Exoplanet Detection: A Comparison of Three Statistics *or* How Long Should It Take To Find a Small Planet?

by

Floyd Bullard

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Michael Lavine, Advisor

_____
Jim Berger

_____
Merlise Clyde

_____
Tom Loredo

_____
Dalene Stangl

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2009

Abstract

(Statistics)

Exoplanet Detection: A Comparison of Three Statistics *or* How Long Should It Take To Find a Small Planet?

by

Floyd Bullard

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Michael Lavine, Advisor

_____
Jim Berger

_____
Merlise Clyde

_____
Tom Loredo

_____
Dalene Stangl

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2009

Copyright © 2009 by Floyd Bullard
All rights reserved

Abstract

If a distant star happens to host an orbiting exoplanet, then that planet will exert a gravitational influence on the star that may be detectable from the earth by the apparent "stellar wobble"—regular, periodic variations in the observed radial velocities. Two complicating factors, however are "stellar jitter" and measurement errors, both of which generate radial velocity "noise". As a result, it is not always obvious whether variations in radial velocities are due to an orbiting planet or to mere noise. The problem of model selection—is that a planet or not?—may be addressed by using any number of summary statistics, and the choice may be made using frequentist hypothesis testing, Bayesian model selection, or likelihood analysis.

Here we compare the planet-detecting ability of three different statistics: the "Maximal Periodogram Spike", or the height of the tallest spike in the Lomb-Scargle periodogram; the Maximum Likelihood Ratio; and the Bayes Factor. We measure the usefulness of each statistic by how many observations would be required before the statistic would able to distinguish accurately the no-planet model from the one-planet model when presented with data from a star hosting a planet of one of six different types—we vary the magnitude of the radial velocity function's semi-amplitude and the eccentricity of the orbit. The data are simulated, using observations made at random times.

We find that the current practice of examining the tallest spike in the periodogram cannot easily be improved upon using either the Maximum Likelihood Ratio (which

performs comparably) or the Bayes Factor (which is exceedingly difficult to estimate for eccentric orbits unless it is already plain that a planet is present). We also conclude that a planet having a semiamplitude of 3 m/s (which could potentially correspond to the smallest planet yet discovered outside of a multi-planet system) could in fact be detected using current methods, if there was an initial commitment to make at least 150 or 200 radial velocity observations on the same candidate star, even in the absence of growing evidence for planetary presence.

Dedicated To My Parents

Contents

List of Figures

List of Abbreviations and Symbols

Symbols and Abbreviations

Frequently used symbols and abbreviations in this paper include the following:

- X The radial velocity observations on a star includes these components:
  - n The number of observations made on a star.
  - t_i The time of the i^{th} observation, usually measured in days since an arbitrary starting date.
  - V_i The i^{th} measured radial velocity of the star, usually measured in m s^{-1}.
  - σ_i The measurement error associated with V_i, usually measured in m s^{-1}.
- C The "background velocity" of the star, usually measured in m s^{-1} or km s^{-1}.
- s The stellar jitter, usually measured in m s^{-1}.
- θ The five additional orbital parameters of a star are:
  - K The radial velocity semiamplitude, usually measured in m s^{-1}.
  - P The period of the orbit, usually measured in days.
  - e The eccentricity of the orbit—a dimensionless quantity in the interval $[0, 1)$, zero indicating a circular orbit.
  - ω The argument of periatron, unitless and measured in radians.
  - M_0 The mean anomaly at time $t = 0$, unitless and measured in radians.

  Transformed parameters:

x	$x = e\cos(\omega)$, a Poincaré variable.
y	$y = e\sin(\omega)$, another Poincaré variable.
γ	$\gamma = \omega + M_0$.
M0	The no-planet model.
M1	The one-planet model.
$\Upsilon(P)$	The periodogram function value at a candidate period P.
MPS	Maximal Periodogram Spike
ROC curve	Receiver Operating Characteristic curve
IMDA	Index of model distinguishing ability (ranges from 0 to 1)
MLE	Maximum Likelihood Estimate
Λ	The Maximum Likelihood Ratio of M1 to M0.
BFE	Bayes Factor Estimate
ESS	Effective Sample Size

Acknowledgements

This dissertation would never have been possible without the generous financial support I received from the Duke University Graduate School, the Department of Statistical Science, and the Statistical and Applied Mathematical Sciences Institute (SAMSI), for which I am extremely grateful.

I am also grateful for all the help I received from all the members of my committee—Michael Lavine, Jim Berger, Merlise Clyde, Tom Loredo, and Dalene Stangl—as well as other people involved in exoplanet research, including Bill Jefferys (thank you for the books!), Eric Ford, and Barbara McArthur. Their patience with a novice in the fields of both statistics and astronomy is remarkable.

My partner Rick has seen me through the entire graduate school process from beginning to end, and has never failed to give me completely selfless support. I would never have completed this journey without him.

Finally, I thank my parents. From my childhood they helped me to see that learning is fun and rewarding—even if it sometimes requires much hard work! This dissertation is dedicated to them.

Introduction

In December of 2000 I was teaching mathematics full-time at the North Carolina School of Science and Mathematics. As part of a guest lecture series, the physicist Philip Morrison visited our school and gave a talk about topics that were then hot in science. One of them was the search for extra-solar planets, or exoplanets. I remember him showing us a plot of the measured radial velocities of 51 Pegasi—the first star other than the Sun confirmed to host an orbiting planet—against observation times. The plot was clearly sinusoidal, so to me it cried out to be turned into an application of trigonometry that we at my school could share with our precalculus students. Creating "The Exoplanet Lab" was lots of fun, and I was particularly pleased to receive, after a request by email for more raw data, a personal reply from Geoff Marcy, who was then, and still remains, one of the leaders in exoplanet detection research. He was delighted that the research he was involved in would be shared with high school students.

At the same time, my interest in Bayesian statistical methods was growing, after having been introduced to the concept in the summer of 1999 by Dalene Stangl, who was a guest speaker at a workshop I attended for high school statistics teachers. When I decided three years later to apply to be a graduate student in Duke University's Statistics Department, I had a notion that, though it seemed unlikely to come to fruition, I might be able to turn "The Exoplanet Lab" into a doctoral dissertation. In its present form, it is not what I first imagined—in particular, it is

not very Bayesian—yet it is an outgrowth of that early research I did in exoplanet detection while preparing a project for my precalculus students.

There are two basic questions that I've addressed in this dissertation:

1. When trying to decide whether radial velocity perturbations are due to a planet or to mere noise, which of three summary statistics is the best tool for making that judgment:

    - the height of the tallest spike in the periodogram (this is what is used in practice);
    - the maximum likelihood ratio, comparing the one-planet model to the no-planet model; or
    - the Bayes factor, also comparing the one-planet model to the no-planet model?

2. For each of the statistics above, how many observations must one be prepared to make if one's goal is to detect a planet of a certain type? I decided to consider six variations on the orbital parameters, which are as follows:

    - circular orbit, with semiamplitude $K = 50$ m s^{-1} (for small periods, this would be a very "loud" star, practically screaming "planet!");
    - circular orbit, with semiamplitude $K = 20$ m s^{-1};
    - circular orbit, with semiamplitude $K = 7$ m s^{-1};
    - circular orbit, with semiamplitude $K = 3$ m s^{-1} (even for small periods, this would be a "whisper-quiet" star, not eager to give away its beautiful secret);
    - eccentric orbit having eccentricity $e = 0.3$ and semiamplitude $K = 20$ m s^{-1}; and finally,

- eccentric orbit having eccentricity $e = 0.6$ and semiamplitude $K = 20$ m s^{-1}.

Although the actual mass of a planet depends upon the orbital period and the mass of the host star, the first four of these scenarios above correspond roughly to planets whose sizes range from very large to fairly small. I use the term "fairly small" loosely, because even allowing for a short orbital period and a small host star, which both favor small planet detection, a planet inducing an orbital period having semiamplitude 3 m s^{-1} on its host star would still be somewhat larger than the Earth and, due to its proximity to its sun, probably too hot to host life. But even so, it would be one of the smallest planets ever detected. The semiamplitude of 3 m s^{-1} is comparable in magnitude to typical stellar jitter magnitudes, making planetary detection difficult. The next-smallest semiamplitude I considered, 7 m s^{-1} is comparable in magnitude to the combined noise due to stellar jitter and measurement errors.

I included the eccentric orbit scenarios in the list above to see whether eccentricity made a planet harder to detect and if so, how much harder? It is not obvious that eccentricity would in fact make a planet harder to detect at all. There are competing factors to consider. On the one hand, a star with an eccentric orbit may spend a great deal of time sitting (apparently) relatively still, with only occasional bursts of motion as the planet passes close to the star. In such a case, the detection of the planet would depend on whether any of the observation times had been "lucky" enough to catch the burst of motion. On the other hand, if that burst of motion *is* caught, then it can show a much greater radial velocity differential during the burst than would ever be exhibited by a planet of comparable mass that had a circular orbit; the observation of such a large burst would favor planetary detection. Finally, I wanted to see whether the periodogram as a detection tool would be rendered relatively blind to eccentric motion, since it is designed to detect sinusoidal patterns.

As such, it was of particular interest whether either of the other two statistics—the maximum likelihood ratio or the Bayes factor—would outperform the periodogram spike when looking at eccentric orbits.

For each of the six scenarios above, I simulated data for a star having an orbital period of 7 days, with observations made at random times over six periods: 42 days. I did not vary the period at all but treated it as a constant. So long as the observations are made at random times over a duration spanning several periods, the actual period will not affect the detectability of the planet. Only the time scale would change, not the radial velocity measurements.

For each of the scenarios, I also simulated data sets of different sizes. In particular, these eleven sample sizes were used throughout this dissertation: 4, 6, 10, 16, 25, 40, 63, 100, 159, 252, and 400. These numbers are evenly separated on a logarithmic scale, and as it happens, the appropriate way to look at the relative usefulness of the different sample sizes is on a logarithmic scale.

