million (ppm) per six hours of observations (Gilliland et al., 2011). Although they generally do not interfere with searches for transiting planets, larger systematic effects exist on different timescales. One of the most serious of these is spacecraft pointing: If the detector flat-field is not known with very high accuracy, then tiny changes to the relative illumination of pixels caused by a star's motion in the focal plane will lead to changes in the measured or inferred brightness of the star.

The great stability of the original *Kepler* Mission came to an end with the failure of a critical reaction wheel. The *K2* Mission (Howell et al., 2014) is a follow-on to the primary Mission, observing about a dozen fields near the ecliptic plane, each for ~ 75 days at a time. Because of the degraded spacecraft orientation systems, the new *K2* data exhibit far greater pointing variations—and substantially more pointing-induced variations in photometry—than the original *Kepler* Mission data. This makes good data-analysis techniques even more valuable.

Good photometry relies on either a near-perfect flat-field and pointing model or else data-analysis techniques that are insensitive to these instrument properties. The flat-field for *Kepler* was measured on the ground before the launch of the spacecraft, but is not nearly as accurate as required to make pointing-insensitive photometric measurements at the relevant level of precision. In principle direct inference of the flat-field might be possible; however, because point sources are observed with relatively limited spacecraft motion, and only a few percent of the data are actually stored and downloaded to Earth, there isn't enough information in the data to derive or infer a complete or accurate flat-field map. Therefore, work on *K2* is sensibly focused on building data-analysis techniques that are pointing-insensitive.

Previous projects have developed methods to work with *K2* data. Both Vanderburg & Johnson (2014) and Armstrong et al. (2014) extract aperture photometry from the pixel data and decorrelate with image centroid position, producing light curves for each star that are "corrected" for the spacecraft motion. These data have produced the first confirmed planet found with *K2* (Vanderburg et al., 2014). Both Aigrain et al. (2015) and Crossfield et al. (2015) use a Gaussian Process model for the measured flux, with pointing measurements as the inputs, and then "de-trend" using the mean prediction from that model. Other data-driven approaches have been developed and applied to the data from space missions (for example, Ofir et al., 2010; Stumpe et al., 2012; Smith et al., 2012; Petigura et al., 2013a; Wang et al., 2015) and ground-based surveys (for example, Kovács et al., 2005; Tamuz et al., 2005; Berta et al., 2012) but they have yet to be generalized to *K2*.

In all of these light-curve processing methodologies, the authors follow a traditional procedure of "correcting" or "de-trending" the light curve to remove systematic and stellar variability as a step that happens *before* the search for transiting planets. Fit-and-subtract is dangerous: Small signals, such as planet transits, can be partially absorbed into the best-

fit stellar variability or systematics models, making each individual transit event appear shallower. In other words, the traditional methods are prone to over-fitting. Because over-fitting will in general reduce the amplitude of true exoplanet signals, small planets that ought to appear just above any specific signal-to-noise or depth threshold could be missed because of the de-trending. This becomes especially important as the amplitude of the noise increases.

The alternative to this approach is to *simultaneously fit* both the systematics and the transit signals. Simultaneous fitting can push the detection limits to lower signal-to-noise while robustly accounting for uncertainties about the systematic trends. In particular, it permits us to *marginalize* over choices in the noise model and propagate any uncertainties about the systematic effects to our confidence in the detection. This marginalization ensures that any conclusions we come to about the exoplanet properties are conservative, given the freedom of the systematics model.

In this Chapter we present a data-analysis technique for exoplanet search and characterization that is insensitive to spacecraft-induced trends in the light curves. We assume that the dominant trends in the observed light curves in each star are caused by the spacecraft and are, therefore, shared with other stars. We reduce the dimensionality by running PCA on stellar light curves to obtain the dominant modes. The search for planets proceeds by modeling the data as a linear combination of 150 of these basis vectors and a transit model. Our method builds on the ideas behind previous data-driven de-trending procedures such as the *Kepler* pipeline pre-search data conditioning (*PDC*; Stumpe et al., 2012; Smith et al., 2012), but (because of our simultaneous fitting approach) we can use a much more flexible systematics model while being less prone to over-fitting.

The methods developed within this paper are highly relevant to both *K2* and the upcoming *TESS* mission (Ricker et al., 2014). *TESS* will feature pointing precision of ~ 3 arc-

seconds[1], similar to the level of pointing drift with *K2*. Moreover, the typical star will be only observed for one month at a time, and the typical transit detection will be at a similar signal-to-noise ratio as with *K2*.

Catalogs of transiting planets found in the *K2* data will be important to better understand the physical properties, formation, and evolution of planetary systems. These planets, especially when they orbit bright or late-type stars, will be useful targets for ground-based and space-based follow-up, both for current facilities and those planned in the near future such as *JWST*. They will also deliver input data for next-generation population inferences (Foreman-Mackey et al., 2014), especially for the population of planets around cool stars (for example, Dressing & Charbonneau, 2015).

This project follows in the tradition of independently implemented transit search algorithms applied to publicly available datasets (such as Petigura et al., 2013b,a; Sanchis-Ojeda et al., 2014; Dressing & Charbonneau, 2015). These efforts have been hugely successful, especially in the field of exoplanet population inference because, thanks to their relative simplicity, the efficiency and behavior of these pipelines can be quantified empirically. The work described in this Chapter is built on many of the same principles as the previous projects developed for studying *Kepler* data but our main intellectual contribution is a computationally tractable framework for simultaneously fitting for the trends and the transit signal even when searching for planets.

The Chapter is organized as follows. In Section 3.3, we describe our method of extracting aperture photometry from the calibrated *K2* postage stamp time series. In Section 3.4 (with details in Appendix 3.9), we describe our data-driven model for the systematic trends in the photometric light curves and our method for fitting this model simultaneously with a transit signal. In Section 4.6, we give the detailed procedure that we use for discovering

[1] `http://tess.gsfc.nasa.gov/documents/TESS_FactSheet_Oct2014.pdf`

and vetting planet candidates. To quantify the performance and detection efficiency of our pipeline, we test (in Section 3.6) the recovery of synthetic transit signals, spanning a large range of physical parameters, injected into real *K2* light curves. Finally, in Section 3.7, we present a catalog of 36 planet candidates orbiting 31 stars from the publicly available *K2* Campaign 1 dataset.

3.3 Photometry and eigen light curves

The starting point for analysis is the raw pixel data. We download the full set of 21,703 target pixel files for *K2*'s Campaign 1 from MAST[2]. We extract photometry using fixed, approximately circular, binary apertures of varying sizes centered on the predicted location of the target star based on the world coordinate system. For each target, we use a set of apertures ranging in radius from 1 to 5 pixels (in steps of 0.5 pixels). Following Vanderburg & Johnson (2014), we choose the aperture size with the minimum CDPP (Christiansen et al., 2012) with a 6 hour window.[3]

All previous methods for analyzing *K2* data involve some sort of "correction" or "de-trending" step based on measurements of the pointing of the spacecraft (Vanderburg & Johnson, 2014; Aigrain et al., 2015; Crossfield et al., 2015). In our analysis, we do not do any further preprocessing of the light curves because, as we describe in the next Section, we fit raw photometric light curves with a model that includes both the trends and the transit signal.

One key realization that is also exploited by the official *Kepler* pipeline is that the systematic trends caused by pointing shifts and other instrumental effects are shared—with

[2] `https://archive.stsci.edu/k2/`

[3] Note that although we chose a specific aperture for each star, photometry for every aperture radius is available online: `http://bbq.dfm.io/ketu`.

different signs and weights—by all the stars on the focal plane. To capitalize on this, the *PDC* component of the *Kepler* pipeline removes any trends from the light curves that can be fit using a linear combination of a small number of "co-trending basis vectors". This basis of trends was found by running Principal Component Analysis (PCA) on a large set of (filtered) light curves and extracting the top few (~ 4) components (Stumpe et al., 2012; Smith et al., 2012). Similarly, we ran PCA on the full set of our own generated *K2* Campaign 1 light curves to determine a basis of representative trends but, unlike *PDC*, we retain and use a larger number of these components (150). For clarity, we will refer to our basis as a set of "eigen light curves" (ELCs) and the full set is made available online[4]. The top ten ELCs for Campaign 1 are shown in Figure 3.1.

3.4 Joint transit & variability model

The key insight in our transit search method that sets it apart from most standard procedures is that no de-trending is necessary. Instead, we can fit for the noise (or trends) and exoplanet signals simultaneously. This is theoretically appealing because it should be more sensitive to low signal-to-noise transits and similar methods have been shown to be effective for finding transits in ground-based surveys (Berta et al., 2012). The main motivation for this model is that the signal is never precisely orthogonal to the systematics and any de-trending will over-fit. This will, in turn, decrease the amplitude of the signal and distort its shape. In order to reduce these effects, most de-trending procedures use a very rigid model for the systematics. For *K2*, this rigidity has been implemented by effectively asserting that centroid measurements contain all of the information needed to describe the trends (Vanderburg & Johnson, 2014; Aigrain et al., 2015; Crossfield et al., 2015). In the *Kepler* pipeline, this

[4] `http://bbq.dfm.io/ketu`

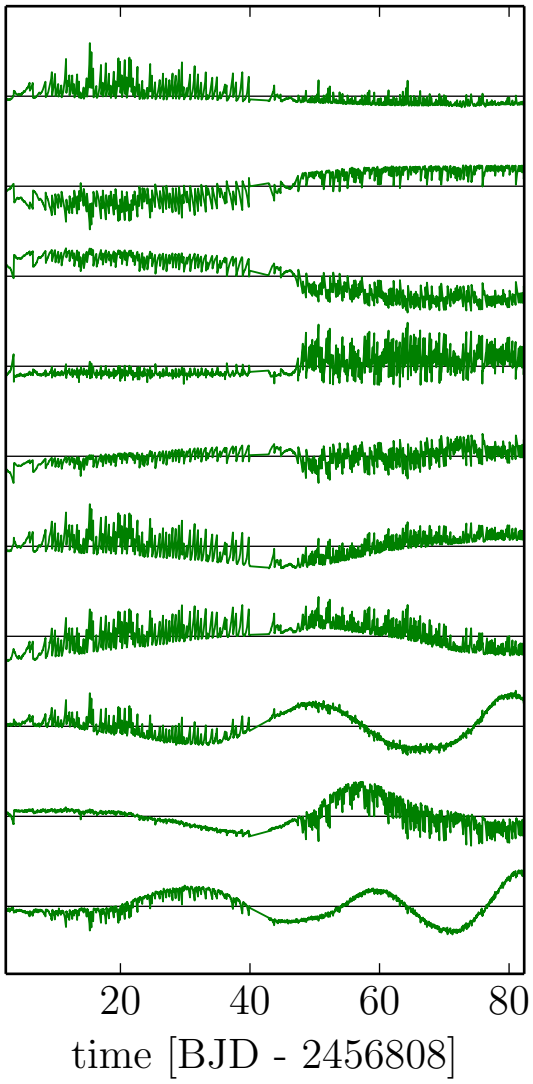

Figure 3.1: The top 10 eigen light curves (ELCs) generated by running principal component analysis on all the aperture photometry from Campaign 1.

is implemented by allowing only a small number of PCA components to contribute to the fit in the *PDC* procedure. Instead, we will use a large number of ELCs—a very flexible model—and use a simultaneous fitting and marginalization to avoid over-fitting.

Physically, the motivation for our model—and the *PDC* model—is that every star on the detector should be affected by the same set of systematic effects. These are caused by things like pointing jitter, temperature variations, and other sources of PSF modulation. Each of these effects will be imprinted in the light curves of many stars with varying amplitudes and signs as a result of the varying flat field and PSF. Therefore, while it is hard to write down a physical generative model for the systematics, building a data-driven model might be possible. This intuition is also exploited by other methods that model the systematics using only empirical centroids (Vanderburg & Johnson, 2014; Armstrong et al., 2014; Aigrain et al., 2015; Crossfield et al., 2015), but our more flexible model should capture a wider range of effects, including those related to PSF and temperature. For example, Figure 3.2 shows the application of our model—with 150 ELCs—to a light curve with no known transit signals and the precision is excellent.

If we were to apply this systematics model alone (without a simultaneous fit of the exoplanet transit model) to a light curve with transits, we would be at risk of over-fitting and decreasing the amplitude of the signal. Figure 3.3 demonstrates this effect on a synthetic transit injected into the light curve of a typical bright star. The middle two panels in this Figure show the light curve de-trended using 10 and 150 ELCs respectively. When only 10 ELCs are used, the measured transit depth is relatively robust but this model is clearly not sufficient for removing the majority of the systematic trends. The model with 150 ELCs does an excellent job of removing the systematics but it also distorts the transit shape and decreases the measured transit depth, hence reducing the signal strength in the *BLS* spectrum (Kovács et al., 2002).

In our pipeline we simultaneously fit for the transit signal and the trends using a rigid model for the signal and a relatively flexible model for the systematic noise. Specifically, we model the light curve as being generated by linear combination of 150 ELCs and a "box" transit model at a specific period, phase, and duration. The mathematical details are given in Appendix 3.9, but in summary, since the model is linear, we can analytically compute the likelihood function—conditioned on a specific period, phase, and duration—for the depth *marginalizing out the systematics model*. The signal-to-noise of this depth measurement can then be used as a quality of fit metric or candidate selection scalar. This computation is expensive but, as described in the following Sections, it is possible to scale the method to a *K2*-size dataset. The bottom panel of Figure 3.3 shows the application of this joint transit–systematics model to the synthetic transit discussed previously. When the joint model is used, the correct transit depth is measured—the transit is not distorted—but the systematics are also well-described by the model.

It is worth noting that this model can be equivalently thought of as a (computationally expensive) generalization of the "Box Least Squares" (*BLS*; Kovács et al., 2002) method to a more sophisticated description of the noise and systematics. Therefore, any existing search pipeline based on *BLS* could, in theory, use this model as a drop-in replacement, although some modifications might be required for computational tractability.

The choice to use 150 basis functions is largely arbitrary and we make no claims of optimality. This value was chosen as a trade-off between the computational cost of the search—the cost scales as the third power of the size of the basis—and the predictive power of the model. In some preliminary experiments, we found that using a larger basis did, as expected, lead to a marginally higher sensitivity to small transit signals but the gain wasn't sufficient to justify the added cost.

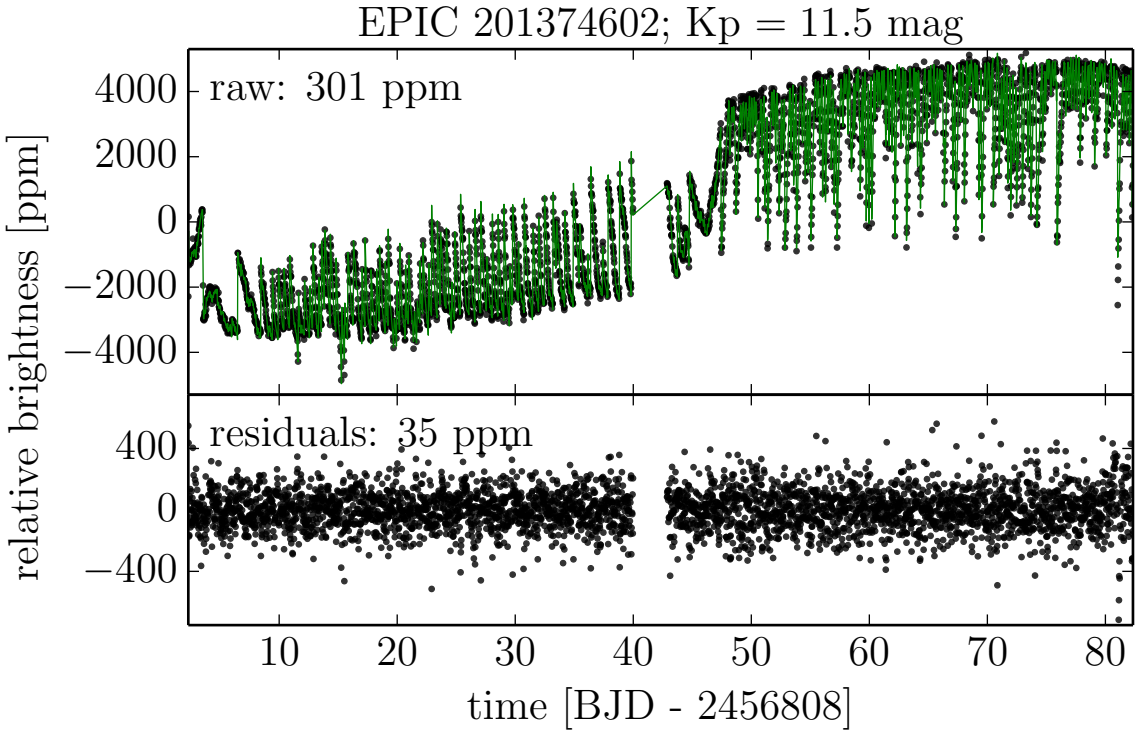

Figure 3.2: A demonstration of the ELC fit to the aperture photometry for EPIC 201374602. *Top:* The black points show the aperture photometry and the green line is the maximum likelihood linear combination of ELCs. The estimated 6-hour precision of the raw photometry is 264 ppm. *Bottom:* The points show the residuals of the data away from the ELC prediction. The 6-hour precision of this light curve is 31 ppm. Note that although we show a "de-trended" light curve to give a qualitative understanding of the model, this is not a product of the analysis. In this search for transits, *the data are never de-trended*.