The simulations included both stellar jitter and measurement errors, of magnitudes comparable to those that have been observed on stars in the past. To my knowledge, no analysis has been done that evaluates the effectiveness of the periodogram spike in planet detection when both stellar jitter and measurement errors are included in the actual data.

After simulating at least 1000 data sets[1] for each scenario and for each sample size, it was then necessary to evaluate how well each statistic performed at recognizing there really was a planet there. So for comparison, sets of data using the same sample size were simulated for a star that had no orbiting planet. Then each of two evaluation tools was used to measure the usefulness of the statistic for detecting a planet:

[1] It would have been preferable to perform 100,000 or even a 1,000,000 simulations; however, the computation time prohibited it. This will be discussed further at a later point in this dissertation.

- The "Index of Model Distinguishing Ability", or "IMDA" (my own term). This ranges from 0 (for no planet detection ability) to 1 (for perfect detection ability) and is derived from the "receiver operating characteristic" curve, or ROC curve, which plots true positive rates against false positive rates—or in hypothesis testing lingo, it plots the power of a test against the significance level of the test.

- The probability of obtaining "strong evidence for" the one-planet model over the no-planet when the one-planet model is the correct one, with "strong evidence for" being defined as having a likelihood ratio of at least 32 to 1 in favor of the one-planet model.

The estimation of the density of the statistic required non-parametric kernel density estimation. For this I simply used the normal density as the kernel, and used a bandwidth chosen automatically by the Matlab function **ksdensity**.

Unfortunately, the Bayes factor was so difficult to compute for the one-planet model with eccentric orbits, that at that point I had to restrict myself to considering circular orbits. This still provided an interesting look at how many observations would be required to detect large and small planets using the Bayes factor as a detection tool.

In the end I came to the conclusion that the maximal periodogram spike is much easier to compute than the other two statistics and performs comparably well or better when it comes to detecting planets. This does not mean that these other two statistics may not be superb tools for other purposes, such as estimating orbital parameters or adaptively determining optimal future observation times. But for the purpose of planetary detection only, the current practice of examining the periodogram for stand-out spikes seems to be hard to beat.

This dissertation contains two quite lengthy chapters that are tangential to the

purpose outlined above. Although much work went into them and they have their own peculiar topics of interest, I have decided not to incorporate them into the body of the dissertation proper, but to relegate them to an appendix—an appendix comprising the two longest chapters of the paper. In this way, I was able to keep the narrative flow of the paper brisk and focused on the two main research questions.

The primary chapters in the dissertation are as follows:

1. An introduction to the one-planet model and no-planet model, incorporating stellar jitter and measurement errors, and including prior distributions on parameters (for when they are relevant);

2. A brief discussion about why the problem of exoplanet detection is statistically challenging;

3. An analysis of the usefulness of the maximal periodogram spike height (the tool currently used to detect planets);

4. An analysis of the usefulness of the maximum likelihood ratio;

5. An analysis of the usefulness of the Bayes factor—but only for circular orbits; and finally,

6. A brief summary of the findings in the dissertation.

And the two chapters in the appendix are these:

1. A discussion of a possible improvement upon the maximal periodogram spike height by conditioning on the estimated stellar jitter—the improvement was not realized;

2. An excursion into an integral estimation approach called *nested sampling*, which I initially hoped might make it possible to estimate the Bayes factor efficiently,

but which turned out to be ill-suited to the multi-modal and multi-dimensional integration problem at hand;

This introductory section is meant to outline the broad goals of the dissertation and to give the reader a fairly clear idea of the methods used to address the two questions of interest. Greater detail will be given in each chapter to the methods described therein and to the results.

I conclude this introduction in the spirit of its beginning, with a personal statement. I took four years off to be a full-time graduate student at Duke University but have since returned to teaching full-time at the North Carolina School of Science and Mathematics while I complete this dissertation. During all this time there have been many ups and downs, but upon reflection I am somewhat startled by how much I have learned from my graduate school work. I went in not knowing what a periodogram was, nor what an MCMC sampling procedure was, nor what a conjugate distribution was. During my preliminary oral exam in January of 2005, I still hadn't clearly sorted out in my mind why the likelihoods of posterior samples from an MCMC chain couldn't be averaged to estimate the marginal likelihood of a model.

When I was in college I once, on a whim, sought and found my father's doctoral dissertation (on "The Hypostasis of the Archons"—a Coptic text) and read the first chapter or two. When I told him I had done so, he said, "Why would you want to do a thing like *that*?" It has been over ten years since I completed my master's thesis in statistics at UNC-Chapel Hill, on a non-parametric estimator of the home range of an animal. Since then, I have occasionally been contacted by email from people as far away as the Netherlands who have found references to my thesis on the Internet, but have not found the paper itself. (It was inadvertently lost by the library at UNC, but I have since sent them another copy.) So now I wonder now about the fate of

this present dissertation. Will it end up covered with virtual dust on a dark unused corner of the Internet? Will the occasional person interested in exoplanets will pull it out and read a page or two? ("Why would you want to do a thing like *that*?")

But for you, present reader who is perhaps contemplating reading further, I have tried to make the dissertation flow like a narrative, with as few stumbling blocks as possible along the way. I have tried to make the topic clear and concise while still including sufficient detail to make my methods and reasoning explicit. A dear friend of mine who counseled me on the writing process assured me that the doctoral dissertation need not be the greatest piece of writing a person will ever produce. The most important characteristic it requires is being finished. And whatever its shortcomings may be, it at last does at least, and at last, possess that necessary quality.

1

Overview of the Radial Velocity Detection Method

The *radial velocity* of a star is its velocity relative to the solar system barycenter, measured along the radius from the sun to the star.[1] Incredibly, radial velocities are measurable with a precision of below one meter per second, due to the Doppler shift in the star's spectrum—*i.e.*, its tendency towards higher frequencies ("blue-shifting") when the star is moving towards us and lower frequencies ("red-shifting") when the star is moving away from us.

If a star does not host any orbiting planets, then the measured radial velocities will be roughly constant over any period of time, varying only due to *stellar jitter* (the inconstancy of a star's surface) and measurement errors. A typical plot of radial velocity versus time for a star hosting no planets might therefore look something like that in Figure 1.1.

If a star hosts a single orbiting planet, then the measured radial velocities may reflect the Keplerian orbit of the star, which may appear roughly sinusoidal (if the orbit is nearly circular) or more asymmetrically periodic (if the orbit is more eccentric).

[1] Although measurements are made from the earth, the motion of the earth is easily subtracted out. We certainly don't want the earth's own movements to complicate the search for faraway planets.

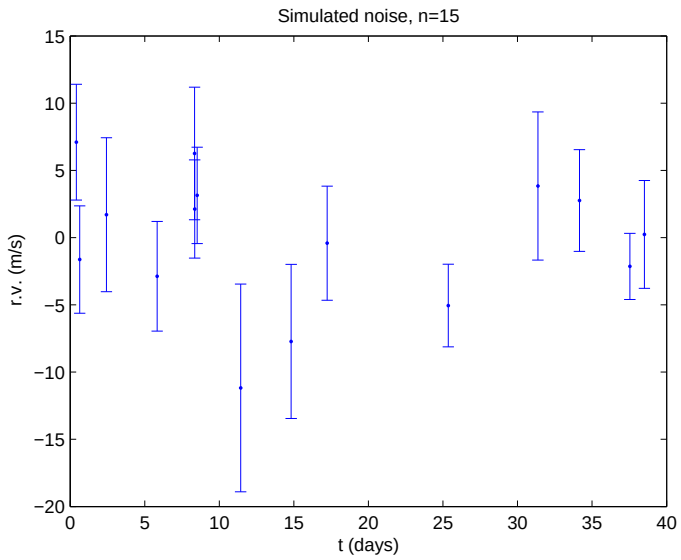

FIGURE 1.1: 15 radial velocity observations, with measurement errors, simulated from a star hosting no planet.

In Figure 1.2, we see a typical radial velocity plot for a star with a nearly circular orbit. Superimposed on the data is the radial velocity function corresponding to the maximum likelihood estimate of the orbital parameters. The data are shown modulo 3.024 days on the time scale so that each period will "wrap around" and make the periodic behavior clearer. Figure 1.3 and Figure 1.4 show stars hosting planets with eccentric orbits. The orbit of the star in Figure 1.2 has such a long period (about 576 days) that is easy to see the pattern in a single period.

In this section we will describe the mathematical models for both of these types of data—the no-planet model and the single-planet model.[2]

1.1 The Keplerian Orbit Model for Radial Velocities

When the i^{th} radial velocity is measured for a star, three quantities are derived: the time t_i of the observation, which is known precisely; the radial velocity measurement

[2] In this dissertation we will not consider the multiple-planet model. It is of great interest to scientists, but the complexity of the model is beyond the scope of this present research.

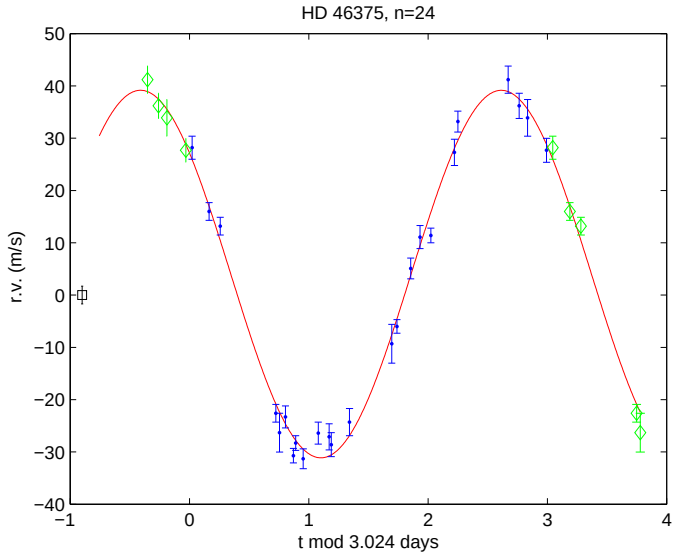

FIGURE 1.2: HD 46375, n=24. This star hosts a planet with a nearly circular orbit.

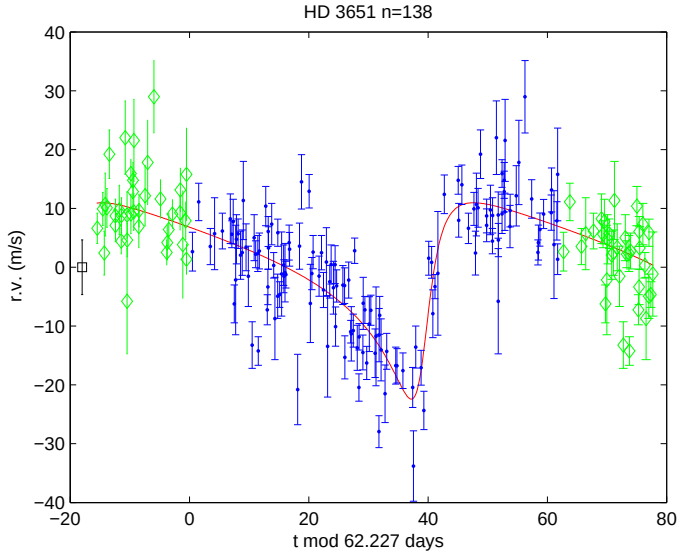

FIGURE 1.3: HD 3651, n=138. The eccentricity of this orbit is around $e = 0.63$.

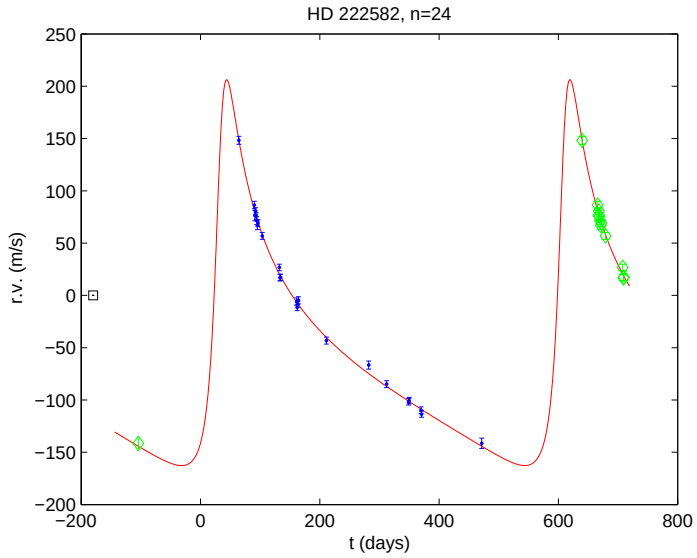

FIGURE 1.4: HD 222582, n=24. The eccentricity of this orbit is around $e = 0.71$. The long period makes the Keplerian orbit pattern clear even before a complete period has been observed.

V_i relative to a template velocity for the observatory; and an estimate σ_i of the measurement error associated with the radial velocity. t_i is usually measured in days since some reference day; V_i and σ_i are usually measured in m s^{-1}.

The actual radial velocity $V(t)$ is composed of a constant component C that reflects merely the relative motion between the center of mass of the candidate planetary system and the observer's reference frame, and which contains no information about possible orbiting planets; and a possible periodic component $\Delta V(t)$—the perturbation component—that may be due to the gravitational influence of one or more orbiting planets.

In addition to the known measurement error σ_i, there is an additional uncertainty in V_i due to "stellar jitter"—inherent variability in radial velocity due to an inconstant stellar surface. This stellar jitter we take to have a variance of s^2. If we assume that the radial velocity measurements are normally distributed and unbiased, then our preliminary model for the data is:

$$V_i \sim \mathsf{N}\left(C + \Delta V(t_i|\theta), \sigma_i^2 + s^2\right). \tag{1.1}$$

The nature of the function ΔV is determined by the orbiting planets, if there are any.

1.1.1 The no-planets model

If there are no planets orbiting a star, then the function ΔV is zero for all values of t:

$$\Delta V(t) = 0, \forall t \in \mathbb{R} \tag{1.2}$$

In that case, there are only two parameters in the model to be estimated: C and s^2, which are described below.

$C =$ a constant velocity, usually measured in m s^{-1}, sometimes km s^{-1}. C can be positive or negative. *Note:* If data on the same star are measured relative to different template observations (*e.g.*, at different observatories), then each one requires its own unique C_j, where j is an observatory index. In this paper, we will assume all observations are made at the same observatory and will only consider a single constant C.

$s^2 =$ variance of additional noise, or "stellar jitter"; s is usually measured in m s^{-1}.

1.1.2 The single planet model

If there is a single planet orbiting a star, then we introduce a parameter vector θ containing five components: $\theta = (K, P, e, \omega, M_0)$, where:

$K =$ velocity semi-amplitude, usually measured in m s^{-1}. K is, by convention, non-negative. (Allowing $K < 0$ would create an identifiability problem if M_0 and ω remained unrestricted.)

$P =$ period, usually measured in days.

$e =$ eccentricity. Unitless, $0 \leq e < 1$. $e = 0$ corresponds to a circular orbit. Larger values of e correspond to more eccentric orbits. For comparison, Earth's eccentricity is about 0.017. In our Solar system, Mercury ($e = 0.206$) and Pluto ($e = 0.248$) have the largest eccentricities.

$\omega =$ argument of periastron. (Periastron is the point at which the planet is closest to the star.) $0 \leq \omega < 2\pi$. Essentially, ω measures the angle at which we happen to be observing the elliptical orbit, and it is uninteresting from a physical point of view. ω is often measured in degrees rather than radians.

$M_0 =$ mean anomaly at time $t = 0$. $0 \leq M_0 < 2\pi$. (Mean anomaly is an angular distance of a planet from periastron. Note that for near-circular orbits, both M_0 and ω are difficult to pin down, and they become unidentifiable at $e = 0$, a perfectly circular orbit, although M_0 then becomes a simple phase shift in the circular orbit model (see the following section).

The value of the perturbation function ΔV at a time t is now conditional upon θ and we denote it $\Delta V(t|\theta)$. We also define $V(t|\theta) \equiv C + \Delta V(t|\theta)$. The following equations indicate how $\Delta V(t|\theta)$ is determined. Note that it does not have a closed expression but must be evaluated numerically at every particular time t.

$$\Delta V(t|\theta) = K\left[\cos(\omega + T(t)) + e\cos\omega\right], \text{ where} \tag{1.3}$$

$T(t)$ is the *true anomaly* at time t, and is related to $E(t)$—the *eccentric anomaly* at time t—in the following way:

$$T(t) = 2\arctan\left[\tan\left(\frac{E(t)}{2}\right)\sqrt{\frac{1+e}{1-e}}\right] \tag{1.4}$$

The eccentric anomaly $E(t)$ is related to $M(t)$—the *mean anomaly* at time t—in the following way:[3]

$$E(t) - e\sin(E(t)) = M(t) \tag{1.5}$$

And finally, the mean anomaly $M(t)$ is linearly related to t and is determined directly by the equation:

$$M(t) = \frac{2\pi}{P}t + M_0 \tag{1.6}$$

So at a given time t, from the parameters P and M_0 we obtain the mean anomaly $M(t)$; from $M(t)$ and e we obtain the eccentric anomaly $E(t)$; from $E(t)$ (and e again) we obtain the true anomaly $T(t)$, and from $T(t)$ and the other parameters we obtain the predicted radial velocity $V(t) = C + \Delta V(t|\theta)$.

Note that the equation (1.5) is transcendental, so a closed expression for $\Delta V(t|\theta)$ is not possible. There are a variety of approaches one may use to determine $E(t)$ from e and $M(t)$. Our approach here is to set $E(t) - e\sin E(t) - M(t)$ equal to 0 and solve using Newton's method. Doing so with a quadratic term usually leads to a solution with at least ten decimal place accuracy in 3 or 4 iterations.

1.1.3 Special single-planet case: a circular orbit

If $e = 0$ exactly, then equation (1.5) is no longer transcendental, ω becomes meaningless, M_0 becomes a simple phase shift, and we have $T(t) = E(t) = M(t)$. In that case:

$$\Delta V(t|\theta) = K\cos\left(\frac{2\pi}{P}t + M_0\right), \text{ when } e = 0 \tag{1.7}$$

[3] This equation is Kepler's equation, a famous transcendental equation.

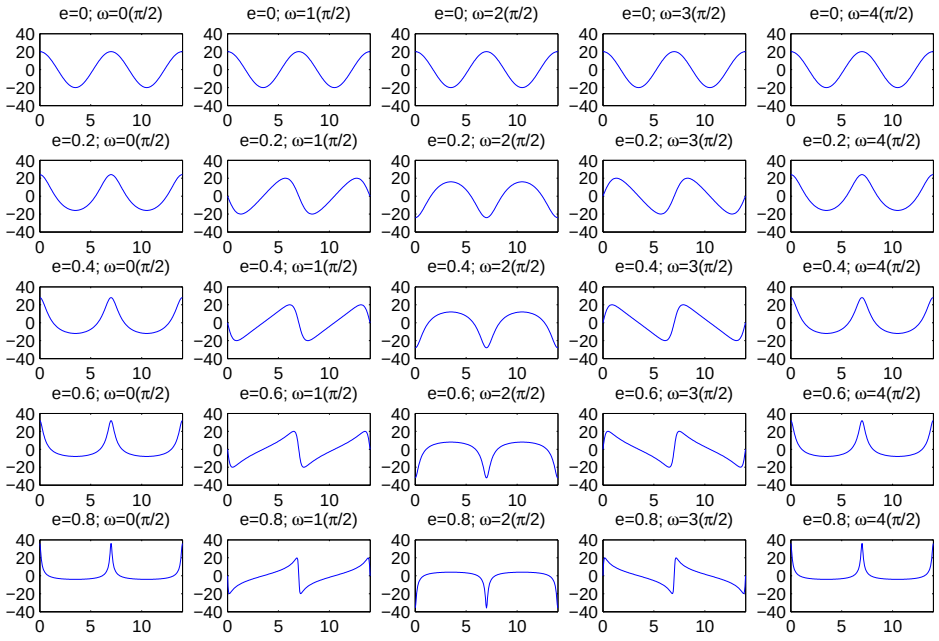

FIGURE 1.5: A star's orbital eccentricity alone does not determine the apparent shape of the radial velocity function: it is also determined by ω, the argument of periastron.