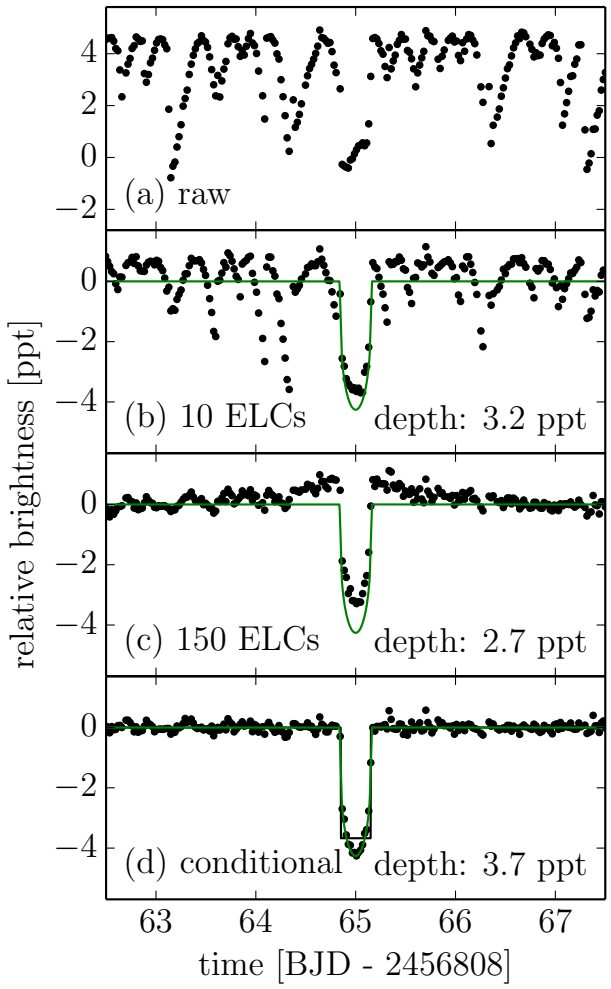

Figure 3.3: A comparison between de-trending using different numbers of ELCs and a simultaneous fit of the systematics and transit model. *(a)* The raw photometry for EPIC 201374602 with a synthetic transit injected at 65 days. In all plots, the photometry is measured in parts-per-thousand (ppt). *(b)* The black points show the photometry de-trended using a linear combination of 10 ELCs and the green line shows the true transit model. The maximum likelihood transit depth is computed following *BLS* (Kovács et al., 2002). While some of the systematics are removed by this model, there is still a lot of residual noise. *(c)* The same plot as panel *(b)* but using 150 ELCs to de-trend. This model removes the majority of the systematics but also distorts the transit and weakens the signal; it reduces the measured transit depth. *(d)* The final panel shows the results of simultaneously fitting for the transit and the systematics using 150 ELCs. The maximum likelihood depth is computed as described in Appendix 3.9. Like panel *(c)*, this model removes most of the systematics but does not distort the transit or reduce the measured transit depth.

3.5 Search pipeline

In principle, the search for transit signals simply requires evaluation of the model described above on a fine three-dimensional grid in period, phase, and duration, and then detection of high significance peaks in that space. In practice, this is computationally intractable for any grids of the required size and resolution. Instead, we can compute the values on this grid approximately, but at very high precision, using a two-step procedure that is much more efficient.

Specifically, we must evaluate the likelihood function for the light curve of star n given a period P, reference transit time T^0, duration D, and depth Z

$$p(\{f\}_n \,|\, P,\, T^0,\, D,\, Z) \quad . \tag{3.1}$$

We make the simplifying assumption that each transit enters this quantity independently. This is not true; as we change beliefs about each transit, we change beliefs about the systematics model, which in turn affects the other transits. However, this simplifying assumption is approximately satisfied for all but the shortest periods and leads to a huge computational advantage. Under this assumption, this likelihood function can be rewritten as

$$p(\{f\}_n \,|\, P,\, T^0,\, D,\, Z) \;=\; \prod_{m=1}^{M(P,T^0)} p(\{f\}_n \,|\, T_m(P,\, T^0),\, D,\, Z) \tag{3.2}$$

where $T_m(P,\, T^0)$ is the time of the m-th transit given the period P and reference time T^0, and $M(P,\, T^0)$ is the total number of transits in the dataset for the given P and T^0. Equation (3.2) can be efficiently computed for many periods and phases if we first compute

a set of likelihood functions for single transits on a grid in T_l and duration D_k

$$\{p(\{f\}_n \,|\, T_l,\, D_k,\, Z)\}_{l=1,\,k=1}^{L,\,K} \quad . \tag{3.3}$$

Then, we can use these results as a look-up table—with nearest-neighbor interpolation—to approximately evaluate the full likelihood in Equation (3.1).

In the remainder of this Section, we give more details about each step of the search procedure. In summary, it breaks into three main steps: linear search, periodic search, and vetting. In the **linear search** step, we evaluate the likelihood function in Equation (3.3) on a two-dimensional grid, coarse in transit duration D_k and fine in transit time T_m. Then in the **periodic search** step, we use this two-dimensional grid to approximately evaluate the likelihood (Equation 3.2) for a three-dimensional grid of periodic signals. Then, we run a peak detection algorithm on this grid that discards signals with substantially varying transit depths. These transit candidates are then passed along for machine and human **vetting**.

Linear search The linear search requires hypothesizing a set of transit signals on a two-dimensional grid in transit time and duration. For each point in the grid, we use the model described in Section 3.4 to evaluate *the likelihood function* for the transit depth at that time and duration. Since the model is linear, the likelihood function for the depth (marginalized over the model of the systematics) is a Gaussian with analytic amplitude L, mean $\bar{Z}$, and variance $\delta\bar{Z}^2$, all derived and given in Appendix 3.9. In the linear search, we save these three numbers on a two-dimensional grid of transit times T_l and durations D_k. The transit time grid spans the full length of Campaign 1 with half hour spacing and we choose to only test three durations: 1.2, 2.4, and 4.8 hours. Figure 3.4 shows the maximum likelihood transit depth $\bar{Z}$ as a function of transit time T for the light curve of EPIC 201613023, a transiting planet candidate with a period of 8.3 days.

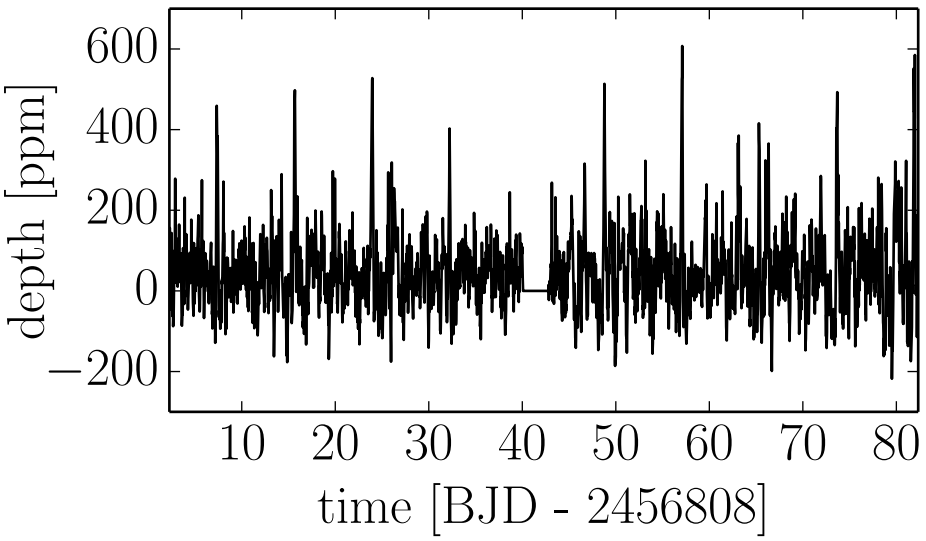

Figure 3.4: The maximum likelihood transit depth as a function of transit time as computed in the linear search of the light curve of EPIC 201613023. After the periodic search and vetting this target is found to have a planet candidate with a period of 8.3 days. The first transit occurs at 7.4 days on this plot.

Periodic search In the period search step, the table of likelihood functions generated in the linear search step are used to compute the likelihood of the periodic model (Equation 3.2) on a three dimensional grid in period P, reference time T^0, and duration D. At each point in this grid, the likelihood function for each transit depth is chosen as the nearest point calculated in the linear search. If the time spacing of the linear search is sufficiently fine, this will give a good approximation of the correct periodic likelihood. For each periodic model, we compute the likelihood of a model where the transit depth varies between transits and the "correct" simpler model where the transit depth is constant. The variable depth likelihood is given by the product of amplitudes from the initial search

$$p_{\rm var}(\{f\}_n \,|\, P,\, T^0,\, D) \;=\; \prod_{m=1}^{M(P,T^0)} L_m \quad . \tag{3.4}$$

Since the likelihood function for the depth at each transit time is known and Gaussian, the likelihood function for the depth under the periodic model can also be computed analytically; it is a product of Gaussians which itself is a Gaussian

$$p_{\rm const}(\{f\}_n \,|\, P,\, T^0,\, D) \;=\; \prod_{m=1}^{M(P,T^0)} \frac{L_m}{\sqrt{2\,\pi\,\delta\bar{Z}_m{}^2}} \exp\left(-\frac{[Z-\bar{Z}_m]^2}{2\,\delta\bar{Z}_m{}^2}\right) \tag{3.5}$$

where the maximum likelihood depth, for the periodic model, is

$$Z \;=\; \sigma_Z{}^2 \sum_{m=1}^{M(P,T^0)} \frac{\bar{Z}_m}{\delta\bar{Z}_m{}^2} \tag{3.6}$$

and the uncertainty is given by

$$\frac{1}{\sigma_Z{}^2} \;=\; \sum_{m=1}^{M(P,T^0)} \frac{1}{\delta\bar{Z}_m{}^2} \quad . \tag{3.7}$$

Note that this result has been *marginalized* over the parameters of the systematics model. Therefore, this estimate of the uncertainty on the depth takes any uncertainty that we have about the systematics into account.

In general, the variable depth model will *always* get a higher likelihood because it is more flexible. Therefore, a formal model comparison is required to compete these two models against each other on equal footing. For computational simplicity and speed, we use the Bayesian Information Criterion (BIC). The traditional definition of the BIC is

$$-\frac{1}{2}\,\mathrm{BIC} \quad = \quad \ln p(\{f\}_n \,|\, P,\, T^0,\, D) - \frac{K}{2}\ln N \tag{3.8}$$

where the likelihood function is evaluated at the maximum, K is an estimate of the model complexity and N is the effective sample size. To emphasize that K and N are tuning parameters of the method, we rewrite this equation as

$$-\frac{1}{2}\,\mathrm{BIC} \quad = \quad \ln p(\{f\}_n \,|\, P,\, T^0,\, D) - \frac{J\,\alpha}{2} \tag{3.9}$$

where J is the number of allowed depths—one for the constant depth model and the number of transits for the variable depth model—and α is chosen heuristically. For the K2 Campaign 1 dataset, we find that $\alpha \sim 1240$ leads to reliable recovery of injected signals while still being fairly insensitive to false signals.

To limit memory consumption, in the periodic search, we profile (or maximize) over T^0 and D subject to the constraint that $\mathrm{BIC}_\mathrm{const} < \mathrm{BIC}_\mathrm{var}$ and requiring that the signal have at least two observed transits. This yields a one-dimensional spectrum of the signal-to-noise of the depth measurement as a function of period using Equations (3.6) and (3.7) to compute Z/σ_Z at each period. The result is a generalization of the *BLS* frequency spectrum (Kovács et al., 2002) to a light curve model that includes both a transit and the trends. For example,

Figure 3.5 shows the spectrum for a planet candidate transiting EPIC 201613023.

After selecting the best candidate based on the signal-to-noise of the depth, we mask out the sections of the linear search corresponding to these transits and iterate the periodic search. This permits us to find second transiting planets in light curves in which we have already found a more prominent signal. Under our assumption of independent transits, this masking procedure is equivalent to removing the sections of data that have a transit caused by the exoplanet that produces the highest peak. For the purposes of this Chapter, we iterate the periodic search until we find three peaks for each light curve. This will necessarily miss the smallest and longest period planets in systems with more than three transiting planets but given the conservative vetting in the next Section, three peaks are sufficient to discover all the high signal-to-noise transits.

Initial candidate list The periodic search procedure returned three signals per target so this gave an initial list of 65,109 candidates. The vast majority of these signals are not induced by a transiting planet: there are many false positives. Therefore to reduce the search space, we estimate the signal-to-noise of each candidate by comparing the peak height to a robust estimate of variance in BIC values across period. This is not the same criterion used to select the initial three peaks but we find that it produces a more complete and pure sample. A cut in this quantity can reject most variable stars and low signal-to-noise candidates that can't be reliably recovered from the data. To minimize contamination from false alarms but maximize our sensitivity, we choose a threshold of 15. We also find that the signals with periods $\lesssim$ 4 days are strongly contaminated by false alarms. This might be because of the fact that our independence assumption (Equation 3.2) breaks down at these short periods. Therefore, we discard all signals with periods shorter than 4 days, acknowledging this will cause us to miss some planets (Sanchis-Ojeda et al., 2014). After these cuts, 741 candidates

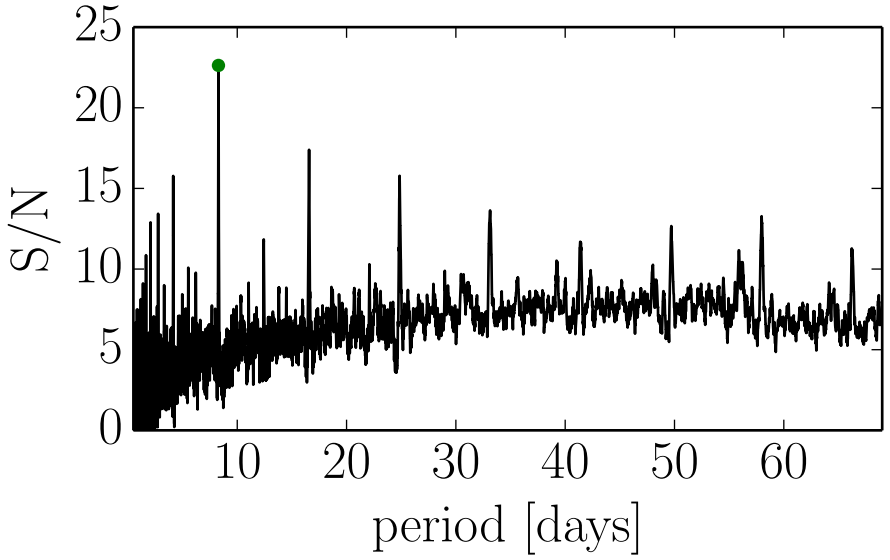

Figure 3.5: The signal-to-noise spectrum as a function of period for the light curve of EPIC 201613023. This is the generalization of the *BLS* spectrum (Kovács et al., 2002) to this simultaneous model of the transit and the systematic trends. To compute this spectrum, the results of the linear search (Figure 3.4) were used as described in Section 4.6. The top peak (at a period of 8.3 days) is indicated with a green dot. Iterating the periodic search found no other transit signals above the signal-to-noise threshold.

remain; we examine these signals by hand. The full list of peaks and their relevant meta data is available online at[5].

Hand vetting After our initial cuts on the candidate list, the majority of signals are still false alarms mostly due to variable stars or single outlying data points. It should be possible to construct a more robust machine vetting algorithm that discards these samples without missing real transits but for the purposes of this Chapter, we simply inspect the light curve for each of the 741 candidates *by hand* to discard signals that are not convincing transits. The results of this vetting can be seen online[6].

Although de-trended light curves are never used in the automated analysis of the data, when conditioned on a specific set of transit parameters, the model produces an estimate of what the light curve would look like in the absence of systematic effects. This prediction is one of the plots that we examine when vetting candidates by hand. For example, Figure 3.6 shows the maximum likelihood light curve for EPIC 201613023 evaluated at the candidate period, phase, duration, and depth. Similarly, Figure 3.7 shows the same prediction folded on the 8.3 day period of this candidate.

After visually inspecting 741 signals, 101 candidate transits pass and are selected as astrophysical events. Many of these signals are due to "false positives" such as eclipsing binary systems, either as the target star or as a background "blend." We address this effect in the following Section, where we separate the list of candidates into a list of astrophysical false positives and planet candidates.