1.1.4 A comment about eccentricity

Before continuing, it may be worth pausing to consider what the radial velocity functions looks like under different eccentricities. The graphs in Figure 1.5 show $V(t|\theta)$ over 14 days given $C = 0$, $K = 20$, $P = 7$, $M_0 = 0$, and e and ω vary from 0 to 0.8 and from 0 to 2π, respectively.

1.1.5 Reparameterization

There are two reasons why correlations in the posterior distribution between the parameters are troublesome. First, when using the Metropolis algorithm with independent parameter proposal transitions, correlations make it difficult to explore the space efficiently. Second, if approximating the integral under a mode with a Gaussian

distribution, then any curvature in the posterior density will introduce error.

The following reparameterizations can help reduce correlations between the parameters substantially, particularly for systems with small eccentricities:

- Translate the data t_i (essentially reparameterizing M_0) so that $t = 0$ occurs in the middle of the observations—here we place it at the weighted mean of the observation times, with the weights being inversely proportional to the measurement errors. This reduces correlations between M_0 and P.

- Use Poincaré variables $x = e\cos\omega$ and $y = e\sin\omega$ instead of e and ω to reduce (very strong) correlations between M_0 and ω. This is particularly important for low eccentricity orbits.

- Use $z = \omega + M_0$ instead of M_0 to further reduce correlations between these two parameters when $e \ll 1$; for larger e, it still doesn't hurt.

1.2 Priors

In this section we will describe the prior distributions placed on the parameters, and the prior probabilities placed on the models. These priors were recommended by Eric Ford and Philip Gregory [1] based in part on mathematical convenience, and in part on their approximate realism. (The realism can be no more than approximate, since we don't really know what the distribution of orbital parameters is like out there, nor how many stars actually host planets. Even those we've examined are not representative, since selection bias favors large planets over small ones, small periods over long ones, and small eccentricities over high ones. And although only around 12% of observed stars have been found to host planets, it's possible—perhaps even probable—that observations on some star-hosting planets ceased before the planets made themselves known.)

1.2.1 No-planet model

Recall that this model is

$$V_i \sim \mathsf{N}\left(C, \sigma_i^2 + s^2\right),$$

so there are only two parameters: C and s. We choose a prior in which they are independent, and their marginal distributions are given by:

$$\pi_C(C) = \begin{cases} \frac{1}{C_{max} - C_{min}} & \text{for} \quad C_{min} \leq C \leq C_{max} \\ 0 & \text{otherwise} \end{cases}, \tag{1.8}$$

and

$$\pi_s(s) = \begin{cases} \frac{1}{\log\left(1+\frac{s_{max}}{s_0}\right)} \cdot \frac{1}{s_0+s} & \text{for} \quad 0 < s \leq s_{max} \\ 0 & \text{otherwise} \end{cases} \tag{1.9}$$

So the complete prior is given by:

$$\pi_0(C, s) = \kappa_0 \cdot \frac{1}{s_0 + s},$$

where

$$\kappa_0 = \frac{1}{C_{max} - C_{min}} \cdot \frac{1}{\log\left(1 + \frac{s_{max}}{s_0}\right)}$$

and where $C_{min} \leq C \leq C_{max}$, $0 < s \leq s_{max}$.

1.2.2 One-planet model with no eccentricity

Recall that this model is

$$V_i \sim \mathsf{N}\left(C + \Delta V(t_i|K, P, M_0), \sigma_i^2 + s^2\right),$$

where

$$\Delta V(t|K,P,M_0) = K\cos\left(\frac{2\pi}{P}t + M_0\right),$$

so there are five parameters: C and s as before, plus the new parameters K, P, and M_0.

We take the five parameters to be independent, and the marginal priors on C and s to be as in the no-planet model, given by equations (1) and (2). The marginal priors on K, P, and M_0 are given by:

$$\pi_K(K) = \begin{cases} \frac{1}{\log\left(1+\frac{K_{max}}{K_0}\right)} \cdot \frac{1}{K_0+K} & \text{for} \quad 0 < K \le K_{max} \\ 0 & \text{otherwise} \end{cases}, \tag{1.10}$$

$$\pi_P(P) = \begin{cases} \frac{1}{\log\left(\frac{P_{max}}{P_{min}}\right)} \cdot \frac{1}{P} & \text{for} \quad P_{min} \le P \le P_{max} \\ 0 & \text{otherwise} \end{cases} \tag{1.11}$$

(this prior is uniform on $\log P$ over the interval $\log P_{min} \le \log P \le \log P_{max}$), and

$$\pi_{M_0}(M_0) = \begin{cases} \frac{1}{2\pi} & \text{for} \quad 0 \le M_0 < 2\pi \\ 0 & \text{otherwise} \end{cases}. \tag{1.12}$$

So the complete prior is given by:

$$\pi_{1c}(C,K,P,M_0,s) = \kappa_{1c} \cdot \frac{1}{K_0+K} \cdot \frac{1}{P} \cdot \frac{1}{s_0+s},$$

where

$$\kappa_{1c} = \frac{1}{C_{max}-C_{min}} \cdot \frac{1}{\log\left(1+\frac{K_{max}}{K_0}\right)} \cdot \frac{1}{\log\left(\frac{P_{max}}{P_{min}}\right)} \cdot \frac{1}{2\pi} \cdot \frac{1}{\log\left(1+\frac{s_{max}}{s_0}\right)},$$

and where $C_{min} \le C \le C_{max}$, $0 < K \le K_{max}$, $P_{min} \le P \le P_{max}$, $0 \le M_0 < 2\pi$, and $0 < s \le s_{max}$.

1.3 One-planet model with a full Keplerian orbit

Recall that this model is

$$V_i \sim \mathsf{N}\left(C + \Delta V(t_i|\theta), \sigma_i^2 + s^2\right),$$

where

$$\theta = (K, P, e, \omega, M_0),$$

and

$$\Delta V(t|\theta) = K\left[\cos(\omega + T(t|\theta)) + e\cos\omega\right].$$

We take the seven parameters to be independent, and the marginal priors on the five earlier parameters C, K, P, M_0, and s to be as in the one-planet circular orbit model, given by equations (1) - (5). The marginal priors on the new parameters e, and ω are given by:

$$\pi_e(e) = \begin{cases} 1 & \text{for} \quad 0 \le e < 1 \\ 0 & \text{otherwise} \end{cases}, \tag{1.13}$$

and

$$\pi_\omega(\omega) = \begin{cases} \frac{1}{2\pi} & \text{for} \quad 0 \le \omega < 2\pi \\ 0 & \text{otherwise} \end{cases}. \tag{1.14}$$

So the complete prior is given by:

$$\pi_1(C, \theta, s) = \kappa_1 \cdot \frac{1}{K_0 + K} \cdot \frac{1}{P} \cdot \frac{1}{s_0 + s},$$

where

$$\kappa_1 = \frac{1}{C_{max} - C_{min}} \cdot \frac{1}{\log\left(1 + \frac{K_{max}}{K_0}\right)} \cdot \frac{1}{\log\left(\frac{P_{max}}{P_{min}}\right)} \cdot \left(\frac{1}{2\pi}\right)^2 \cdot \frac{1}{\log\left(1 + \frac{s_{max}}{s_0}\right)},$$

and where $C_{min} \leq C \leq C_{max}$, $0 < K \leq K_{max}$, $P_{min} \leq P \leq P_{max}$, $0 \leq e < 1$, $0 \leq M_0 < 2\pi$, $0 \leq \omega < 2\pi$, and $0 < s \leq s_{max}$.

Transformed parameter space

In many analyses, such as MCMC, it is convenient to work in a parameter space where the parameters are not so highly correlated as they are in the space defined by $(C, K, P, e, \omega, M_0, s)$. It is also convenient to have the posterior be as Gaussian as possible. For these reasons, one transformation of variables that is useful is the following:

$$\begin{aligned} x &\equiv e\cos\omega \\ y &\equiv e\sin\omega \\ z &\equiv (\omega + M_0) \mod 2\pi. \end{aligned}$$

And, for a more Gaussian posterior it is useful to work in $\dot{P} \equiv \log P$ rather than in P, in $\dot{K} \equiv \log K$ rather than in K, and in $\dot{s} \equiv \log s$ rather than in s.[4]

In this transformed space, the prior equivalent to the one given for the one-planet model with a full Keplerian orbit is given by:

$$\pi_1^*(C, \dot{K}, \dot{P}, x, y, z, \dot{s}) = \kappa_1^* \cdot \exp\dot{K} \cdot \frac{1}{\left(1 + \frac{\exp\dot{K}}{K_0}\right)} \cdot \frac{1}{\sqrt{x^2+y^2}} \cdot \exp\dot{s} \cdot \frac{1}{\left(1 + \frac{\exp\dot{s}}{s_0}\right)},$$

where

$$\kappa_1^* = \frac{1}{C_{max} - C_{min}} \cdot \frac{1}{\log\left(1 + \frac{K_{max}}{K_0}\right)} \cdot \frac{1}{K_0} \cdot \frac{1}{\log\left(\frac{P_{max}}{P_{min}}\right)} \cdot \left(\frac{1}{2\pi}\right)^2 \cdot \frac{1}{\log\left(1 + \frac{s_{max}}{s_0}\right)} \cdot \frac{1}{s_0},$$

[4] Note that $\dot{P}$, $\dot{K}$, and $\dot{s}$ are not derivatives with respect to time, they are simple log transformations.

and where $C_{min} \leq C \leq C_{max}$, $\dot{K} \leq \log(K_{max})$, $\log P_{min} \leq \dot{P} \leq \log P_{max}$, $x^2+y^2 < 1$, $0 \leq z < 2\pi$, and $\dot{s} \leq \log s_{max}$.