Astrophysical false positives A major problem with any transit search is the potential confusion between transiting planets and stellar eclipsing binaries (EBs). Of particular

[5] `http://bbq.dfm.io/ketu`
[6] `http://bbq.dfm.io/ketu`

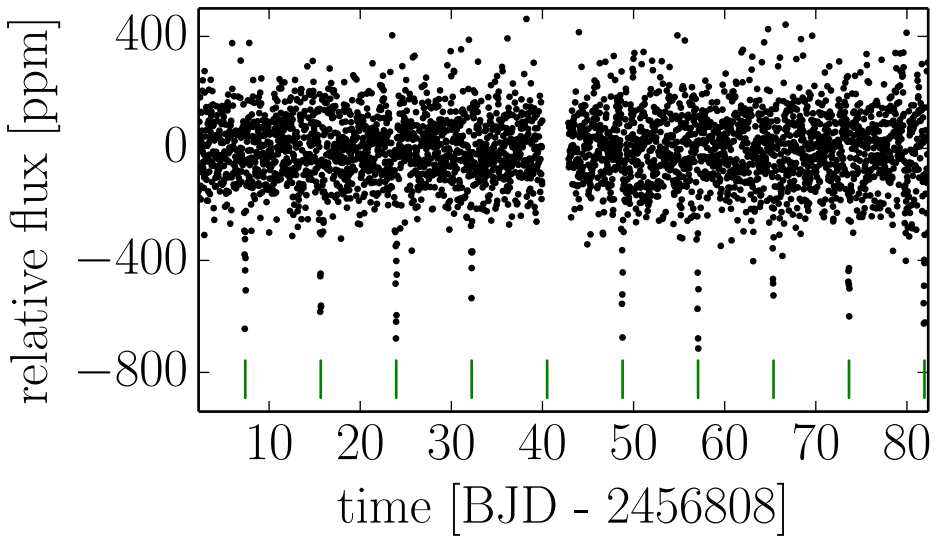

Figure 3.6: The maximum likelihood "de-trended" light curve for EPIC 201613023 evaluated at the planet candidate's period, phase, duration, and depth. The transit times are indicated by the green ticks below the light curve. This Figure is only generated for qualitative hand vetting and in the search procedure, the model is always marginalized over any choices about the systematic trends.

concern are grazing stellar eclipses or stellar eclipses that contribute only a small fraction of the total light in a photometric aperture, resulting in greatly diluted eclipse depths able to mimic the signals of small planets.

Ground-based transit surveys have experienced false-positive rates well over 50 percent. For example, Latham et al. (2009) reported eight eclipsing binaries and one transiting planet among the sample of transit candidates in one field of the Hungarian Automated Telescope Network transit search. In fact, the follow-up process to try to rule out such astrophysical false positives is a large portion of the effort that goes into a transit survey (e.g., O'Donovan et al., 2006; Almenara et al., 2009; Poleski et al., 2010).

Despite this large fraction of astrophysical false positives in ground-based surveys, the primary *Kepler* Mission saw a much lower false positive rate of only 5-10% (Morton & Johnson, 2011; Fressin et al., 2013), primarily due to three major factors. First, the superior precision of the *Kepler* photometry enables detection of secondary stellar eclipses, odd-even transit depth variations, and ellipsoidal variations (Batalha et al., 2010) to a much lower level than ground-based surveys. Second, the relatively small pixels and stable pointing of the *Kepler* telescope has enabled the identification of many spatially distinct blended eclipsing binaries by means of detailed pixel-level analysis (Bryson et al., 2013) to identify shifts in the center of light during transits. And finally, *Kepler* is sensitive to much smaller planets than ground-based surveys, and small planets are much more common than the Jupiter-sized planets able to be detected from the ground.

In *K2*, the precision of the photometric tests used to vet for such false positives is lower and they must be applied with care. There are typically only a handful of transits, meaning differences between "odd" and "even" transits must be large to create a significant difference. Searching for ellipsoidal variations is hindered by the short time baseline and the increased photometric uncertainty in *K2* data. Centroid variations are feasible in *K2* but

must be treated differently than in the original *Kepler* mission where this effect was generally measured using difference imaging (Batalha et al., 2010; Bryson et al., 2013).

To do first-pass vetting for blended EBs among our catalog of planetary candidates, we test for significant centroid offsets using the machinery that we have already established for modeling the systematic trends in the data, inspired by the methods used to vet *Kepler* candidates (Bryson et al., 2013). This is only an initial vetting step and a more complete characterization of our catalog's reliability is forthcoming (Montet, *et al.* in preparation).

To measure *centroid offsets*, we start by empirically measuring the pixel centroid time series for each candidate by modeling the pixels near the peak as a two-dimensional quadratic and finding the maximum at each time. This method has been shown to produce higher precision centroid measurements than center-of-light estimates (Vakili *et al.*, in preparation). Figure 3.8 shows the measured x and y pixel coordinate traces for EPIC 201613023. Much like the photometry, this signal is dominated by the rigid body motion of the spacecraft and we can, in fact, model it identically. In our analysis, we model the light curve as a linear combination of ELCs and a simple box transit model at a given period, phase, and duration (Equation 3.10). Under this model, the maximum likelihood depth can be computed analytically. If we apply *exactly the same model* to the centroid trace, the "depth" that we compute becomes the centroid motion in transit in units of pixels. Since the motions won't necessarily point in a consistent direction across transits, we treat each transit independently and report the average offset amplitude weighted by the precision of each measurement. To compute the significance of a centroid offset, we bootstrap the offset amplitude for models at the same period and duration but randomly oriented phases. If the centroid measured for the candidate transit is substantially larger than the random realizations, we label the candidate as a false positive. In practice, the precision of the centroid measurements isn't sufficient to robustly reject many candidates, but two candidates—EPIC 201202105 and

EPIC 201632708—have offsets 3-σ above the median out-of-transit offset amplitude so they are removed from the final catalog. For example, Figure 3.9 shows the in-transit centroid offset measured for EPIC 201202105 and compares it to the distribution of out-of-transit offset amplitudes.

A quick *a priori* estimate of the background blended eclipsing binary rate serves as a good sanity check. A query to the TRILEGAL (TRIdimensional modeL of thE GALaxy, Girardi et al., 2005) galaxy line-of-sight simulation software reveals that the typical density of field stars along the line of sight to the Campaign 1 field is about 7.8×10^{-4} arcsec^{-2}. This gives a probability of about 0.16 that a background star might be blended within a 8 arcsec radius (two pixels) from a target star. Allowing that $\sim$10% of stars might host close binary companions within the period range accessible by this survey, this gives a probability of 0.016 that a blended binary star might be chance-aligned within two pixels of any given target star. Noting that the average number of planets per star with periods less than 30 days is about 0.25 (Fressin et al., 2013), we can roughly estimate that we expect <10% of our candidates to be caused by nearby contaminating EBs. This estimate suggests that such astrophysical false positives should be rare in our sample, consistent with our detection of only 2 candidates with clear centroid offsets.

3.6 Performance

To test the performance and detection efficiency of our method, we conducted a suite of injection and recovery tests, five per star for all 21,703 target stars. For each test, we inject the signal from a realistic planetary system into the raw aperture photometry of a random target and run the resulting injected light curve through the full pipeline (except the manual vetting). If the search returns a planet candidate—passing all of the same cuts as we apply in

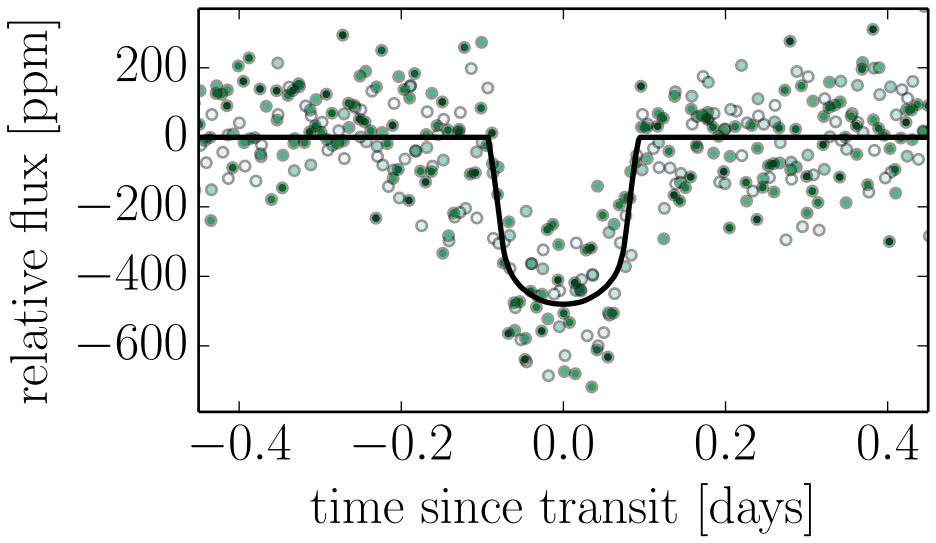

Figure 3.7: The maximum likelihood prediction for the light curve of EPIC 201613023 (see also Figure 3.6) folded on the 8.3 day period of this planet candidate. The points are color-coded by time and the median *a posteriori* transit model is overplotted as a black line.

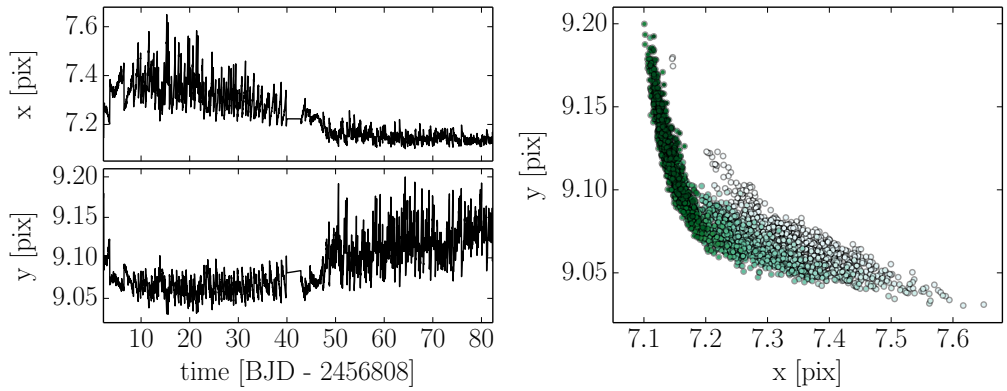

Figure 3.8: The centroid motion for EPIC 201613023. *Left:* The measured x and y pixel coordinates as a function of time. *Right:* The pixel coordinates color-coded by time. As identified by Vanderburg & Johnson (2014), the centroid motions fall in a slowly time variable locus. If the centroid coordinates in transit are inconsistent with the out-of-transit motions, the candidate is likely to be an astrophysical false positive.

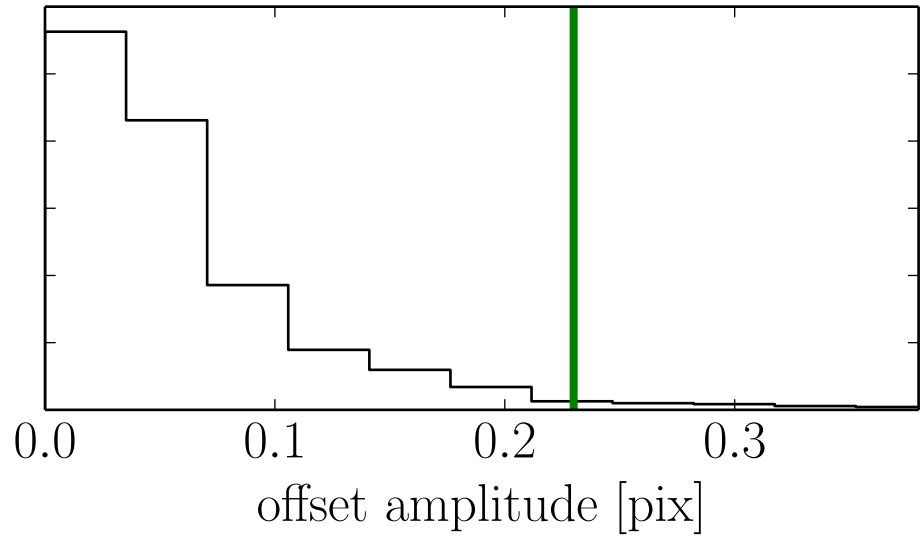

Figure 3.9: The estimated in-transit centroid offset for EPIC 201202105 (green line) compared to the distribution of 1000 centroid offests computed for randomly assigned phases (black histogram). The in-transit measurement is 3-σ larger than the median out-of-transit offset so it is rejected from the final catalog.

the main search (except the manual vetting)—with period and reference transit time within 6 hours of the injected signal, we count that injection as recovered. The detection efficiency of the search is given approximately by the fraction of recovered injections as a function of the relevant parameters.

To generate the synthetic signals, we use the following procedure:

1. Draw the number of transiting planets based on the observed multiplicity distribution of KOIs (Burke et al., 2014).

2. Sample—from the distributions listed in Table 3.1—limb darkening parameters and, for each planet, an orbital period, phase, radius ratio, impact parameter, eccentricity, and argument of periapsis.

3. Based on the chosen physical parameters, simulate the light curve, taking limb darkening and integration time into account (Mandel & Agol, 2002; Kipping, 2010), and multiply it into the raw aperture photometry.

We then process these light curves using exactly the pipeline that we use for the light curves without injections. Finally, we test for recovery after applying the cuts in signal-to-noise and period. We should, of course, also vet the results of the injection tests by hand to ensure that our measurements of detection efficiency aren't biased by the hand-vetting step but, since we chose to limit our sample to very high signal-to-noise candidates, it seems unlikely that our hand vetting removed any true transit signals. Any estimates of the false alarm rate will, however, be affected by this negligence but we leave a treatment of this for future work.

Figures 3.10 and 3.11 show the fraction of recovered signals as a function of the physical parameters of the injection and the magnitude of the star in the *Kepler* bandpass as reported

in the Ecliptic Plane Input Catalog (EPIC[7]). As expected, the shallower transits at longer periods are recovered less robustly and all signals become harder to detect for fainter stars. It is worth noting that these Figures are projections (or marginalizations) of a higher dimensional measurement of the recovery rate as a function of all of the input parameters. For example, this detection efficiency map is conditioned on our assumptions about the eccentricity distribution of planets and it is marginalized over the empirical distribution of stellar parameters. It is possible to relax this assumption and apply different distributions by re-weighting the simulations used to generate this figure. Therefore, alongside this Chapter, we publish the full list of injection simulations[8] to be used for population inference (occurrence rate measurements).

While we argue that the most relevant quantity to use to quantify the performance of a transit search pipeline is the efficiency with which it discovers transits, it is also useful to consider some other standard metrics. In particular, while de-trended light curves are never used at any stage of the analysis, our method does make a prediction for the systematics model and we can measure the relative precision of the residuals away from this model. These residuals are what would be used as de-trended light curves if that was the goal. Figure 3.12 shows, as a function of the *Kepler* magnitude reported in the EPIC, the 6-hour CDPP (Christiansen et al., 2012) for each light curve after subtracting the best fit linear combination of 150 ELCs.

[7] `http://archive.stsci.edu/k2/epic.pdf`
[8] `http://bbq.dfm.io/ketu`

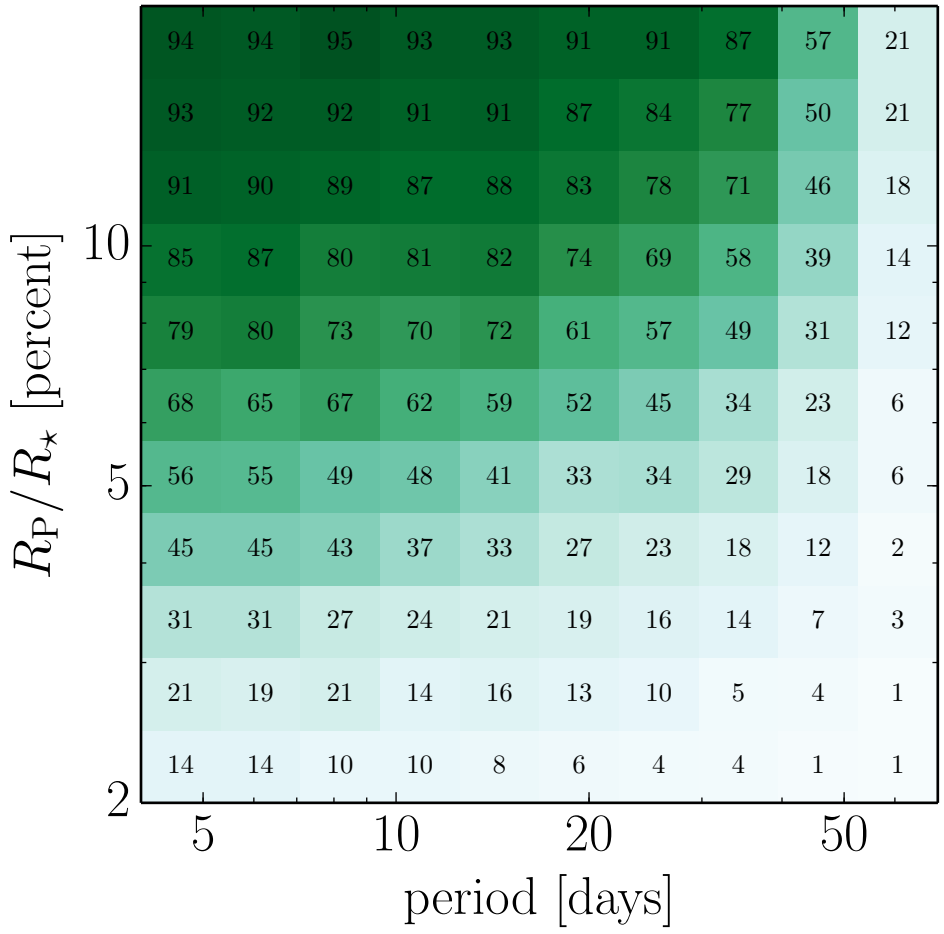

Figure 3.10: The detection efficiency of the search procedure as a function of the physical transit parameters computed empirically by injecting synthetic transit signals into the raw light curves and measuring the fraction that are successfully recovered. These tests were performed on the entire set of stars so these numbers are marginalized over all the stellar properties, including magnitude.

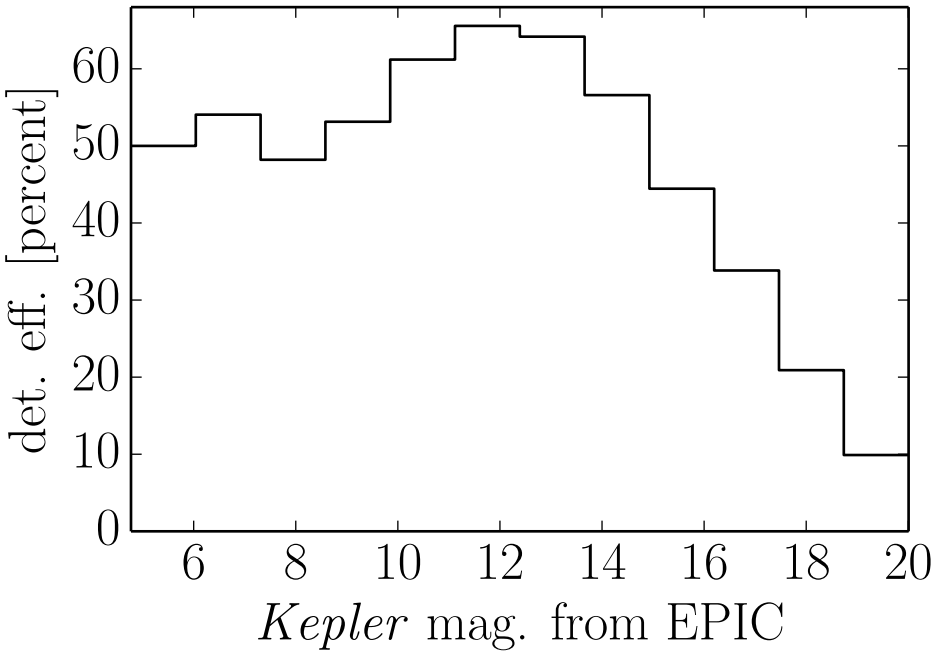

Figure 3.11: Like Figure 3.10, the empirically measured detection efficiency of the search procedure as a function of stellar magnitude as reported by the EPIC. This never reaches 90 percent because these numbers are marginalized over the range of physical parameters shown in Figure 3.10. Even for the brightest stars, the long period, small transits cannot be detected.

Parameter	Units	Distribution
limb darkening parameters q_1 and q_2	—	$q \sim U(0,\, 1)$
orbital period P	days	$\ln P \sim U(\ln 0.5,\, \ln 70)$
reference transit time T^0	days	$T^0 \sim U(0,\, P)$
radius ratio $R_P/R_\star$	—	$\ln R_P/R_\star \sim U(\ln 0.02,\, \ln 0.2)$
impact parameter b	—	$b \sim U(0,\, 1)$
eccentricity e	—	$e \sim \mathrm{Beta}(0.867,\, 3.03)$
argument of periapsis ω	—	$\omega \sim U(-\pi,\, \pi)$

Table 3.1: The distribution of physical parameters for the injected signals. The eccentricity distribution is based on Kipping (2013b) and the limb darkening parameterization is given by Kipping (2013a).

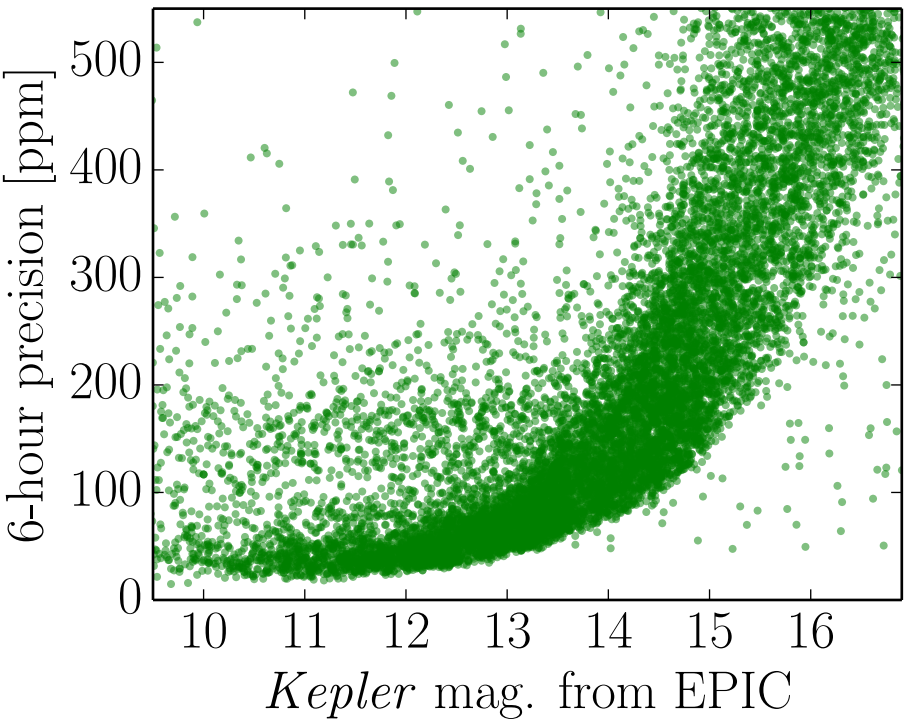

Figure 3.12: The 6-hour CDPP (Christiansen et al., 2012) for each light curve in Campaign 1 after subtracting the best fit linear combination of 150 ELCs. For each star, the precision is plotted as a function of the *Kepler* magnitude reported in the EPIC. The "outliers" in the bottom right corner of the plot are caused by a bright star within the photometric aperture and the points in the top left corner of the plot are variable stars where the major trends in the light curve are not caused by systematic effects, making the ELC model a bad fit.

3.7 Results

Out of the 21,703 Campaign 1 light curves, our pipeline returns 741 signals that pass the signal-to-noise and period cuts. After hand vetting by the two first authors, this list is reduced to 101 convincing astrophysical transit candidates. Of these, 36 signals—in 31 light curves—have no visible secondary eclipse and are deemed planet candidates. These planet candidates are listed in Table 3.2. The two candidates transiting EPIC 201367065 were previously published (Crossfield et al., 2015) and the third planet in that system is found as the third signal by our pipeline but it falls just below the signal-to-noise cut so it is left out of the catalog for consistency. This suggests that a less conservative cut in signal-to-noise and more aggressive machine vetting could yield a much more complete catalog at smaller radii and longer periods even with the existing dataset.

The remaining signals are caused by EBs with visible secondary eclipses. In most cases, the search reports the secondary eclipse as a candidate and in a few very high signal-to-noise cases, the period reported by the pipeline is incorrect and multiple candidates correspond to the same transit. It is important to note, however, that the choices made in the search were heuristically tuned to find planets, not binaries, so our results are not complete or exhaustive, especially at short orbital periods. There are other methods specifically tuned to find EBs in *K2* (such as Armstrong et al., 2014, 2015a) and these catalogs contain our full sample of EBs and more.

For the planet candidates, we perform a full physical transit fit to the light curve. To do this fit, we use Markov Chain Monte Carlo (MCMC; Foreman-Mackey et al., 2013) to sample from the posterior probability for the stellar and planetary parameters taking limb darkening and integration time into account. In this fit, we continue to model the trends in the data as a linear combination of the 150 ELCs but, at this point, we combine this with a realistic light curve model (Mandel & Agol, 2002; Kipping, 2013a). Even though we have no

constraints on the stellar parameters, we also sample over a large range in stellar mass and radius so that future measurements can be applied by re-weighting the published samples. In Table 3.2 we list the sample quantiles for the observable quantities and the full chains are available electronically[9]. Figure 4.2 shows the observed distribution of planet candidates in the catalog.

In a follow-up to this Chapter, we will characterize the stars for each of the candidates in detail but for now it's worth noting that many of the planet candidates are orbiting stars selected for *K2* as M-type stars. If this rate remains robust after stellar characterization and if these numbers are representative of the yields in upcoming *K2* Campaigns, the *K2* Mission will substantially increase the number of planets known to transit cool stars.

3.8 Discussion

We have searched the *K2* Campaign 1 data set for exoplanet transit signals. Our search is novel because it includes a very flexible systematics model, which is fit simultaneously with the exoplanet signals of interest (and marginalized out). By this method, we find 36 transiting exoplanets, which we have vetted by both automatically and manually and characterized by probabilistic modeling. The candidates are listed in Table 3.2 and posterior distributions of planet candidate properties are available[10].

The flexible systematics model we employ is a 150-parameter linear combination of PCA components derived from the full set of 21,703 stellar light curves. That is, it presumes that the systematics afflicting each star are shared in some way across other stars. It is our belief—although not a strict assumption of our model—that these systematics are caused

[9] `http://bbq.dfm.io/ketu`
[10] `http://bbq.dfm.io/ketu`

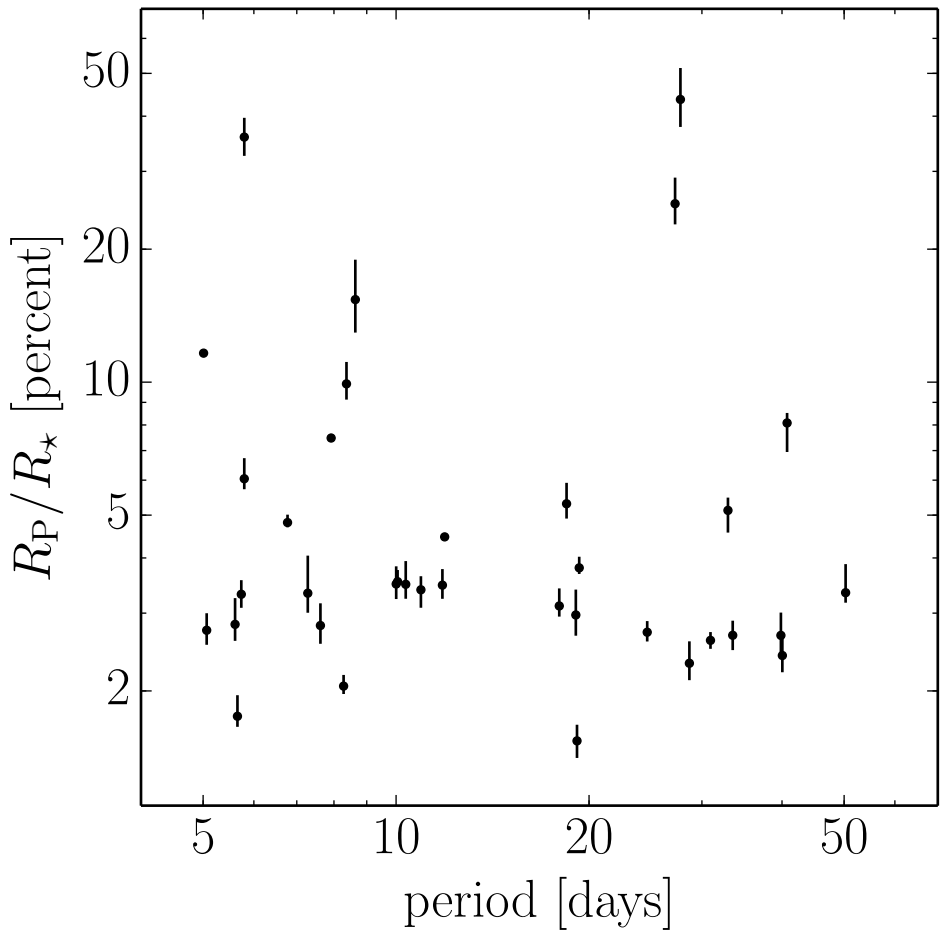

Figure 3.13: The *a posteriori* distribution of planet candidates in the catalog. The error bars indicate the 0.16 and 0.84 posterior sample quantiles for the radius ratios.

EPIC	RA (J2000)	Dec (J2000)	P [days]	t_0 [BJD-2456808]	$R_{\rm P}/R_\star$
201208431	174.745640	-3.905585	$10.0040^{+0.0018}_{-0.0016}$	$7.5216^{+0.0098}_{-0.0090}$	$0.0349^{+0.0034}_{-0.0026}$
201257461	178.161109	-3.094936	$50.2677^{+0.0083}_{-0.0074}$	$20.3735^{+0.0147}_{-0.0098}$	$0.0334^{+0.0054}_{-0.0017}$
201295312	174.011630	-2.520881	$5.6562^{+0.0007}_{-0.0007}$	$3.7228^{+0.0086}_{-0.0091}$	$0.0175^{+0.0020}_{-0.0009}$
201338508	169.303502	-1.877976	$10.9328^{+0.0022}_{-0.0021}$	$6.5967^{+0.0088}_{-0.0081}$	$0.0339^{+0.0025}_{-0.0030}$
201338508	169.303502	-1.877976	$5.7350^{+0.0006}_{-0.0006}$	$0.8626^{+0.0054}_{-0.0055}$	$0.0331^{+0.0025}_{-0.0023}$
201367065	172.334949	-1.454787	$10.0542^{+0.0004}_{-0.0004}$	$5.4186^{+0.0018}_{-0.0018}$	$0.0354^{+0.0022}_{-0.0011}$
201367065	172.334949	-1.454787	$24.6470^{+0.0014}_{-0.0016}$	$4.2769^{+0.0030}_{-0.0029}$	$0.0272^{+0.0016}_{-0.0013}$
201384232	178.192260	-1.198477	$30.9375^{+0.0029}_{-0.0052}$	$19.5035^{+0.0053}_{-0.0039}$	$0.0260^{+0.0011}_{-0.0011}$
201393098	167.093771	-1.065755	$28.6793^{+0.0105}_{-0.0116}$	$16.6212^{+0.0305}_{-0.0177}$	$0.0231^{+0.0028}_{-0.0020}$
201403446	174.266344	-0.907261	$19.1535^{+0.0050}_{-0.0050}$	$7.3437^{+0.0116}_{-0.0143}$	$0.0154^{+0.0014}_{-0.0013}$
201445392	169.793665	-0.284375	$10.3527^{+0.0011}_{-0.0011}$	$5.6110^{+0.0047}_{-0.0051}$	$0.0349^{+0.0045}_{-0.0025}$
201445392	169.793665	-0.284375	$5.0644^{+0.0006}_{-0.0006}$	$5.0690^{+0.0059}_{-0.0064}$	$0.0274^{+0.0025}_{-0.0020}$
201465501	176.264468	0.005301	$18.4488^{+0.0015}_{-0.0015}$	$14.6719^{+0.0035}_{-0.0032}$	$0.0531^{+0.0061}_{-0.0039}$
201505350	174.960319	0.603575	$11.9069^{+0.0005}_{-0.0004}$	$9.2764^{+0.0013}_{-0.0015}$	$0.0446^{+0.0009}_{-0.0006}$
201505350	174.960319	0.603575	$7.9193^{+0.0001}_{-0.0001}$	$5.3840^{+0.0006}_{-0.0008}$	$0.0747^{+0.0016}_{-0.0013}$
201546283	171.515165	1.230738	$6.7713^{+0.0001}_{-0.0001}$	$4.8453^{+0.0012}_{-0.0011}$	$0.0481^{+0.0020}_{-0.0012}$
201549860	170.103081	1.285956	$5.6083^{+0.0005}_{-0.0006}$	$4.1195^{+0.0045}_{-0.0047}$	$0.0283^{+0.0041}_{-0.0023}$
201555883	176.075940	1.375947	$5.7966^{+0.0002}_{-0.0002}$	$5.3173^{+0.0027}_{-0.0050}$	$0.0604^{+0.0068}_{-0.0032}$
201565013	176.992193	1.510249	$8.6381^{+0.0003}_{-0.0002}$	$3.4283^{+0.0016}_{-0.0015}$	$0.1538^{+0.0355}_{-0.0243}$
201569483	167.171299	1.577513	$5.7969^{+0.0000}_{-0.0000}$	$5.3130^{+0.0002}_{-0.0003}$	$0.3587^{+0.0379}_{-0.0334}$
201577035	172.121957	1.690636	$19.3062^{+0.0013}_{-0.0013}$	$11.5790^{+0.0025}_{-0.0027}$	$0.0380^{+0.0023}_{-0.0012}$
201596316	169.042002	1.986840	$39.8415^{+0.0136}_{-0.0155}$	$21.8572^{+0.0120}_{-0.0101}$	$0.0267^{+0.0034}_{-0.0022}$
201613023	173.192036	2.244884	$8.2818^{+0.0006}_{-0.0007}$	$7.3752^{+0.0055}_{-0.0052}$	$0.0205^{+0.0012}_{-0.0008}$
201617985	179.491659	2.321476	$7.2823^{+0.0007}_{-0.0008}$	$4.6337^{+0.0050}_{-0.0050}$	$0.0333^{+0.0072}_{-0.0032}$
201629650	170.155528	2.502696	$40.0492^{+0.0186}_{-0.0259}$	$4.5363^{+0.0202}_{-0.0172}$	$0.0241^{+0.0025}_{-0.0020}$
201635569	178.057026	2.594245	$8.3681^{+0.0002}_{-0.0002}$	$3.4514^{+0.0015}_{-0.0014}$	$0.0991^{+0.0120}_{-0.0078}$
201649426	177.234262	2.807619	$27.7704^{+0.0001}_{-0.0001}$	$13.3476^{+0.0001}_{-0.0002}$	$0.4365^{+0.0777}_{-0.0583}$
201702477	175.240794	3.681584	$40.7365^{+0.0026}_{-0.0025}$	$3.5451^{+0.0026}_{-0.0025}$	$0.0808^{+0.0043}_{-0.0114}$
201736247	178.110797	4.254747	$11.8106^{+0.0016}_{-0.0019}$	$3.8483^{+0.0093}_{-0.0071}$	$0.0347^{+0.0030}_{-0.0024}$
201754305	175.097258	4.557340	$19.0726^{+0.0048}_{-0.0049}$	$1.4893^{+0.0128}_{-0.0133}$	$0.0297^{+0.0042}_{-0.0030}$
201754305	175.097258	4.557340	$7.6202^{+0.0012}_{-0.0011}$	$3.6813^{+0.0061}_{-0.0057}$	$0.0281^{+0.0034}_{-0.0026}$
201779067	168.542699	4.988131	$27.2429^{+0.0001}_{-0.0001}$	$12.2599^{+0.0002}_{-0.0003}$	$0.2535^{+0.0369}_{-0.0259}$
201828749	175.654342	5.894323	$33.5093^{+0.0023}_{-0.0018}$	$5.1554^{+0.0037}_{-0.0032}$	$0.0267^{+0.0021}_{-0.0020}$
201855371	178.329775	6.412261	$17.9715^{+0.0015}_{-0.0017}$	$9.9412^{+0.0033}_{-0.0038}$	$0.0311^{+0.0030}_{-0.0017}$
201912552	172.560460	7.588391	$32.9410^{+0.0039}_{-0.0032}$	$28.1834^{+0.0057}_{-0.0105}$	$0.0513^{+0.0035}_{-0.0056}$
201929294	174.656969	7.959611	$5.0084^{+0.0001}_{-0.0001}$	$4.5703^{+0.0022}_{-0.0012}$	$0.1163^{+0.0011}_{-0.0014}$