1.3.1 Values of the constants

The following choices of constants are in part based on physical realities (e.g., an orbit that's too small will result in the planet getting consumed by the star) and are in part arbitrary.

$$\begin{aligned} P_{min} &\equiv 1 \text{ day} \\ P_{max} &\equiv 1000 \text{ years} \\ K_0 &\equiv 1 \text{ ms}^{-1} \\ K_{max} &\equiv 2128 \text{ ms}^{-1} \\ C_{min} &\equiv -K_{max} \\ C_{max} &\equiv K_{max} \\ s_0 &\equiv 1 \text{ ms}^{-1} \\ s_{max} &\equiv K_{max} \end{aligned}$$

1.3.2 Priors on models

For Bayesian model selection, one needs priors on the different models. Since the proportion of stars that have so far been determined to have companion planets is approximately 12%, we will let the prior probability of the one-planet model (with full Keplerian orbit) be $\Pr(M1) = 0.12$ and the prior probability of the no-planet model be $\Pr(M0) = 0.88$. In reality, there is the possibility that a star hosts more than one planet; however, since we are here considering only the no-planet and the one-planet model, we will put a prior probability of zero on multiple-planet systems.

2

Why the Decision of Whether to Continue Observing a Star Is a Hard One

In this section we briefly discuss what makes the problem of planetary detection a difficult one statistically. In theory, it isn't really a hard question at all. Given a set of radial velocity data, we are interested in choosing between two parametric models. M0, the no-planet model, has only two parameters. M1, the one-planet model has seven parameters. Model selection may be done using frequentist hypothesis testing, or Bayesian model selection, or even using maximum likelihood ratios. These decision paradigms are well established and as well-suited to this model-selection problem as to many other signal-detection problems. So what is it about it that makes the problem hard?

There are two factors that, in combination, make the problem difficult to solve—and there is a third factor that doesn't help matters. The two primary factors are multimodality in the likelihood function of model M1, and multidimensionality in the M1 parameter space. Together, these make it hard (1) to determine a suitable statistic for hypothesis testing; (2) to find local maxima in the likelihood function;

(3) to integrate out the seven parameters in model M1 to give a Bayes factor.

The third factor that adds difficulty is that when the eccentricity of an orbit is not zero, then the radial velocity as a function of observation time is a transcendental function—it cannot be expressed in closed algebraic form and must be solved at every time t numerically. This makes it hard to solve equations analytically, such as one would like to do to find local maxima in the likelihood function by setting derivatives equal to zero. It also slows down numerical solutions considerably. If there are 25 radial velocity observations and one is searching for the MLE of the parameters in M1, for example, then no matter what numerical search algorithm one uses—and note that the multimodality means that a simple "climbing" algorithm won't suffice—at every step along the way one must numerically evaluate the radial velocity function at all 25 data points for the new parameter set.

2.1 Multimodality in the likelihood function under M1

Just how multimodal is the likelihood function in model M1? Recall that the model is:

$$V_i \sim \mathsf{N}\left(C + \Delta V(t_i|\theta), \sigma_i^2 + s^2\right), \text{ where} \tag{2.1}$$

$$\Delta V(t|\theta) = K\left[\cos(\omega + T(t)) + e\cos\omega\right], \tag{2.2}$$

and where $T(t)$ is a transcendental function of not only t but also e, P, and M_0.

Happily, the likelihood function is not multimodal in all parameters. In Chapter 4 of this dissertation there will be further discussion about the extent of the multimodality. For now we only remark that the likelihood function is actually conditionally normal in C and K (that is, conditional upon all the other parameters, the log likelihood function is quadratic in each of these parameters) and therefore unimodal. The likelihood function is also conditionally unimodal in ω, since the function ΔV is actually sinusoidal in ω. But the other parameters pose problems.

In some ways, the parameter P is the most problematic. Certainly, the likelihood function fluctuates more violently with changes in P than in any other parameter. A thought experiment will show why. Imagine the radial velocity data for a star spanning one year, 365 days. Now imagine some radial velocity function $\Delta V(t)$ superimposed over the data whose period is $P = 1$ day. The fit may be good or poor, it doesn't matter in this thought experiment. Now maintain the values of all the other parameters in ΔV while allowing P to increase slowly from 1 day to $365/364 \approx 1.0027$ days. For observations near $t = 0$, very little is happening. But for those observations near $t = 365$, they have passed through an entire period of ΔV. The predictions $\Delta V(t_i)$ have risen and fallen once like a cork on the surface of an ocean wave while the actual observations V_i sat still. Thus the residuals swelled and ebbed once, and not all in sync with one another.

Indeed, this is happening all along the interval [0,365]. As P makes the tiny change from 1 day to 1.0027 days, the residuals of every observation are swelling and ebbing, each to its own rhythm, such that the total sum of squared errors (which is, roughly speaking, a monotonic function of the log likelihood function) bobs up and down erratically, the total sum of a cacophony of wavy motions. The greater the range of the data is, the more erratic the log likelihood function behaves as P changes, since points further from $t = 0$ fluctuate more often when P increases. Likewise, the more data there are, the more erratic the log likelihood function behaves, since it is a sum of more individual moving components.

Figure 2.1 shows for the star 51 Peg the log likelihood function in P, conditional upon all the other parameters being held at some arbitrary values ($C = 0$, $K = 40$, $e = 0.3$, $\omega = 0$, $M_0 = 0$, $s = 2$). This should give some idea of what we're up against.

We see from the figure how sensitive the log likelihood function ℓ is to changes in P, particularly when P is small.

The log likelihood function is also multimodal in e, M_0, and s, but much less

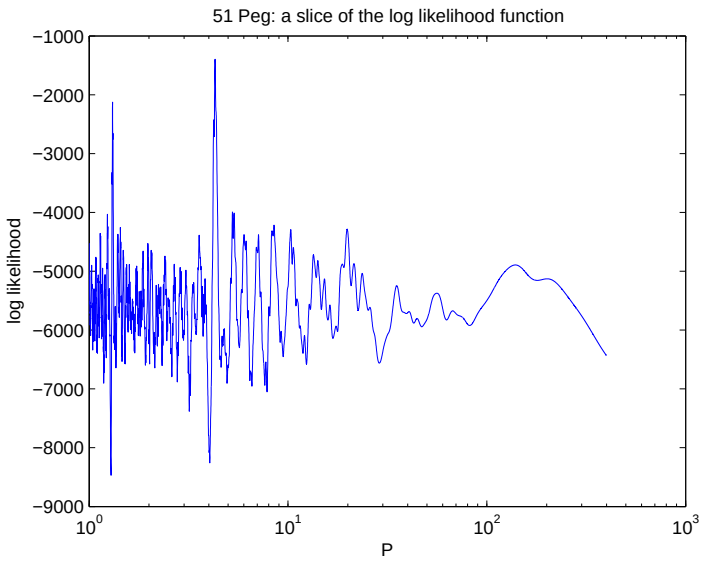

FIGURE 2.1: The star 51 Peg has $n = 111$ observations spanning 324 days, making the log likelihood function very sensitive to changes in P. Here is one cross-section of the log likelihood function $\ell(P)$, when the other parameters are held constant at some arbitrary values ($C = 0$, $K = 40$, $e = 0.3$, $\omega = 0$, $M_0 = 0$, $s = 2$).

violently so. Figure 2.2 shows plots of ℓ with respect to these three parameters, again holding the others constant at their earlier values (and now holding P constant at the arbitrary value of $P = 7$ days.) The data set this time is 10 simulated data from a star hosting no planets, rather than 51 Peg. (51 Peg happens not to illustrate the potential for multimodality in the other parameters so well as a smaller data set with less variability in V_i.)

We see from these three plots that, at least in this case, the log likelihood function may be multimodal in these three parameters—and somewhat unpredictable—but not so violently multimodal as $\ell(P)$.

This would not be a very serious problem if the conditional maximum of ℓ in one parameter was independent of the others. But all it takes is an illustration to show that this is not the case. Figure 2.3 shows the surface ℓ over P and M_0 for a simulated data set of 10 points from a non-planet-hosting star.

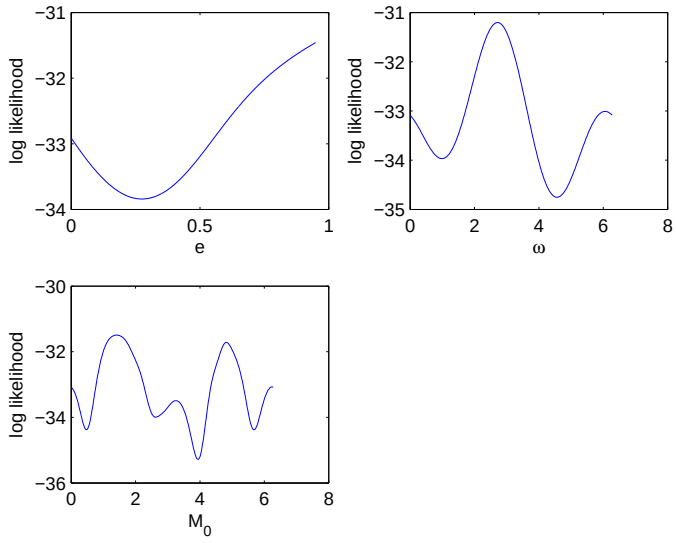

FIGURE 2.2: Ten simulated data points from a star with no orbiting planet yielded these cross-sections of the log likelihood function. Each is multimodal in its parameter, but not nearly so erratic as the earlier one in P.