Table 3.2: The catalog of planet candidates and their observable properties. These values and their uncertainties are derived from MCMC samplings and the numbers are computed as the 0.16, 0.5, and 0.84 posterior sample quantiles. The coordinates are retrieved directly from the EPIC.

primarily by pointing drifts, or movements of the pixels in the focal plane relative to the stars. In principle, if the systematics *are* dominated by pointing issues, the systematics model could require only three parameters—three Euler angles—not 150 amplitudes. However, because (as the pointing drifts) each star sees it's own unique local patch of flat-field variations, the mapping from pointing drifts to brightness variations can be extremely non-linear. Furthermore, because when the pointing is moving fast there is a smearing of the point-spread function, there are effects keyed to the time derivative of the Euler angles as well. The large number (150) of linear coefficients gives the linear model the freedom to model complex non-linear behavior; we are trading off parsimony in parameters with the enormous computational advantages of maintaining linearity (and therefore also convexity). The computational advantages of the linear model are three-fold: Convexity obviates searching in parameter space for alternative modes; linear least-squares optimization can be performed with simple linear algebra; given Gaussian uncertainties and uninformative priors, marginalizations over the linear parameters also reduces to pure linear algebra.

The goal of this Chapter was to get exoplanet candidates out of the *K2* pixel-level data, it was *not* to generate light curves. That is, both the search phase and the characterization phase of the method are approximations to computations of a likelihood function for the pixel data telemetered down from the satellite. We did not generate "corrected" or "pre-search conditioned" light-curves at any stage; we simultaneously fit systematics and the signals of interest to the raw data. For this reason, there is no sense in which this method ever really produces corrected light curves.

In this work, we are agnostic about fundamental properties of the host stars. The only assumptions we make are that the star targeted by the *K2* team is truly the planet host, and that there is no dilution by other stars in any aperture. As a result, these posterior distributions reflect the maximum possible uncertainty in parameters such as the planet

radius, which depend sensitively on properties of the host star. To use these distributions to characterize the properties of specific systems, one could re-weight our samples using a measurement of the inferred stellar properties.

This project does not live in isolation and this is certainly not the last time the *K2* data will be searched! There are other teams searching the *K2* light curves for transiting planets (A. Vanderburg, private comm.) and they are likely to find some planets that we did not and vice versa. We make many heuristic choices and short-cuts in this search. For example, the choice to work at 150 principal components was based on computational feasibility and qualitative tests on a handful of light curves instead of any real model selection or utility optimization.

Another major limitation is that, in principle, the systematics model is designed to describe spacecraft-induced trends, but not intrinsic stellar variability. In practice, the method can still find planets around variable stars but a more sophisticated model should be more robust in this case. One appealing option would be to model the systematics as a Gaussian Process where the input parameters are both time and the same 150 ELCs. Interestingly, while this model isn't linear, the search and marginalization can still be executed efficiently—using optimized linear algebra algorithms (Ambikasaran et al., 2014, Foreman-Mackey et al. in preparation)—inside the search loop.

Additionally, while we apply this systematics model simultaneously with a transiting planet model to search for planet candidates, this scheme is not restricted to planet searches. Any astrophysical event that could be observed in the *K2* data could be searched for in the same way. By modeling a set of ELCs with any arbitrary data model, events in the *K2* data that appear similar to that data model could be identified. Such a technique may be useful in searching for astrophysical events such as ellipsoidal variations induced by orbiting companions, stellar activity, microlensing events, especially in the upcoming Campaign 9, or

active galactic nuclei variability.

A substantial caveat to the reliability of all existing transiting exoplanet searches is that they all include human intervention. This makes quantifying the false alarm rate of these catalogs complicated. There has been some work on automated vetting algorithms using supervised classification algorithms (McCauliff et al., 2014; Jenkins et al., 2014) but these methods rely on hand classified examples for training and the performance is not yet competitive with human classification.

The catalog of planet candidates presented here includes only planets with periods longer than 4 days and at least two transits in the *K2* Campaign 1 footprint. This means that we are necessarily missing many planets with orbital periods outside this range. In particular, planets with a single transit in the dataset must be abundant. These candidates are the most relevant for the study of planetary system formation and for statistical inference of the distribution of habitable zone exoplanets. What's more, given the observing strategy for *TESS*, where each field will only be contiguously observed for one month at a time, methods for finding and characterizing planets with a single transit are vital and the new *K2* light curves are a perfect test bed.

As a supplement to this Chapter, we make all the results, data products, and MCMC chains available at `http://bbq.dfm.io/ketu`. The LaTeX source for this Chapter, complete with the full revision history, is available at `http://github.com/dfm/k2-paper` and the pipeline implementation is available at `http://github.com/dfm/ketu` under the MIT open-source software license. This code and a lot of computation time are all that is needed to reproduce the Figures in this Chapter.

3.9 Appendix: Mathematical model

We model the raw aperture photometry as a linear combination of 150 ELCs and a transit model. Formally, this can be written for the light curve of the k-th star as

$$\boldsymbol{f}_k \;=\; \boldsymbol{A}\,\boldsymbol{w}_k + \text{noise} \tag{3.10}$$

where

$$\boldsymbol{f}_k \;=\; \begin{pmatrix} f_{k,1} & f_{k,2} & \cdots & f_{k,N} \end{pmatrix}^{\mathrm{T}} \tag{3.11}$$

is the list aperture fluxes for star k observed at N times

$$\boldsymbol{t} \;=\; \begin{pmatrix} t_1 & t_2 & \cdots & t_N \end{pmatrix}^{\mathrm{T}} \quad . \tag{3.12}$$

In Equation (3.10), the design matrix is given by

$$\boldsymbol{A} \;=\; \begin{pmatrix} x_{1,1} & x_{2,1} & \cdots & x_{150,1} & 1 & m_{\boldsymbol{\theta}}(t_1) \\ x_{1,2} & x_{2,2} & \cdots & x_{150,2} & 1 & m_{\boldsymbol{\theta}}(t_2) \\ & & \vdots & & & \\ x_{1,N} & x_{2,N} & \cdots & x_{150,N} & 1 & m_{\boldsymbol{\theta}}(t_N) \end{pmatrix} \tag{3.13}$$

where the $x_{j,n}$ are the basis ELCs—with the index j running over components and the index n running over time—and $m_{\boldsymbol{\theta}}(t)$ is the transit model

$$m_{\boldsymbol{\theta}}(t) \;=\; \begin{cases} -1 & \text{if } t \text{ in transit} \\ 0 & \text{otherwise} \end{cases} \tag{3.14}$$

parameterized by a period, phase, and transit duration (these parameters are denoted by $\boldsymbol{\theta}$).

Assuming that the uncertainties on $\boldsymbol{f}_k$ are Gaussian and constant, the maximum likelihood solution for $\boldsymbol{w}$ is

$$\boldsymbol{w}_k{}^* \quad \leftarrow \quad \left(\boldsymbol{A}^{\mathrm{T}}\,\boldsymbol{A}\right)^{-1}\,\boldsymbol{A}^{\mathrm{T}}\,\boldsymbol{f}_k \tag{3.15}$$

and the marginalized likelihood function for the transit depth is a Gaussian with the mean given by the last element of $\boldsymbol{w}_k{}^*$ and the variance given by the lower-right element of the matrix

$$\boldsymbol{\delta w}_k{}^2 \quad \leftarrow \quad \sigma_k{}^2\,\left(\boldsymbol{A}^{\mathrm{T}}\,\boldsymbol{A}\right)^{-1} \tag{3.16}$$

where σ_k is the uncertainty on $\boldsymbol{f}_k$. The amplitude of this Gaussian is given by

$$\mathcal{L}_k \quad = \quad \frac{1}{(2\,\pi\,\sigma_k{}^2)^{N/2}}\,\exp\left(-\frac{1}{2\,\sigma_k{}^2}\,|\boldsymbol{f}_k - \boldsymbol{A}\,\boldsymbol{w}_k{}^*|^2\right) \tag{3.17}$$

evaluated at the maximum likelihood value $\boldsymbol{w}_k{}^*$.

3.10 Chapter acknowledgements

It is a pleasure to thank Eric Agol (UW), Ruth Angus (Oxford), Tom Barclay (Ames), Zach Berta-Thompson (MIT), Daniel Bramich (QEERI, Qatar), Géza Kovács (Konkoly Observatory), Laura Kreidberg (Chicago), Erik Petigura (Berkeley), Roberto Sanchis Ojeda (Berkeley), and Andrew Vanderburg (Harvard) for helpful contributions to the ideas and code presented here.

Chapter 4

Searching for long-period transiting planets in the *Kepler* light curves using supervised classification

This Chapter is joint work with David W. Hogg (NYU) and Bernhard Schölkopf (MPIS).

4.1 Chapter abstract

Many of the most dynamically important planets have orbits longer than the baseline of existing transit surveys (*Kepler*, *K2*). Future surveys (*TESS*, *PLATO*) are planned to have shorter continuous coverage such that even habitable zone planets around M-dwarfs will only present a single transit event. Searches for these long-period transiting planets are plagued by false signals—especially when pushed to low signal-to-noise—and statistical studies of their population are complicated by weak constraints on the physical parameters of the system and high rates of false positives. We develop and present a computationally

expensive but tractable method of searching for single transits using supervised classification methods from the machine learning literature. For each star, we train several Random Forest classifiers on simulated signals injected into different subsets of the data. These models are evaluated on the raw photometry to assign a transit probability at each time. As a proof of concept, we apply this method to 3500 light curves from the *Kepler* archival dataset and discover a previously unknown single transit candidate, KIC 10602068. Assuming a bound Keplerian orbit and using an informative prior on the eccentricity, we measure a radius of $2\,R_\mathrm{J}$ and derive a weak constraint on the orbital period.

4.2 Introduction

The transit method of exoplanet detection and characterization has been demonstrated as the most powerful method for building systematic catalogs of exoplanets. Despite the great success of the *Kepler* Mission, with thousands of planet discoveries (Burke et al., 2014; Rowe et al., 2015), current methods for exoplanet discovery are currently limited in the range of orbital periods that can be studied. Specifically, the standard transit search procedures only discover signals with at least three observed transits (for example Petigura et al., 2013a; Burke et al., 2014; Rowe et al., 2015). This is found to be necessary in order to greatly reduce the number of false signals incorrectly labeled as candidates by the automated pipelines. For *Kepler*, with a baseline of about four years, this sets an upper limit on the detectable periods of just over a year. In the Solar System, Jupiter—with a period of 12 years—dominates the planetary dynamics and, since it would only exhibit at most one transit in the *Kepler* data, it would be missed by most existing transit search procedures. It is possible to discover long-period planets like this using targeted radial velocity (RV) surveys (for example Butler et al., 2006; Knutson et al., 2014) but the cost of implementing a systematic RV search is

substantially higher than searching the existing and forthcoming photometric data for single transits.

There are two main technical barriers to a search for single transit events. The first is that the transit probability for long-period planets is very low; scaling as $\propto P^{-5/3}$ for orbital periods longer than the baseline of contiguous observations. Therefore, even if long-period planets are intrinsically common, they will still be underrepresented in a transiting sample. The second challenge is that there are substantial signals in the observed light curves caused by stochastic processes—both instrumental (pointing jitter, temperature variations, *etc.*) and astrophysical (stellar variability, *etc.*)—that can masquerade as transit signals. In practice, even using the most sophisticated systematics removal methods, these false signal far outnumber the true single transits.

Nearly every transit search algorithm is built on certain common principles and many of the same decisions are made from one study to the next. In particular, at the heart of most methods is a matched filtering step where the likelihood of an approximate transit model is computed on a grid in the physical parameters (*Kepler* Data Processing Handbook[1]; Petigura et al., 2013a; Huang et al., 2013; Dressing & Charbonneau, 2015; Foreman-Mackey et al., 2015). Using these methods—and substantial hand-curation—some long-period transiting candidates have been published (for example Batalha et al., 2013; Huang et al., 2013; Kipping et al., 2014). For more recent data releases, the community has settled on more conservative selection criteria where candidates are required to have multiple transits (for example Petigura et al., 2013a; Burke et al., 2014; Rowe et al., 2015). Even the *QATS* algorithm (Carter & Agol, 2013) for finding quasiperiodic transits builds on much the same infrastructure. This means that there has never been a systematic search for single transits in the full *Kepler* dataset.

[1] `https://archive.stsci.edu/kepler/manuals/KSCI-19081-001_Data_Processing_Handbook.pdf`

A qualitatively different approach to planet search is employed by the *Planet Hunters* (PH) project[2] (Fischer et al., 2012). PH is a "citizen science" project where visitors to the website look at sections of light curve and mark the locations of transits that can be identified visually. Visual inspection can be a useful search technique for large single transits because humans are able to robustly distinguish transit signals from the noise. This project has yielded some promising long-period candidates and confirmed planets (for example Wang et al., 2013). One shortcoming of the PH method is that it can be difficult to fully quantify the performance (completeness and reliability) of the search and PH does not evaluate their users using synthetic transit signals (as is now common practice in the transit search literature). This means that the PH sample of candidates cannot be used for robust population inference.

We propose a transit search method that shares some similarities with the PH model but, instead of human classifiers, we use a supervised classification model trained on large numbers of simulated transit signals injected into real *Kepler* light curves. This procedure is motivated because, by definition, most randomly selected sections of light curve contain no transits and there is an excellent physical model for the signals of interest. If we can train a classification model to robustly separate light curve sections with transits from sections without then we should be able to use this in-place of the humans to find single transit events. This method is novel as a means to transit search and it differs substantially from the standard methods.

Automated classification processes have been hugely successful in the machine learning literature. Many astronomical problems can't be naturally solved by the standard machine learning models because of their heterogeneous character and heteroscedastic uncertainties. The search for single transits, however, is an excellent use case when applied with care. The noise in the *Kepler* light curve of a given star is relatively consistent for the full baseline of

[2] `http://www.planethunters.org/`

observation and we have a very good physical generative model for the transit signal so we can simulate arbitrarily large sets of training examples.

In the training step, a Random Forest (RF) classifier is trained to distinguish—based only on the light curve itself—between sections of light curve that contain transits and sections of the same size that do not. This model is then used to find candidates in left-out sections of the light curve. In principle, this model can capture arbitrarily complicated non-linear structure in the decision space. If the training set is large and complete it will "learn" the shape of a physical transit. In practice, classification algorithms like this also require a complete set of "negative" training examples so it will turn out that further vetting using physical models is necessary.