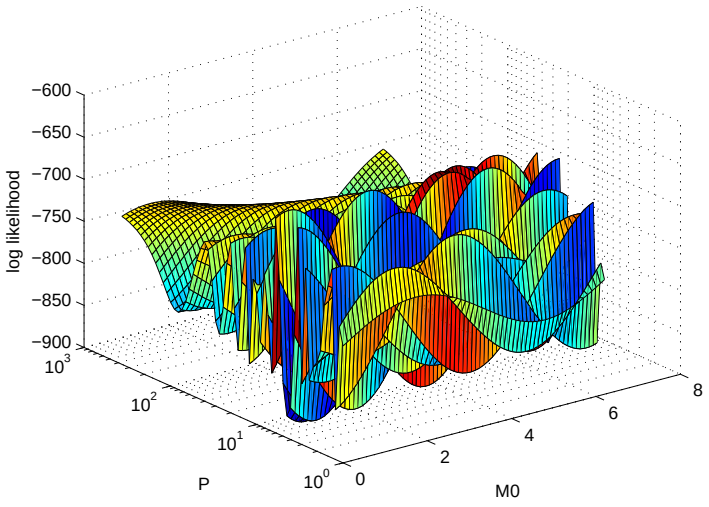

FIGURE 2.3: Ten simulated data points from a star with no orbiting planet yielded this two-dimensional cross-section of the log likelihood function. The grid in this surface plot is coarse so that distinct ridges will show up better; the actual surface has ridges between these ridges, and the local maxima vary in location from ridge to ridge.

We see in this figure that local maxima in the log likelihood function do not occur in predictable places, and certainly not at the same M_0 for every P.

Similar issues arise with e and s. Thus, when either searching for maxima in the likelihood function or (worse) trying to integrate the likelihood surface, we must deal with the fact that there are local maxima scattered all over the place.

2.2 Multidimensionality, and a jump-start

Multimodality would not necessarily be a big problem in one dimension or in two. But in four dimensions, it's a problem. The so-called "curse of dimensionality" means that with every new parameter, any numerical technique applied to that multidimensional surface becomes an order of magnitude more difficult. In this case, it means that trying to find local maxima in this four-dimensional space (we don't count C, K, and ω, since the likelihood function is conditionally unimodal in those parameters) is fairly tricky.

If the likelihood surface had local modes in *completely* unpredictable places, as if it were a mixture of randomly scattered unimodal surfaces, then finding all the modes would require searching literally everywhere in the parameter space, at least down to a granularity as small as the greatest possible separation between two local maxima. Such a search would be exceptionally cumbersome given the great sensitivity that ℓ has with respect to P. But happily, the local modes are not utterly unpredictable. I wrote earlier that P was, *in some ways*, the most problematic parameter. In one way, it's actually a fairly friendly parameter. That is because the *periodogram*, which will be discussed in Chapter 3, is a tool that may be used to suggest good candidates for P. And as we shall eventually see, the periodogram seems to do a remarkably good job of finding those candidate P's, even though it is designed to detect sinusoidal behavior, and the radial velocity function need not be sinusoidal.

Thus, with the periodogram, we will not be required to search for MLE's as

we would needles in a haystack. That will come later. Here, we simply remark that although the combination of multimodality and multidimensionality creates problems, they are not unsurmountable.

2.3 Computational speed

We saw in Chapter 1 that determining the mean radial velocity of a star at time t given all the parameters in the model requires at one point solving a transcendental equation. Throughout this paper, we do that calculation using a quadratic version of Newton's method, which tends to converge to the solution with ten decimal place accuracy in just two or three iterations. So it is not so bad as it might have been. But even so, this must be done for every data point. And one of things we want to look at in this project is whether large data sets (up to 400 observations) may reveal small planets if one is patient enough. A data set with 400 observations requires 400 solutions of the transcendental equation, not to mention all the other computations involved in putting them together, just to compute the log likelihood at one single parameter vector. If you wanted to, say, integrate out the parameter vector altogether using an importance sampler, then this would have to be repeated hundreds or thousands of times, once for each parameter vector that you had sampled from your importance sampler. And then if you wanted to see what the distribution of that integral looked like using simulations, then you would have to repeat *that* process thousands of times. The transcendental nature of the function ΔV might have been of minor concern under other circumstances, but the task at hand makes it quite a nuisance.

2.4 What follows in this paper

Recent papers that have solved problems related to the planetary detection question have often assumed that either $e = 0$ [5], which reduces the number of troublesome

parameters by two (both e and M_0 are eliminated); or that $s = 0$ [4], which reduces the number of troublesome parameters by one (and which greatly simplifies the log likelihood function's analytical form). In this paper we do not make either of these assumptions,[1] and we use simulations rather than analytic calculation to study distributions of sample statistics.

We have seen in Chapter 1 what the problem is ("O Star, art Thou alone?"); and in Chapter 2, why the problem is hard (an axis of computational difficulty: multimodality, multidimensionality, and transcendental equations). The remaining chapters will be devoted to studying three different one-dimensional sample statistics that may be used to address the model selection problem. They are the Maximal Periodogram Spike (or MPS), the Maximum Likelihood Ratio (or MLR) and the Bayes Factor Estimate (or BFE). These are discussed briefly below.

- The Maximal Periodogram Spike, or MPS. This is the statistic used by astronomers to determine whether a set of radial velocity data suggest the presence of an exoplanet. Among its virtues are that it is familiar to those in the field, it is easy to compute, and in computing it, it also suggests plausible values for one of the one-planet model's parameters, the orbital period. Among its deficits are that it is designed to detect sinusoidal periodicity, not necessarily the odd waves exhibited by the radial velocities of eccentric orbits; and that when used for model selection, the criterion typically used to reject the no-planet model is a goodness-of-fit criterion whose alternative is "anything but noise", rather than the rigid structure of the Keplerian model. Additionally, the threshold for planetary detection is based on the hypothetical distribution it would have if the uncertainty in radial velocity measurements were constant;

[1] Exception:when we consider the Bayes factor as a planet detection tool, we restrict our analysis to circular orbits, due to the great difficulty of integrating out the parameters of the full seven-parameter model

this is not the case. The efficiency of its performance is not well understood.

- The Maximum Likelihood Ratio, or MLR. This statistic has the virtue of giving both of the models under consideration the chance to look their best. It essentially estimates how much better the one-planet model is capable of doing compared to the no-planet model when tasked with explaining the observed data. One of its deficits is that it is somewhat more difficult to compute than the MPS, though probably not so much so that it would be prohibitive to use in practice. Like the MPS, its performance efficiency in the context of exoplanet detection is not well-studied.

- The Bayes Factor Estimate, or BFE. The Bayes Factor comparing these two models has the virtue of reflecting the true, overall relative explanatory power of the two models, when the parameters are integrated out. And it is easy to interpret, since its meaning is universal, not local to these exoplanet models. However, in the context of exoplanets it suffers from two problems that we will have to address. First, in order to perform at its best, it requires realistic prior distributions on the parameters in the one-planet model. But unfortunately, sampling bias necessarily prohibits access to those priors. For example, we know that very small planets must be underrepresented in our present surveys, since the perturbations they induce on their host stars are so very hard for us to separate from the much louder noise of stellar jitter and measurement errors. Similarly, very eccentric orbits are probably somewhat underrepresented, since their orientation may be such that they only exhibit short bursts of radial motion between long spells of nearly undetectable lateral motion. And although we will not address the issue at all in this paper, long-period planets are selected against, since our present search has not been going on long enough for us to witness even one complete period of anything that takes more than twelve or

so years to go around its host star. (Jupiter takes twelve years to go around our Sun, Saturn nearly thirty.)

The second problem that afflicts the Bayes Factor in this context is that computing it for a given data set requires integrating out the seven parameters of the one-planet model from a posterior density that is multimodal, multidimensional, and transcendental everywhere. We discuss in Chapters 2 and B why this combination of characteristics makes integration so extremely vexing. And so as not to get your hopes up too high, let me reveal right away that you will not find in Chapter 5 a nice solution to the problem. Rather, I will present an algorithm intended to estimate the Bayes Factor, which is why I call this statistic the Bayes Factor Estimate. I probably should call it *a* Bayes Factor Estimate or *one* Bayes Factor Estimate, since of course there are many others. But in the present paper it is the only one I will use, so I will call it the Bayes Factor Estimate or the BFE to distinguish it from the MPS and the MLR. Because the BFE may not be the true Bayes Factor, we will study it as a sample statistic, just as by that time we shall have already done with the MPS and the MLR.

So now let's return to the first of these three statistics, the Maximal Periodogram Spike—henceforth, the MPS. In the following chapter I'll first introduce it to any reader to whom it is not already familiar, and then we'll look at how well it does at finding planets.

3

The Maximal Periodogram Spike Height (MPS)

3.1 Introduction to the Lomb-Scargle Periodogram

The "classical periodogram" emerges from the discrete Fourier Transform, a basic tool of spectral analysis, and is defined in the following way. For a set of temporal data $(t_j, X_j), j = 1 \ldots N$, let

$$\begin{aligned} P_X(\omega) &= \frac{1}{N}\left|\sum_{j=1}^{N} X_j \exp(-i\omega t_j)\right|^2 \\ &= \frac{1}{N}\left[\left(\sum_j X_j \cos(\omega t_j)\right)^2 + \left(\sum_j X_j \sin(\omega t_j)\right)^2\right] \end{aligned}$$

If the X_j's are in fact sinusoidal with respect to the t_j's, with frequency ω_0, then at or near ω_0, the X_j's and the $\exp(-i\omega t_j)$'s will be in phase—larger at the same time and smaller at the same time, and contributing greatly to the given sum. At other values of ω, the X_j's will have an arbitrary relationship to the $\exp(-i\omega t_j)$'s, and so will contribute comparatively little to the sum. As a result, the classical periodogram when plotted over many candidate frequencies ω will show a "spike"

at any frequency which is a good candidate frequency for the data. The classical periodogram also has the property that if the X_j's are in fact mere Gaussian noise with constant variance, then the distribution of P_X is exponential with known mean (a function of the noise variance) if the t_i's are evenly spaced. The distribution in the case of unevenly spaced data is considerably more complicated. To deal with that case, Jeffrey Scargle [10] suggested the following modification.

Let

$$P_X(\omega) = \frac{1}{2}\left\{ \frac{\left[\sum_j X_j \cos\omega(t_j-\tau)\right]^2}{\sum_j \cos^2\omega(t_j-\tau)} + \frac{\left[\sum_j X_j \sin\omega(t_j-\tau)\right]^2}{\sum_j \sin^2\omega(t_j-\tau)} \right\},$$

where τ is defined by

$$\tan(2\pi\tau) = \frac{\sum_j \sin 2\omega t_j}{\sum_j \cos 2\omega t_j}$$

This new function has the desired property of being exponentially distributed when the X_i's are Gaussian noise, whether the t_i's are evenly spaced in time or not (again, with known mean related to the noise variance). Because the distribution of $P_X(\omega)$ is known in the case of Gaussian noise, the periodogram can be used—indeed *is* used—to perform a classical hypothesis test to see whether a spike in the periodogram truly indicates an orbiting planet or not. However, the behavior of P_X when the data are periodic but not sinusoidal has not been extensively studied. Moreover, the noise in both of our models does not have constant variance from one observation to the next. For these two reasons, we will not in this paper take the distribution of P_X to be known, but will estimate it using simulations.

Both this periodogram and the classical periodogram are time-translation invariant. Developed independently by Nicholas Lomb, the function is now generally called the Lomb-Scargle periodogram. I will only discuss the Lomb-Scargle periodogram in

this paper, not the classical one or any other variation, and from here on I will call it simply the *periodogram*.

Because planetary orbits are more easily and more often considered in terms of the period of their orbits rather than their frequencies, I will henceforth use the following version of the periodogram, a function of periods rather than frequencies:

$$\Upsilon(P) \equiv P_X\left(\frac{2\pi}{P}\right)$$

Readers already familiar with the periodogram may feel free to skip ahead to section A.1. What follows are some simulations and some real data meant to give to readers unfamiliar with it a "feel" for it.

We now will look at some examples of the behavior of Υ.

3.1.1 51 Peg

The first exoplanet ever discovered surprised the world by how very different it was from the planets in our own Solar system. Half as big as Jupiter, it whirls around its host star once every 4 days and 5 hours at a distance 20 times closer than we are to the Sun. Its host star is approximately 50 light-years away from the earth (that makes it a close neighbor) and is called HD217014 in the Henry Draper Catalog of stars[1] or more commonly 51 Pegasi, as it is the 51$^{\text{st}}$ brightest star in the region of the sky called the constellation Pegasi. More familiarly still, it is called simply 51 Peg.

Figure 3.1 shows the radial velocity data for 51 Peg along with error bars showing measurement errors, drawn ±one standard deviation from the estimated radial

[1] 225,300 stars are classified in this catalog, which was first published in 1918. Over 10,00 of the classifications were done by an astronomer named Willamina Fleming, who made many other contributions to astronomy and later became the first American woman to be made an honorary member of the Royal Astronomical Society of London. Yet Fleming had no formal training in astronomy, and was originally the housemaid of astronomer Edward Pickering, who recognized in her a keen intellect and gave her the opportunity to work for the Harvard College Observatory, where in 1899 she was named Curator of Astronomical Photographs.

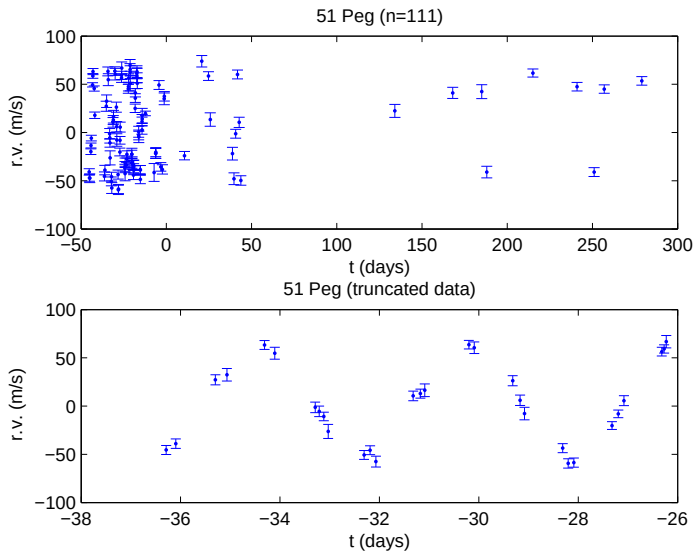

FIGURE 3.1: The upper plot shows all $n = 111$ published radial velocity observations, along with their measurement errors. The lower plot shows only observations 12–41, selected because they clearly reveal the periodicity in the data. Note that the nine latest observations made all have radial velocities near the maximum or minimum of the radial velocity function. That is because such observations serve better to reduce the uncertainty in the orbital parameters.

velocities.

Looking at the lower plot in Figure 3.1, it is clear that the orbital period is approximately four days. The periodogram should therefore show a spike corresponding to that period. In Figure 3.2 we see the periodogram for the given data, with the candidate periods shown on a logarithmic scale.

The periodogram shows two spikes: one at 4.23 days, the other at 1.31 days. We might notice that on the vertical scale, the difference between these two spikes' heights is over 2000, but without yet having any understanding of the meaning of $\Upsilon(P)$, it's unclear whether that difference is significant. Let's plot the data once again, but this time over one period only. We'll make a plot in which the time of the observations are shown modulo 4.23 days, and another plot in which the times of the observations are shown modulo 1.31 days. If either one is correct, perhaps

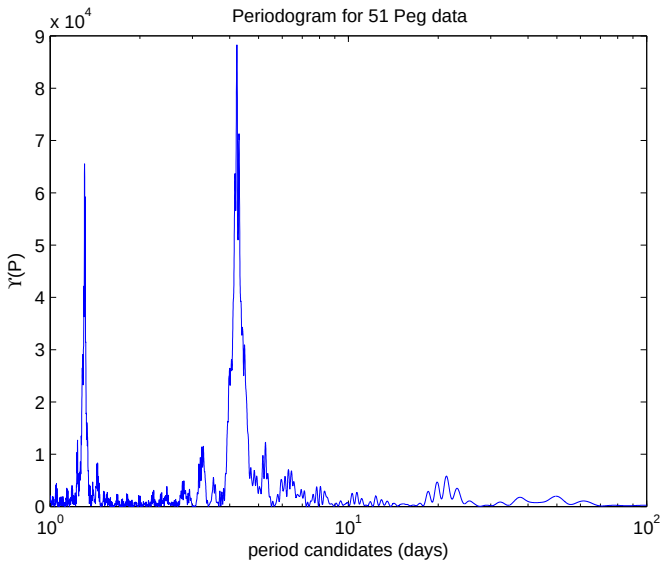

FIGURE 3.2: The periodogram does indeed show a spike at around 4.23 days, as expected; that's the taller of the two spikes. But it also shows a spike at around 1.31 days, suggesting a second plausible candidate period for the data.

such a plot will reveal it. Figure 3.3 shows these two plots, along with one showing a deliberately poor choice for the period.

3.1.2 HD73526

Let's take at look at the radial velocity data for HD73526, first published in 2002 [3]. Figure 3.4 shows the data with errorbars.

The variability in radial velocities is on the order of 200 m s^{-1} and is therefore much too great to be due to measurement errors and stellar jitter. But if there is periodic behavior in the data, this plot does not readily reveal it. So let's take a look at a periodogram for these data, which is shown in Figure 3.5.

We see in the figure that there are three spikes in the periodogram that seem to stand out, located at periods $P = 188$ days, $P = 386$ days, and $P = 128$ days.[2] On the vertical scale, the difference between the height of the tallest and shortest of these

[2] There may be others worth considering as well, but for our illustrative purposes, the these three will suffice.

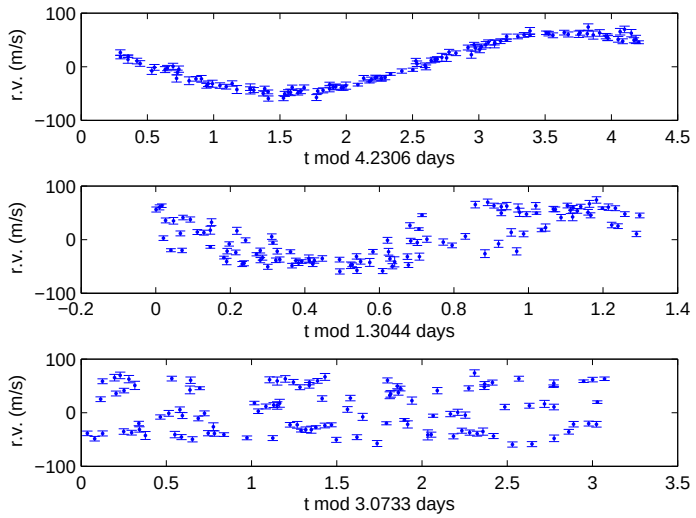

FIGURE 3.3: It is clear that the period of 4.23 days is correct, and the period of 1.31 days is not—though it is equally clear why there is a spike in the periodogram at $P = 1.31$ days; the data are "sort of" sinusoidal at that period. For illustrative purposes, the bottom plot is included, which shows the data modulo 3.07 days—this is the value that produces the *lowest* point in the periodogram.

three is over 11,000, and in our example with 51 Peg, we saw that a difference of just 2000 occurred between $\Upsilon(P)$ for two periods of considerably different plausibility. But we mustn't rush to interpret $\Upsilon(P)$ the same way every time. This time the magnitudes of the radial velocity differences are much greater, and the sample size is much smaller.