In Section 4.3, we extrapolate a recent model of the distribution of exoplanets to the long periods considered in this Chapter. Using an approximate but realistic model for the detection efficiency of a search for single transits, we estimate that ~ 60 single transit events should be detectable in the archival *Kepler* dataset. In Section 4.4, we describe the data and outline the structure exploited by our transit search procedure. In Section 4.5, we briefly describe the RF classification framework and in Section 4.6 we outline the steps used to apply this model for transit search. This method requires the setting of many tuning parameters and the optimization of these choices is both computationally and intellectually challenging so in Section 4.7 we discuss some steps in this direction. In Section 4.8, we demonstrate the feasibility of this method, present a single-transit discovery, and indicate some interesting weaknesses of the method. In Section 4.9 we describe the parameter estimation that is possible given a single transit.

4.3 Estimated yield

Before the launch of the *Kepler* Mission, Yee & Gaudi (2008) predicted the yield of single transit events based on the planet occurrence rates estimated given a small catalog of radial velocity discoveries (Butler et al., 2006) and a fit to their occurrence rate (Cumming et al., 2008). Using these early results and the pre-launch specifications of the *Kepler* Mission, Yee & Gaudi (2008) predicted that *Kepler* would discover ~ 6 single transit events in the full dataset. Now that the Mission is complete and we have better estimates of the occurrence rate and distribution of planets (for example Dong & Zhu, 2013; Petigura et al., 2013a; Foreman-Mackey et al., 2014; Dressing & Charbonneau, 2015), we can update the predicted yield for single transit events in the *Kepler* light curves. To make this estimate, we extrapolate the distribution of large planets on relatively short orbits out to longer periods, taking detection efficiency and the survey and targeting properties into account.

We will base our estimate on an assumed model $Q_k(R_\mathrm{P},\, P)$ for the absolute probability of detecting a planet with radius R_P and period P orbiting the star k, and the occurrence rate distribution

$$\Gamma(R_\mathrm{P},\, P) \quad = \quad \frac{\mathrm{d}N}{\mathrm{d}\ln R_\mathrm{P}\,\mathrm{d}\ln P} \quad , \tag{4.1}$$

the expected number of planets per star, per logarithmic radius, per logarithmic period. Given these two quantities, the expected number of single transits in the *Kepler* data is given by

$$N \quad = \quad \sum_{k=1}^{K} N_k \tag{4.2}$$

where the sum is over the K stars in the sample and N_k is the expected number of observable

transits around star k

$$N_k \;=\; \int Q_k(R_{\rm P},\, P)\, \Gamma(R_{\rm P},\, P)\, {\rm d}\ln R_{\rm P}\, {\rm d}\ln P \quad . \tag{4.3}$$

The integral in Equation (4.3) is over the parameter range of interest.

For the purposes of this discussion, we assume an approximate simple detection efficiency model with three contributions: the geometric transit probability, the temporal transit probability, and signal-to-noise ratio threshold of the search technique. Assuming circular orbits, the geometric transit probability is given by (Winn, 2010)

$$Q_k^{\rm (geom)}(R_{\rm P},\, P) \;=\; \frac{R_k}{a} \tag{4.4}$$

$$\;=\; \left(\frac{4\,\pi^2}{G\,M_k}\right)^{1/3} R_k\, P^{-2/3} \tag{4.5}$$

where R_k and M_k are the radius and mass of the star k respectively. The eccentricity distribution of these long-period planets will affect this transit probability (Kipping, 2014) but this simple prescription should be sufficient for a rough estimate. For long-period orbits, the temporal transit probability will be given by

$$Q_k^{\rm (time)}(R_{\rm P},\, P) \;=\; \frac{T_k}{P} \tag{4.6}$$

where T_k is the total time that *Kepler* spent observing the star k. Finally, the detection threshold depends on the detailed sensitivity of the search procedure but we will approximate it as a simple step function in signal-to-noise ratio of the transit. This contribution will be

given approximately by

$$Q_k^{(\mathrm{detect})}(R_\mathrm{P},\, P) \;=\; \begin{cases} 1 & \text{if } (R_\mathrm{P}/R_k)^2 > f\,\sigma_k \\ 0 & \text{otherwise} \end{cases} \tag{4.7}$$

where σ_k is an estimate of the noise in light curve of star k and f is the detection threshold for the method.

Combining the detection efficiency components, the integral from Equation (4.3) becomes

$$N_k \;=\; \left(\frac{4\,\pi^2}{G\,M_k}\right)^{1/3} R_k\,T_k \int_{R_{\mathrm{P}\,\mathrm{min}}}^{R_{\mathrm{P}\,\mathrm{max}}} \frac{\mathrm{d}R_\mathrm{P}}{R_\mathrm{P}} \int_{P_\mathrm{min}}^{P_\mathrm{max}} P^{-8/3}\,\Gamma(R_\mathrm{P},\, P)\,\mathrm{d}P \tag{4.8}$$

where all the integration limits are set by the target parameter space. Because of the detection probability threshold, $R_{\mathrm{P}\,\mathrm{min}}$ can be no smaller than $R_k\,\sqrt{f\,\sigma_k}$. Dong & Zhu (2013) used the catalog of short period transiting planets found by *Kepler* to constrain a model for the occurrence rate of large planets of the form

$$\Gamma(R_\mathrm{P},\, P) \;=\; C(R_\mathrm{P})\,\left(\frac{P}{10\,\mathrm{d}}\right)^{\beta(R_\mathrm{P})} \tag{4.9}$$

in a set of radius R_P bins. Using this model, the integral in Equation (4.8) becomes

$$N_k \;=\; \left(\frac{4\,\pi^2}{G\,M_k}\right)^{1/3} R_k\,T_k \sum_{j=1}^{J} \frac{C_j}{(10\,\mathrm{d})^{\beta_j}}\,\ln\left(\frac{R_{\mathrm{P}\,\mathrm{max},j}}{R_{\mathrm{P}\,\mathrm{min},j}}\right)\,\left[\frac{{P_\mathrm{min}}^{\beta_j-5/3}}{5/3-\beta_j}\right] \tag{4.10}$$

where the sum is over the J radius bins studied by Dong & Zhu (2013). It's important to note that the model used by Dong & Zhu (2013) is given in base-10 logarithms so the units must be converted to natural logarithms as appropriate.

Assuming the stellar parameters provided by the NASA Exoplanet Archive[3] (Huber et al., 2014) and approximating the stellar noise using the 15-hour CDPP (Christiansen et al., 2012), Figure 4.1 shows the extrapolated number of transiting planets with $1500\,\mathrm{d} < P < 5000\,\mathrm{d}$ based on the Dong & Zhu (2013) power-law model. For a detection threshold $f \sim 10$, the expected number of single transit events in the *Kepler* light curves is ~ 60.

This results is much more optimistic than the pre-launch estimate from Yee & Gaudi (2008) for a number of reasons. The *Kepler* Mission has substantially increased our knowledge of the population of planets of all sizes at periods shorter than a year—allowing extrapolation to the longer periods of interest in this Chapter. The detection capabilities of *Kepler* and the properties of the specific stars targeted by the mission are now understood and we use the most up-to-date measurements in this Section. Another effect is that, the *Kepler* Mission ran for more than four years—longer than the fiducial Mission goal of 3 years—and $\sim 190,000$ stars were targeted for nearly the full baseline instead of the original 100,000.

4.4 Data preparation

The *Kepler* Mission measured photometric time series for about 190,000 stars at half-hour cadence for a baseline of over four years. We aim to search these light curves for single transits of long-period planets and single eclipses of binary stars. These data are made available on MAST[4] and, for each target, we downloaded the full set of long cadence light curve files provided by Data Release 24 (Thompson et al., 2015). From these files, we extracted the PDC time series and split them into "sections" with no more than ten contiguous missing or flagged data points. The PDC light curves have been corrected for the instrumental effects

[3] We downloaded the `q1_q16_stellar` table from `http://exoplanetarchive.ipac.caltech.edu/` on 2015-04-03.

[4] `https://archive.stsci.edu/kepler/`

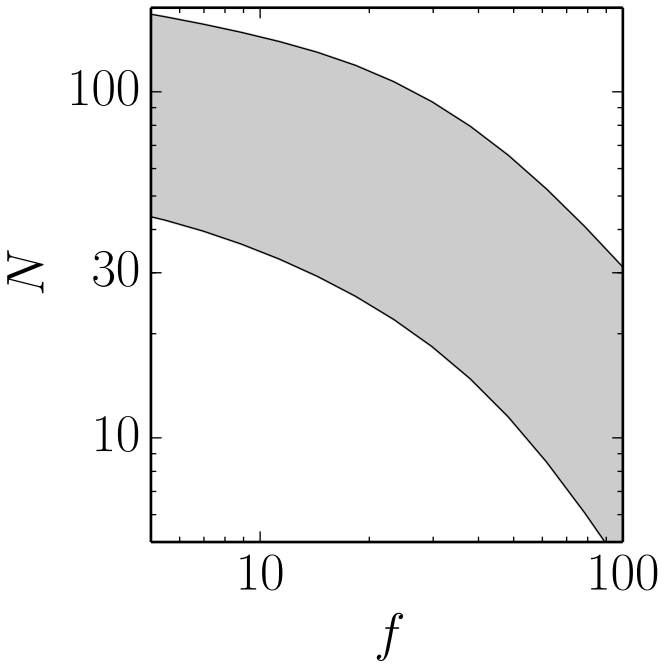

Figure 4.1: The expected number of single transit events—extrapolated from the Dong & Zhu (2013) power-law model fit to the shorter period *Kepler* candidates—as a function of the effective signal-to-noise threshold of the search procedure. The shaded region indicates the uncertainties propagated from the model parameters.

caused by the spacecraft using a data-driven model of the focal plane (Stumpe et al., 2012; Smith et al., 2012). Crucially, an attempt is also made by the PDC procedure to remove sharp instrumental artifacts like "sudden pixel sensitivity dropouts (SPSDs)". The success rate of this correction procedure is much higher than in earlier data releases but, as discussed in Section 4.8, there remain some cases that are not properly accounted for.

The goal of this project is to discover the transits of long-period planets that have not yet been discovered. Therefore, when studying the light curve of an eclipsing binary star or a star with known transiting planet candidates—on shorter periods—we also remove all the in-transit data for the candidate using the parameters provided by the *NASA Exoplanet Archive*[5].

4.5 Random forest classification

A common task in the machine learning literature is *supervised classification*, in which the goal is to separate objects into classes represented by sets of labeled examples. In the astronomy literature, supervised classification has been used for variable star classification (Richards et al., 2011) star–galaxy separation (Fadely et al., 2012), galaxy morphology prediction (Dieleman et al., 2015), and other applications. In each of these problems, there are a set of measurements that have been assigned classes—by some other method; often manually—and the goal is to transfer these labels to a set of observations that have not yet been labeled. For a more in-depth discussion of the application of these techniques in astronomy, the interested reader is directed to Ivezić et al. (2013).

In our problem of transit search, we want to "label" sections of light curve with either

[5] `http://exoplanetarchive.ipac.caltech.edu/`; We downloaded the `cumulative` table of *Kepler* Objects of Interest on 2015-03-25.

the `transit` or `no transit` class. This problem doesn't fall under the standard format of a supervised classification problem because very few long-period transits have actually been observed or classified. We do think, however, that we have a good physical generative model for the signal of interest and most of the observations of each star have no transits. Therefore, we train the model using simulated signals.

The Random Forest (RF) classification model (Breiman, 2001) is a popular model in the machine learning community. It is a common go-to model for basic classification tasks. It has also been applied with great success in astronomy (for example Richards et al., 2011, 2012; Jenkins et al., 2014). The RF model works by fitting an ensemble of decision trees to randomly selected subsamples of the training dataset. Each tree in the forest is "grown" by greedily choosing decision boundaries that optimize the separation of the training data in randomly selected subsets of the features until the separation is complete, each leaf contains only one class or, at most, a fixed number of samples. At each branch, the decision boundary is set by maximizing either the entropy or the "gini" coefficient.

We use the *scikit-learn* (Pedregosa et al., 2011) version of the RF classification algorithm[6] implemented in *Python*. This implementation has state-of-the-art speed and performance[7] while maintaining a flexible and user-friendly interface.

4.6 Search methodology

We apply the RF classification model to search for single transits in *Kepler* light curves. The basic structure of the problem, applied to the light curve of a single star, is:

1. **train** a classifier on simulated signals injected into a subset of the light curves sections,

[6] Specifically we use the `RandomForestClassifier` object; `http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`.

[7] `http://blog.explainmydata.com/2014/03/big-speedup-for-random-forest-learning.html`

2. **validate** the model on signals injected into a different subset of the data to estimate the precision of the results,

3. **test** the classifier by applying it to a final subset of the data—disjoint from the previous two sets—to predict the class (`transit` or `no transit`) of each light curve section, and

4. iterate until the entire light curve has been classified.

This means that searching a single light curve involves training, validating, and testing at least 3 RF models. In practice, we find that it is sufficient to split the data (intelligently) into 3 disjoint subsets so that we need only train 3 classifiers for each star.

A key assumption of this model is that the classes are sufficiently represented in feature space. In other words, the training set must span both of the classes. Given a large enough training set, this is not difficult for the `transit` class; we can simulate a transit for any combination of physical parameters. This is much more difficult for the `no transit` class. Since we don't have a generative model for the stellar and instrumental variability or other artifacts, we cannot simulate negative examples that cover the full space of possibilities. Instead we are using sections of real light curves as the `no transit` examples. This means that there are a finite number of samples available for training and validation, and if there is a qualitatively different signal in the test set that is not generated by a transit, the model is at risk of unpredictably misclassifying that section. This problem can be largely mitigated by careful splitting of the dataset—as described below—but there are inevitably some false signals that are not properly accounted for.

The following Sections detail the application of this search procedure from the training set simulations through to the candidate selection. These candidates are then vetted by hand although more robust methods should be possible as discussed in Section 4.6.4. Finally, Section 4.6.5 explores alternative representations of the data or features.

4.6.1 Splitting the dataset

It is crucial that the model used to test for transits in a raw *Kepler* light curve not be trained on the same data. We must split the light curve of each star into 3 disjoint sets. For the model to work well, we require that the signals induced by variability, noise, and other systematics be consistent between splits. It is, therefore, necessary that we assign the splits carefully. As discussed in Section 4.4, a light curve can be separated naturally into "sections" of nearly contiguous measurements; these are the base unit for splitting the dataset. For the typical *Kepler* target, this division is finer than month long chunks and there are about 100 sections for each light curve. Each light curve section is cut out of a specific "quarter" and each quarter was observed during a specific "season".

These divisions are relevant to this discussion because the spacecraft pointing across a quarter was extremely precise and the photometric aperture is fixed to the same pixels for the full quarter. Therefore, the instrumentally-induced systematics tend to be qualitatively consistent on quarter-long time scales. Every three months, the *Kepler* spacecraft rolled by 90 degrees causing the stars to end up on a different detector every quarter. This means that while the astrophysical variability should be shared between quarters at some level, the instrumental effects can be very different. A year later, however, the spacecraft returned to the original pointing. As a result, the data quality is also similar for observations made in the same season.

In order to take advantage of the natural structure of the observations, we try to evenly distribute observations from each quarter and season across the three sets. When this is not possible (with Quarter 0, for example), we randomly assign the section to a split.

4.6.2 Transit simulations

As mentioned previously, we train the RF classification models on a set simulated transit signals injected into real light curves and a set of light curve sections assumed to contain no transits. For each data split, we construct a set of 20,000 `transit` samples and an equal number of `no transit` examples. Each of these samples is generated by randomly selecting a section of 201 contiguous flux measurements from the split. Since we expect most of the systematic variability to be symmetric in time, we augment the training set by reversing the time series for half of the samples. This is analogous to image processing methods that exploit rotation and translational symmetries (for an example from astronomy see Dieleman et al., 2015).

To generate the `transit` examples, we simulate the transit signal induced by the physical orbit of a large planet. The specific distribution of simulation parameters is listed in Table 4.1. In these simulations, we take limb darkening and integration time into account[8] (Mandel & Agol, 2002; Kipping, 2010). For the purposes of this Chapter, we only simulate circular orbits under the assumption that a single transit of a planet on an elliptical orbit would be nearly identical to a circular orbit at a different period.

The dataset will still have some missing or flagged fluxes. In order to account for these missing points, we linearly interpolate these values based on their neighbors. For consistency, we perform the interpolation on the training set *after injecting the synthetic transit signals.* Before passing the light curve sections to the model at any stage, we normalize the section by its empirical median and take the logarithm of the fluxes. This normalization appears to yield better performance in some test cases but better performance might be achieved using optimized methods. This choice is discussed further in Section 4.9.

[8] `https://github.com/dfm/transit`

Parameter	Units	Distribution
limb darkening parameters q_1 and q_2	—	$q \sim U(0,\, 1)$
orbital period P	days	$\ln P \sim U(\ln 1500,\, \ln 5000)$
reference transit time δt	days	$\delta t \sim U(-0.15,\, 0.15)$
planet radius R_P	$R_\oplus$	$\ln R_\mathrm{P} \sim U(\ln 5,\, \ln 20)$
dimensionless impact parameter b	—	$b \sim U(0,\, 1)$

Table 4.1: The distribution of physical parameters for the injected signals. The limb darkening parameterization is given by Kipping (2013a).

4.6.3 Training, validation, and testing

For each subset of the data, we train a RF classifier on the 40,000 training examples generated from the light curves in that split. The `RandomForestClassifier` implementation from *scikit-learn* has several tuning parameters that could be optimized using cross-validation but we find acceptable performance in most cases using fixed values. Specifically, we use a forest with 1000 trees by setting `num_estimators = 1000`. We discuss this choice and the settings for other parameters in more detail in Section 4.9.

Using this classifier, fit to one of the light curve subsets, we validate its performance on each set of simulations from the two other splits. This results in two precision–recall curves for the classifier. The *precision* is an empirical measurement of the sample purity (or false positive rate) and *recall* is a measurement of the completeness (the probability of detecting a true signal). Since the RF classifier scores each class and with a continuous value S between 0 and 1, both precision and recall can be tuned by changing the threshold value of S above which the signal is considered a candidate. For each model–validation set pair, we ambitiously choose the threshold in order to obtain a precision of 100 percent. The result of this procedure applied to the three splits is three classifiers with two score thresholds each.

Each of these six models has been fit using two splits so we then apply it to the raw light curves (without injected signals) from the remaining, untouched splits. This yields a

`transit` class score as a function of transit time and any points above the detection threshold are passed along to the candidate selection procedure.

At this point, there are two class predictions from two different models for every light curve section. Any section that is classified as `transit` by *both models* is accepted as a single transit candidate.

4.6.4 Candidate vetting

While the search procedure described in the previous sections is quite robust to false alarms, single transit events are extremely rare (as discussed in Section 4.3) and because of substantial violations of the model assumptions, false signals continue to outnumber the true signals in this catalog of candidate transits. To mitigate this problem, we manually inspect the light curves of candidate signals. As future work, we expect to automate this procedure by fitting a physical transit model and comparing it to simple models of stellar variability and instrumental artifacts. For this Chapter, however, the catalog of candidates is sufficiently small that manual inspection is tractable. As demonstrated in Section 4.8, the most common misclassifications are caused by astrophysical variability but some remaining systematic effects are also incorrectly labeled as candidates.

4.6.5 Feature selection

In many supervised classification problems, a sophisticated feature extraction method is used to improve the performance of the procedure. For this specific problem, we find excellent performance using the 201 normalized flux values as the features. The depth of the transit and the noise in the light curve is not irrelevant to detection so we don't normalize the features to unit variance in the standard way. Instead, we normalize each sample (feature vector) by it's empirical median value and take the logarithm. Another option for a feature set

could be the Fourier transform or wavelet transform of the light curve section. Alternatively, the spectrogram of the fluxes in a larger window. We have experimented briefly with these options but find similar performance on the cases that we tested.

4.7 Tuning parameters

Many choices need to be made in building this search procedure. Each of these choices can be specified by a set of parameters that can be tuned to optimize the method. In order to tune these parameters, we must decide on a quantitative measurement of the "performance". The goal is, of course, to *robustly discover transiting planets*. In other words, the aim is a procedure that maximizes both *recall* and *precision*.

As discussed previously, recall is the probability, integrated across the full parameter space of interest that a true transit signal will be detected by the method. In astronomy, this quantity is generally called the "completeness" or "detection efficiency" and it is routinely measured for transit surveys (Petigura et al., 2013a; Dressing & Charbonneau, 2015; Foreman-Mackey et al., 2015). On the other hand, precision is the probability that a positive classification will be true. In astronomy, precision is generally hard to measure because we rarely have access to the ground truth. For this problem, correctly computing the recall is impossible because we never know that *there is definitely no transit* at a given time in the light curve.

What's more, it is not in general possible to maximize both recall and precision because transit signals are not completely separable from false positives. This means that, as the recall of the method improves, the false positive rate will also increase, reducing the precision. We must, therefore, choose a trade-off between recall and precision that represents our objective. The informal tradition in the transit search literature is to build a procedure that

maximizes the recall at an extremely high ($\sim$ 100 percent) fixed precision.

The main tuning parameters of the method are as follows:

- *the number of training and validation examples for each data split* — The performance of most classification models improves drastically as the size of the training set increases but in this case we are limited by the number of `no transit` samples that are available since we have no generative model for the variability and systematics. For this Chapter, we choose 20,000 positive `transit` samples per split and an equal number of negative examples.

- *the range of physical parameters used to simulate the transit signals* — The search will be most sensitive to transits with parameters spanned by the training set but training on too many low signal-to-noise examples in noisy light curves leads to decreased precision. The parameter ranges used for the demonstrations here are given in Table 4.1.

- *tuning parameters of the RF implementation* — The `RandomForestClassifier` implementation has a few tuning parameters. The most important of these seems to be `num_estimators`, the number of decision trees used in the forest. Increasing `num_estimators` leads to substantial computational cost with diminishing returns on the search performance. Otherwise, there are a few parameters (for example, `max_features`, `min_samples_split`, and `min_samples_leaf`) that can be used to regularize the model but we have found that the performance is insensitive to these choices. As a trade-off between performance and computational cost, we set `num_estimators = 1000` in all the examples shown in this Chapter.

In detail, the optimal set of decisions and parameter settings will be different for every target—and probably even for every permutation of the data splits. In practice, we find

that the performance of the search is fairly insensitive to most of these choices so we choose parameters that seem to work well in most cases and leave optimization for future work.

Estimating the precision We estimate the precision of each RF classification model on the validation set as described in Section 4.6.3. As discussed in that Section, we choose a target precision to 100 percent and set the score threshold accordingly. This estimate is only applicable to the test set under two major assumptions. The first is that all transiting planets with more than one transit and a large signal-to-noise ratio have been previously discovered by one of the many transit searches applied the *Kepler* dataset (Burke et al., 2014; Rowe et al., 2015). This is a safe assumption because, for large planets, these surveys are all largely complete out to periods yielding only 2 or 3 transits in the *Kepler* baseline (for example Petigura et al., 2013a). The second assumption is not always valid so it makes this estimate of the precision only approximate and not completely reliable. This assumption is that the noise processes in the data are stationary. In other words, the transit-like noise signals in the training and validation sets must be "similar" to the signals in the test section. This constraint is nearly satisfied when the PDC light curves are used (Section 4.4) and the data are split carefully (Section 4.6.1). As demonstrated by the mis-classifications shown in Section 4.8, however, it is clear that violations to this rule do exist in the data and final vetting of the candidates must be applied using a physical understanding of the signals.

4.8 Preliminary results

As a demonstration of the feasibility of this method to search for single transit events, we selected 3500 bright, Sun-like stars based on the revised stellar parameters collected and measured for a recent *Kepler* data release (Huber et al., 2014). The targets were selected in the parameter range

- effective temperature: $4100\,\mathrm{K} < T_{\mathrm{eff}} < 6100\,\mathrm{K}$,
- surface gravity: $4.0 < \log g < 4.9$, and
- *Kepler* magnitude: $15 > K_{\mathrm{p}} > 10$.

This search results in 596 candidate transits in the light curves of these 3500 stars. The vast majority of these candidates are misclassified false positives and many of them are caused by astrophysical sources. To weed out known astrophysical sources, we cross-match this list against known eclipsing binary stars (Matijevič et al., 2012) and stars with substantial variability (McQuillan et al., 2014). Removing these signals results in 273 candidates. Of these candidates only one (KIC 10602068) is a convincing transit and the others all appear to be caused by systematic effects, instrumental or astrophysical.

Figure 4.2 shows some representative candidate signals from this final list. For most of the incorrectly labeled candidates, it is obvious to the human eye that the signal is not caused by a transit. The false signals seemed to be mostly caused by "sudden pixel sensitivity dropouts" (SPSDs Christiansen et al., 2013) or non-stationary variability of the star. The SPSDs are misclassified because they are actually extremely rare so the training set contains no similar signals. This problem could be mitigated by simulating these sorts of signals and injecting them into real light curves as negative examples. The non-stationarity of the stellar variability is much harder to simulate but it might be possible to vet these candidates in a post-processing step.

4.9 KIC 10602068: A discovery

In the 3500 light curves we searched for this Chapter, we discovered one convincing single transit signal at 830.8093 ± 0.0002 KBJD in the light curve of the 14.9 Kep-mag G-dwarf

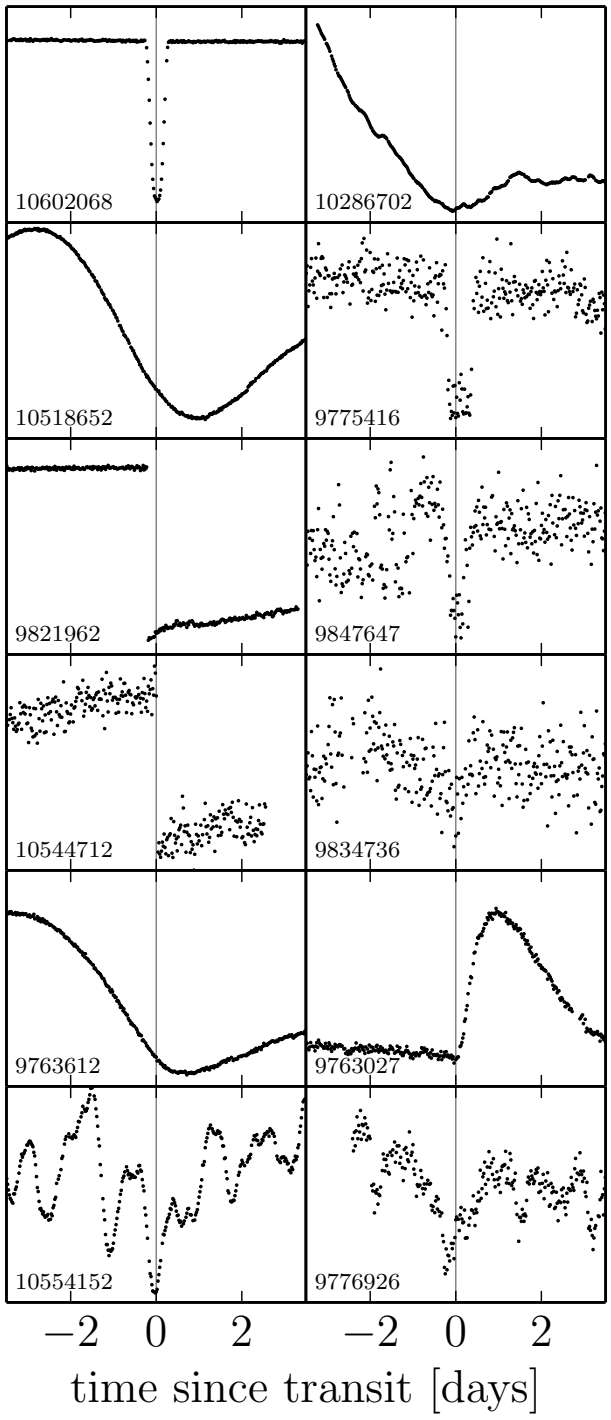

Figure 4.2: Some representative light curve sections that were labeled as candidates by the single transit search. The light curve in the top-left panel is a convincing transit signal but the other panels all appear to be caused by stellar variability or instrumental effects.

KIC 10602068. Despite the fact that this is a very large signal, its discovery has never been reported. Given the estimate of ~ 60 detectable single transits in the full *Kepler* archival dataset, we do expect about one discovery in 3500 light curves.

Even though only one transit is observed, if we assume that the transit is caused by a body on a Keplerian orbit around the host star, we can place constraints on the physical properties—even the orbital period—of the transiting candidate. This technique is similar in spirit to the "photoeccentric effect" (Dawson & Johnson, 2012) and "asterodensity profiling" (Kipping et al., 2012). The fundamental principle is that the transit duration is a measurement of the instantaneous velocity of the planetary motion and this yields a constraint on the orbital period when combining this with a measurement of the stellar density and a prior on the eccentricity of the orbit.

The photometrically derived physical properties for this star place it as a G-dwarf with a mass and radius of about 90 percent Solar (Huber et al., 2014). Using these constraints and a beta function prior on the eccentricity of the orbit (Kipping, 2013b), we run Markov Chain Monte Carlo (Foreman-Mackey et al., 2013) to sample the posterior probability for the physical parameters of the system. In this analysis, we take limb darkening and the finite exposure time into account (Mandel & Agol, 2002; Kipping, 2010, 2013a). The results of this chain marginalized into the relevant physical dimensions are shown in Figure 4.3. The radius of the transiting body is measured to be $2.1 \pm 0.4\, R_{\rm J}$. The period is very weakly constrained above a lower limit of 760 days—set by the fact that a second transit is not detected—with 68 percent of the posterior mass in the range $760\,{\rm d} < P < 1347\,{\rm d}$.

Given the large radius of this candidate, it is probably stellar in nature instead of planetary but this could be easily confirmed using radial velocity follow-up.

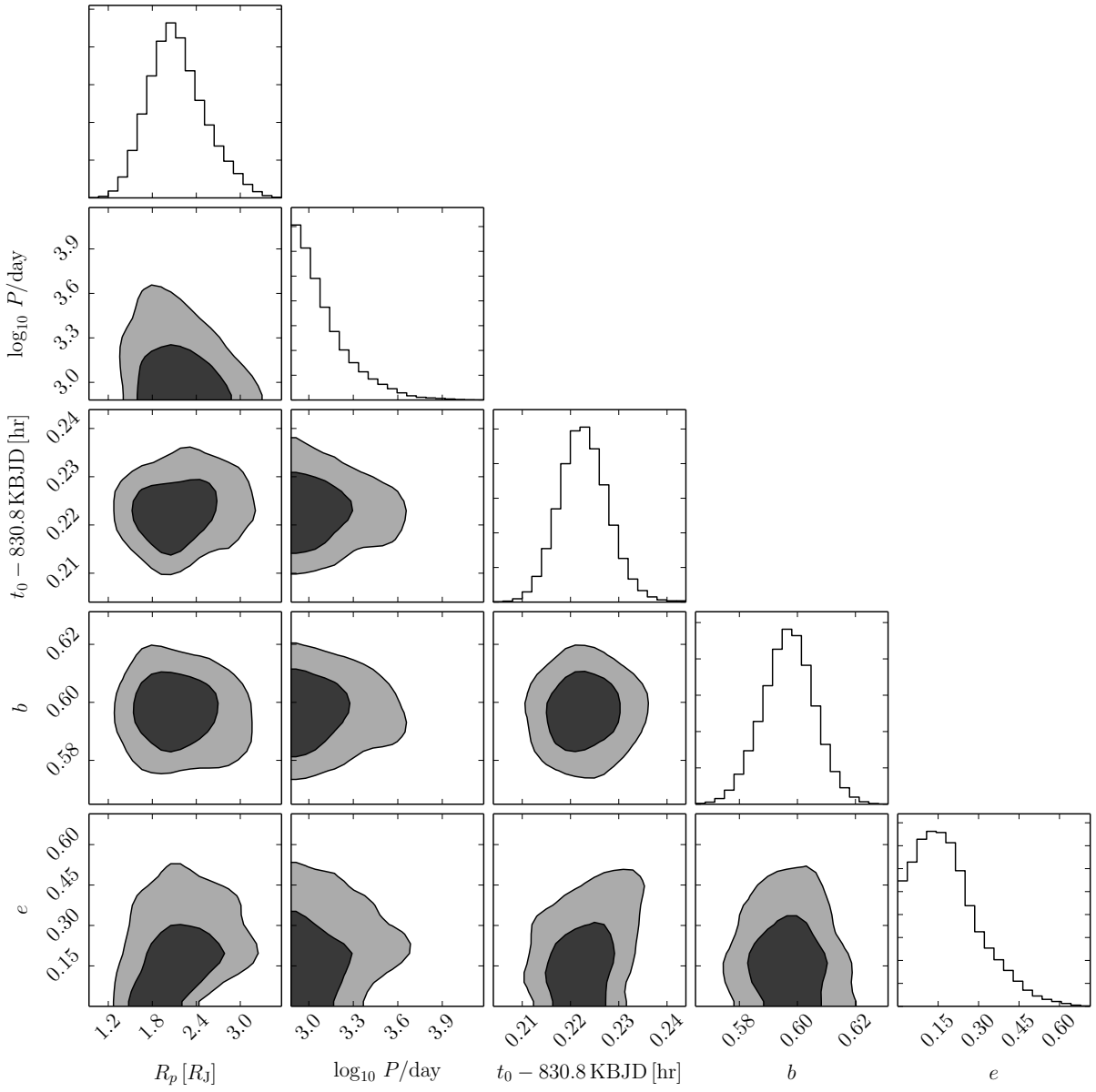

Figure 4.3: Posterior constraints on the physical parameters of the body transiting KIC 10602068 assuming a bound Keplerian orbit.

4.10 Discussion

The discovery and characterization of transiting planets based on a single transit event is crucial for the future of transiting exoplanet surveys. Many of the most dynamically influential planets—like Jupiter in our Solar system—exhibit only a single transit in the full observational baseline of the *Kepler*. This will become even more of a problem as upcoming surveys move to shorter contiguous observations. For example, the *TESS* Mission is planned to get full-sky coverage at half-hour cadence but most of the sky will only be targeted for a month. This means that even habitable zone planets orbiting cool stars will transit their host *at most once in the entire lifetime of the Mission*!

To date, no methods exist for systematically and robustly discovering single transit events based on large photometric surveys. In this Chapter, we present a novel and conceptually unique solution to this problem drawing on machine learning methods for supervised classification. This method has immense potential because it can be designed to be very robust to false positives and it can exploit the detailed shape of physical transits.

Despite the fact that single transits are unlikely even if these long-period planets are intrinsically common, we estimate that ~ 60 events should be detectable in the *Kepler* archival dataset by extrapolating recent models of planet occurrence rates and taking selection effects into account. When applied to 3500 light curves from the *Kepler* dataset, this method recovers one previously unknown single transit event at 830.8093 ± 0.0002 KBJD in the light curve of KIC 10602068. This rate is consistent with the predicted yield of 60 events in the full dataset of 190,000 light curves.

Assuming a bound Keplerian orbit, we place constraints on the physical properties of this transit candidate KIC 10602068.01. Using photometrically derived stellar properties, we find that this candidate has a radius of $2.1 \pm 0.4\,R_{\rm J}$ placing it as a very large planet or brown dwarf or a small star.

This method for transit search is built using supervised classification and its performance relies on several strong assumptions about the datasets and these assumptions are sometimes violated leading to some transit-like signals that appear to be caused by noise to be misclassified as transits. The most severe assumption is that the noise properties of the data are stationary. In other words, we assume that variability in one subset of a light curve is completely spanned by the variability in the other sections. This assumption is, in general, false because the stochastic processes that cause stellar variability are complicated and non-stationary and the detector is plagued by non-negligible catastrophic changes in sensitivity and response. We attempt to mitigate this problem by using light curves that have been preprocessed to remove most of the instrumental effects and carefully dividing the data into subsets but some false positives are still incorrectly identified as candidates.

One possible method for reducing the false positive rate would be to augment the training dataset using heuristic simulations of common false positives or the light curves of "similar" stars. Another option is to recognize that this search drastically reduces the parameter space requiring evaluation and comparing the predictive power of a transit model to other heuristic models including stellar variability or instrumental effects.

As discussed in Section 4.9, many decisions were made in the application of this method and the related hyperparameters were set heuristically. Instead, substantial gains could be made by optimizing these choices objectively, especially on the edge cases and methodological failures. In particular, some different combinations of feature selection and regularization should improve the performance.

Conclusion

In this dissertation, we study the population of exoplanets using data from NASA's *Kepler* Mission and the re-purposed *K2* Mission. We develop and apply novel techniques to discover previously unknown planets and planet candidates (Chapters 3 and 4). We present a robust probabilistic framework for making inferences about the population of exoplanets based on the noisy and incomplete catalogs derived from transit surveys (Chapter 2). The main contributions of this dissertation are methodological and each Chapter is accompanied by open source software implementing the methods.

In the spirit of tool development and open source software, Chapter 1 is describes *emcee*, a general purpose Markov Chain Monte Carlo sampler that, since its release (Foreman-Mackey et al., 2013), has become one of the most popular tools for probabilistic inference in astronomy. This method was originally proposed by Goodman & Weare (2010) and it was designed to sample problems efficiently with little tuning even when the parameter space is poorly conditioned. This feature is especially useful for problems in astronomy where the physical parameters often vary (and covary) over many orders of magnitude. The *emcee* implementation offers a small performance gain by deriving a parallelizable version of the original algorithm and a user-friendly and well documented Python interface. In practice, this method doesn't scale well to large numbers of dimensions ($\gtrsim 50$) but it has been shown to work out-of-the-box on a large class of typical astronomy problems.

In Chapter 2, we derive a hierarchical method for inferring the population of exoplanets based on a catalog of planets with a non-trivial completeness function and large measurement uncertainties. This method builds on the importance sampling technique originally derived by Hogg et al. (2010b) to make a clean histogram from noisy measurements. Applying this population inference method to a catalog of planet candidates transiting Sun-like stars (Petigura et al., 2013a), we make a prediction for the rate of Earth analogs. This prediction is substantially lower than earlier predictions based on the same catalog. We demonstrate that this discrepancy is caused by both the treatment of the observational uncertainties and the choice of extrapolation function.

In Summer 2014, the *Kepler* spacecraft was re-purposed and it began taking data for the *K2* Mission. The pointing accuracy in this mode is substantially degraded relative to the original Mission but, in Chapter 3, we demonstrate that these light curves can still be used to systematically search for transiting exoplanets. By building a flexible data-driven model for the systematic variability in the light curves of the stars and combining this with an approximate linear transit model, we derive a transit search algorithm where the systematics model is marginalized for every hypothesis. This enables the discovery of transit signals with amplitudes smaller than the pointing-induced variability. In Chapter 3, we announce the discovery of 36 planet candidates transiting 33 stars. Of these candidates, 18 have been validated as bona fide planets and 6 have been identified as likely astrophysical false positives (Crossfield et al., 2015; Montet et al., 2015; Armstrong et al., 2015b).

Finally, in Chapter 4, we present a novel method for detecting the transits of planets with orbital periods longer than the baseline of observations. Existing transit search methods are blind to these long periods because it is technically difficult to distinguish a single transit from coincidental variability in light curves. This constraint is not acceptable for forthcoming surveys like *K2* and *TESS* where the observation baselines are shorter than the periods of

the most important planets for studies of dynamics and habitability. We apply a supervised classification algorithm, implemented using a set of Random Forest classifiers trained on simulated transits, to predict the "class" (`transit` or `no transit`) of every section of light curve. Using this method, we announce the discovery of a convincing single transit candidate with a radius of $\sim 2\,R_{\rm J}$.

The ultimate goal of this research program is an improved understanding of the population of exoplanets at the currently uncharted extremes of parameter space, especially pushing to long periods. This dissertation represents a step in this direction but there are some conspicuously missing pieces in the methods presented in these pages. One major shortcoming is that neither Chapter 3 or Chapter 4 realized the dream of a fully automated search. In both projects, a final stage of manual vetting was required to reach the target precision. This is unacceptable if we want to make rigorous inferences of the population of planets because human components of a pipeline can't be stress-tested and characterized for consistency and performance. The main barrier to completely automated search is that we don't have an acceptable generative model for the signals that are mis-classified by the search algorithms and we can never be completely sure that any section of light curve *does not have any transits*. This goal of fully automated transit discovery will become even more important as new datasets continue to roll in from *K2*, *TESS*, and *PLATO*. This should be a focus of large scale transit programs over the next years.

Bibliography

Aigrain, S., Hodgkin, S. T., Irwin, M. J., Lewis, J. R., & Roberts, S. J. 2015, MNRAS, 447, 2880

Almenara, J. M., Deeg, H. J., Aigrain, S., et al. 2009, A&A, 506, 337

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

Armstrong, D. J., Osborn, H. P., Brown, D. J. A., et al. 2014, ArXiv e-prints, arXiv:1411.6830

Armstrong, D. J., Kirk, J., Lam, K. W. F., et al. 2015a, ArXiv e-prints, arXiv:1502.04004

Armstrong, D. J., Veras, D., Barros, S. C. C., et al. 2015b, ArXiv e-prints, arXiv:1503.00692

Barclay, T., Rowe, J. F., Lissauer, J. J., et al. 2013, Nature, 494, 452

Basri, G., Walkowicz, L. M., & Reiners, A. 2013, ApJ, 769, 37

Bastien, F. A., Stassun, K. G., & Pepper, J. 2014, ApJ, 788, L9

Batalha, N. M., Rowe, J. F., Gilliland, R. L., et al. 2010, ApJ, 713, L103

Batalha, N. M., Rowe, J. F., Bryson, S. T., et al. 2013, ApJS, 204, 24

Berta, Z. K., Irwin, J., Charbonneau, D., Burke, C. J., & Falco, E. E. 2012, AJ, 144, 145

Borucki, W. J., Koch, D., Basri, G., et al. 2010, Science, 327, 977

Breiman, L. 2001, Machine learning, 45, 5

Brewer, B. J., Pártay, L. B., & Csányi, G. 2011, Statistics and Computing, 21, 649

Bryson, S. T., Jenkins, J. M., Gilliland, R. L., et al. 2013, PASP, 125, 889

Burke, C. J., Bryson, S. T., Mullally, F., et al. 2014, ApJS, 210, 19

Butler, R. P., Wright, J. T., Marcy, G. W., et al. 2006, ApJ, 646, 505

Carter, J. A., & Agol, E. 2013, ApJ, 765, 132

Catanzarite, J., & Shao, M. 2011, ApJ, 738, 151

Christiansen, J. L., Jenkins, J. M., Caldwell, D. A., et al. 2012, PASP, 124, 1279

Christiansen, J. L., Clarke, B. D., Burke, C. J., et al. 2013, ApJS, 207, 35

Cowles, M. K., & Carlin, B. P. 1996, Journal of the American Statistical Association, 91, 883

Crossfield, I. J. M., Petigura, E., Schlieder, J., et al. 2015, ArXiv e-prints, arXiv:1501.03798

Cumming, A., Butler, R. P., Marcy, G. W., et al. 2008, PASP, 120, 531

Dawson, R. I., & Johnson, J. A. 2012, ApJ, 756, 122

Dieleman, S., Willett, K. W., & Dambre, J. 2015, ArXiv e-prints, arXiv:1503.07077

Dong, S., & Zhu, Z. 2013, ApJ, 778, 53

Dressing, C. D., & Charbonneau, D. 2013, ApJ, 767, 95

—. 2015, ArXiv e-prints, arXiv:1501.01623

Dunkley, J., Bucher, M., Ferreira, P. G., Moodley, K., & Skordis, C. 2005, MNRAS, 356, 925

Fadely, R., Hogg, D. W., & Willman, B. 2012, ApJ, 760, 15

Fischer, D. A., Schwamb, M. E., Schawinski, K., et al. 2012, MNRAS, 419, 2900

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306

Foreman-Mackey, D., Hogg, D. W., & Morton, T. D. 2014, ApJ, 795, 64

Foreman-Mackey, D., Montet, B. T., Hogg, D. W., et al. 2015, ArXiv e-prints, arXiv:1502.04715

Fressin, F., Torres, G., Charbonneau, D., et al. 2013, ApJ, 766, 81

Gelman, A., Roberts, G., & Gilks, W. 1996, Bayesian statistics 5

Gibson, N. P., Aigrain, S., Roberts, S., et al. 2012, MNRAS, 419, 2683

Gilliland, R. L., Chaplin, W. J., Dunham, E. W., et al. 2011, ApJS, 197, 6

Girardi, L., Groenewegen, M. A. T., Hatziminaoglou, E., & da Costa, L. 2005, A&A, 436, 895

Goodman, J., & Weare, J. 2010, Communications in Applied Mathematics and Computational Science, 5, 65

Gregory, P. C. 2005, Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with 'Mathematica' Support (Cambridge University Press)

Hogg, D. W., Bovy, J., & Lang, D. 2010a, ArXiv e-prints, arXiv:1008.4686

Hogg, D. W., Myers, A. D., & Bovy, J. 2010b, ApJ, 725, 2166

Hou, F., Goodman, J., Hogg, D. W., Weare, J., & Schwab, C. 2012, ApJ, 745, 198

Howard, A. W., Marcy, G. W., Bryson, S. T., et al. 2012, ApJS, 201, 15

Howell, S. B., Sobeck, C., Haas, M., et al. 2014, PASP, 126, 398

Huang, X., Bakos, G. Á., & Hartman, J. D. 2013, MNRAS, 429, 2001

Huber, D., Silva Aguirre, V., Matthews, J. M., et al. 2014, ApJS, 211, 2

Ivezić, Ż., Connolly, A., VanderPlas, J., & Gray, A. 2013, Statistics, Data Mining, and Machine Learning in Astronomy (Princeton University Press)

Jenkins, J. M., McCauliff, S., Burke, C., et al. 2014, in IAU Symposium, Vol. 293, IAU Symposium, ed. N. Haghighipour, 94–99

Kipping, D. M. 2010, MNRAS, 408, 1758

—. 2013a, MNRAS, 435, 2152

—. 2013b, MNRAS, 434, L51

—. 2014, MNRAS, 444, 2263

Kipping, D. M., Dunn, W. R., Jasinski, J. M., & Manthri, V. P. 2012, MNRAS, 421, 1166

Kipping, D. M., Torres, G., Buchhave, L. A., et al. 2014, ApJ, 795, 25

Knutson, H. A., Fulton, B. J., Montet, B. T., et al. 2014, ApJ, 785, 126

Kovács, G., Bakos, G., & Noyes, R. W. 2005, MNRAS, 356, 557

Kovács, G., Zucker, S., & Mazeh, T. 2002, A&A, 391, 369

Latham, D. W., Bakos, G. Á., Torres, G., et al. 2009, ApJ, 704, 1107

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

Lissauer, J. J., Ragozzine, D., Fabrycky, D. C., et al. 2011, ApJS, 197, 8

MacKay, D. J. 2003, Information theory, inference, and learning algorithms, Vol. 7 (Citeseer)

Mandel, K., & Agol, E. 2002, ApJ, 580, L171

Matijevič, G., Prša, A., Orosz, J. A., et al. 2012, AJ, 143, 123

McCauliff, S., Jenkins, J. M., Catanzarite, J., et al. 2014, ArXiv e-prints, arXiv:1408.1496

McQuillan, A., Mazeh, T., & Aigrain, S. 2014, ApJS, 211, 24

Montet, B. T., Morton, T. D., Foreman-Mackey, D., et al. 2015, ArXiv e-prints, arXiv:1503.07866

Morton, T. D. 2012, ApJ, 761, 6

Morton, T. D., & Johnson, J. A. 2011, ApJ, 738, 170

Morton, T. D., & Swift, J. 2014, ApJ, 791, 10

Murray, I., & Adams, R. P. 2010, Advances in Neural Information Processing Systems, 23, 1723

Murray, I., Adams, R. P., & MacKay, D. J. 2010, JMLR: W&CP, 9, 541

O'Donovan, F. T., Charbonneau, D., Torres, G., et al. 2006, ApJ, 644, 1237

Ofir, A., Alonso, R., Bonomo, A. S., et al. 2010, MNRAS, 404, L99

Pasarica, C., & Gelman, A. 2010, Statistica Sinica, 20, 343

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Petigura, E. A., Howard, A. W., & Marcy, G. W. 2013a, Proceedings of the National Academy of Science, 110, 19273

Petigura, E. A., Marcy, G. W., & Howard, A. W. 2013b, ApJ, 770, 69

Poleski, R., McCullough, P. R., Valenti, J. A., et al. 2010, ApJS, 189, 134

Press, W. H. 2007, Numerical recipes 3rd edition: The art of scientific computing (Cambridge university press)

Rasmussen, C. E., & Williams, C. K. I. 2006, Gaussian processes for machine learning (The MIT Press)

Richards, J. W., Starr, D. L., Miller, A. A., et al. 2012, ApJS, 203, 32

Richards, J. W., Starr, D. L., Butler, N. R., et al. 2011, ApJ, 733, 10

Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2014, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9143, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 20

Rogers, L. A. 2015, ApJ, 801, 41

Rowe, J. F., Coughlin, J. L., Antoci, V., et al. 2015, ApJS, 217, 16

Sanchis-Ojeda, R., Rappaport, S., Winn, J. N., et al. 2014, ApJ, 787, 47

Smith, J. C., Stumpe, M. C., Van Cleve, J. E., et al. 2012, PASP, 124, 1000

Stumpe, M. C., Smith, J. C., Van Cleve, J. E., et al. 2012, PASP, 124, 985

Swift, J. J., Johnson, J. A., Morton, T. D., et al. 2013, ApJ, 764, 105

Tabachnik, S., & Tremaine, S. 2002, MNRAS, 335, 151

Tamuz, O., Mazeh, T., & Zucker, S. 2005, MNRAS, 356, 1466

Thompson, S. E., Jenkins, J. M., Caldwell, D. A., et al. 2015, Kepler Data Release 24 Notes (KSCI-19064-001)

Traub, W. A. 2012, ApJ, 745, 20

Tremaine, S., & Dong, S. 2012, AJ, 143, 94

Vanderburg, A., & Johnson, J. A. 2014, PASP, 126, 948

Vanderburg, A., Montet, B. T., Johnson, J. A., et al. 2014, ArXiv e-prints, arXiv:1412.5674

Wang, D., Foreman-Mackey, D., Hogg, D. W., & Schölkopf, B. 2015, in American Astronomical Society Meeting Abstracts, Vol. 225, American Astronomical Society Meeting Abstracts, 258.08

Wang, J., Fischer, D. A., Barclay, T., et al. 2013, ApJ, 776, 10

Weiss, L. M., & Marcy, G. W. 2014, ApJ, 783, L6

Widrow, L. M., Pym, B., & Dubinski, J. 2008, ApJ, 679, 1239

Winn, J. N. 2010, ArXiv e-prints, arXiv:1001.2010

Yee, J. C., & Gaudi, B. S. 2008, ApJ, 688, 616

Youdin, A. N. 2011, ApJ, 742, 38