Let's do as we did with 51 Peg and look at the data with the observation times shown first modulo 188 days, then modulo 386 days, and then modulo 128 days. Figure 3.6 shows these three plots.

At first we are disappointed: these plots are not nearly so obviously indicative of a planet as the one at the top of Figure 3.3. But that's in part because there are fewer data to fill in all the gaps of the orbital cycle, and also because 51 Peg's orbit is nearly circular, and this one apparently is not. Perhaps the plots will look

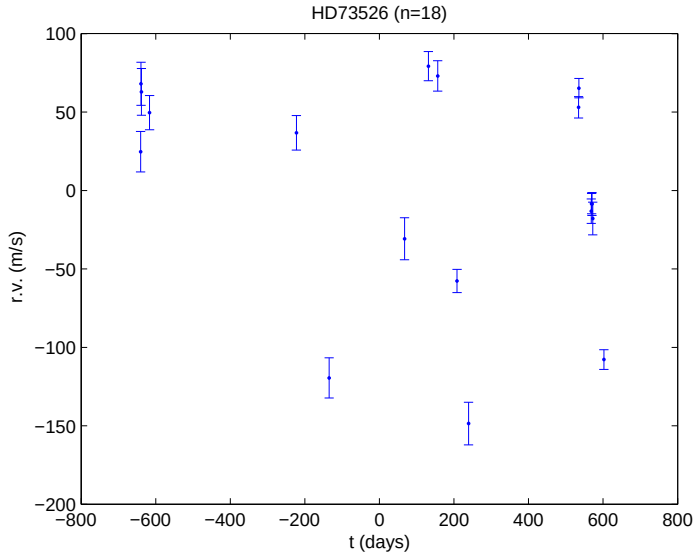

FIGURE 3.4: HD73536 data. While the variations among radial velocities are too great to explain by error measurements alone—or by stellar jitter either, which seldom exceeds 10 m s^{-1}—it is not obvious that the data are periodic at all, much less what their period might be.

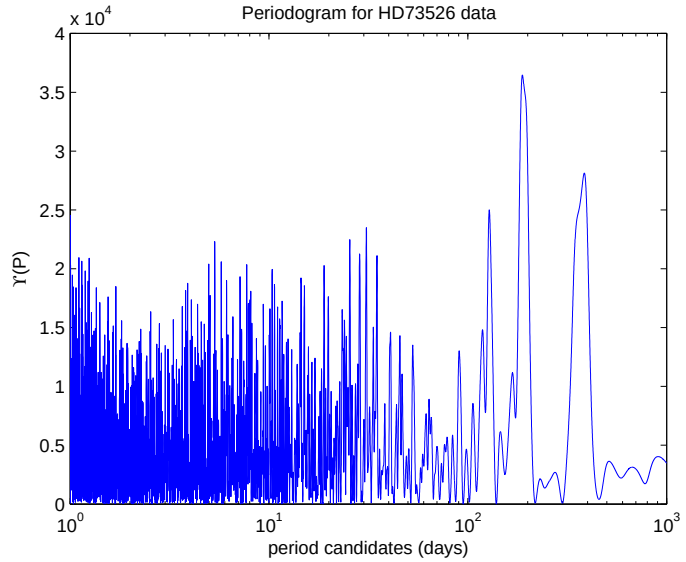

FIGURE 3.5: HD73526. The three tallest spikes in this periodogram occur at periods $P = 188$ days, $P = 386$ days, and $P = 128$ days.

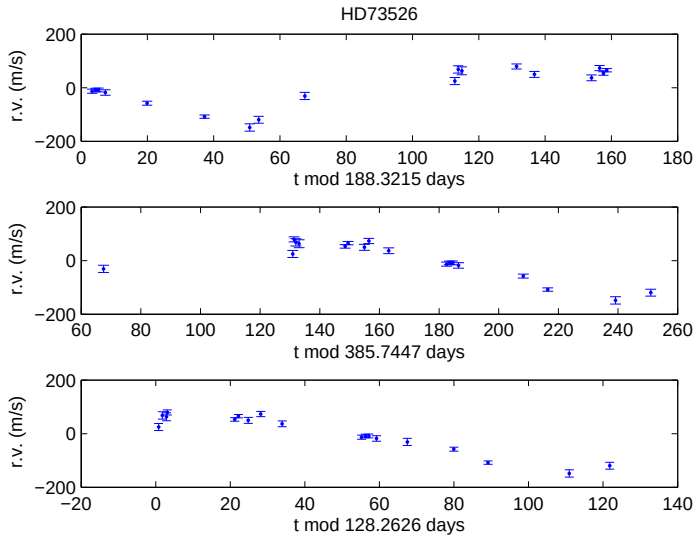

FIGURE 3.6: All of these three periods seem plausible when one recalls that the star's orbit need not be circular, and therefore the radial velocity plot need not be perfectly sinusoidal.

better when each is shown against a radial velocity function selected to match its undulations.[3] These are shown in Figure 3.7.

These three plots suggest that all three candidate periods are plausible. In fact, the original publication in which this planet was unveiled gave the period as 188 days. (That corresponds to the tallest spike in the periodogram.) Philip Gregory [11] later did a Bayesian analysis on the orbital parameters and concluded that the 386-day period was more probable. Now, after gathering further data, researchers believe that there are two planets orbiting HD73526, one having period 188 days and the other having period 378 days.[4] [12]

[3] I'm here using the local maximum likelihood estimator of the parameters corresponding to each period.

[4] Although the periodogram spike occurs at 386 days, the maximum likelihood estimate of the period local to this spike is closer to 374.1 days, and 378 days has its justification as well.

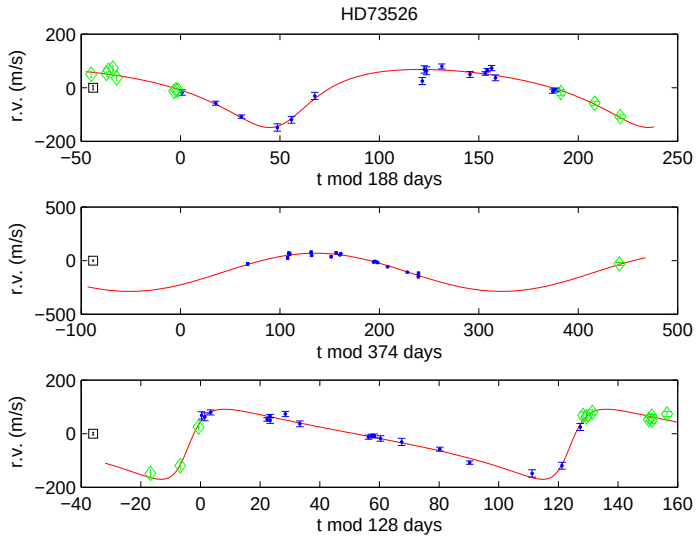

FIGURE 3.7: The three plots show the same data under three different moduli. The actual data are shown as blue dots. Repetitions of the data are shown in green diamonds on either side, to create the sense of a continuum. Inside a black square at the left of each graph is an estimate of the stellar jitter magnitude shown as a vertical line segment—here, so tiny relative to the amplitude of the waves that it is only barely visible. Note that although a spike in the periodogram occurred at $P = 386$ days, the corresponding plot here is shown now with time modulo 374 days. That is because although the periodogram suggested $P = 386$ days, the local maximum likelihood estimate of the period corresponding to that spike occurs at $P = 374.1$ days. The other two plots are also shown under slightly different moduli from those in Figure 3.6, the old ones coming directly from the periodogram, the new ones "tweaked" to correspond to the local MLE.

3.1.3 Noise

Recall that our present purpose is simply to familiarize readers with the periodogram and the MPS. We've probably achieved that at this point, but it may be helpful before looking at the MPS statistic in more detail to consider what the periodogram may look like when in fact there is no planet present—when the data result from nothing but noise due to stellar jitter and measurement errors. Because very small planets (and sometimes planets with highly eccentric orbits) contribute relatively little to the variations in radial velocities relative to noise sources, this section should also be

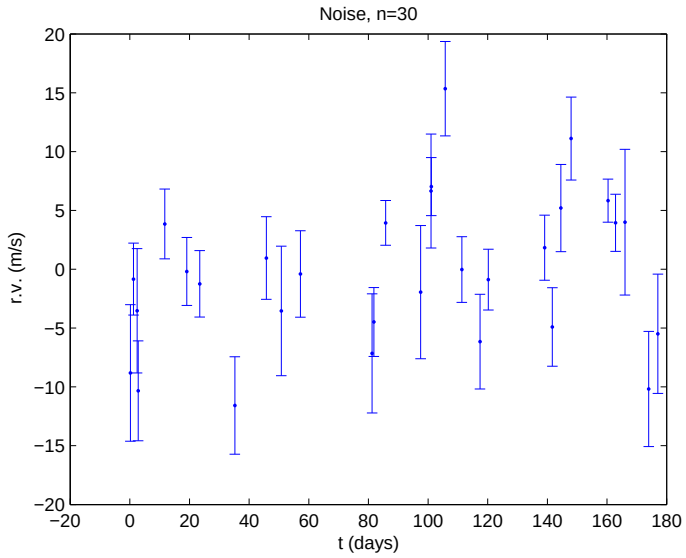

FIGURE 3.8: These 30 simulated radial velocity measurements look like mere noise—and they are exactly that—but they could just as easily result from a very small planet whose perturbations on its host star are drowned out by stellar jitter and measurement errors. The Earth, for example, causes velocity changes in the Sun no greater than 0.1 m s^{-1}.

taken as indicative of what the periodogram may look like in those cases.

We'll use a hypothetical star having stellar jitter of $s = 3$ m s^{-1}, which is typical of many stars. The measurement errors σ_i we will generate from a $\mathsf{Gam}(8, 0.5)$ distribution, independently for each observation. (That is a reasonable approximation of the distribution of actual measurement errors among all published radial velocities, though they have grown somewhat smaller in recent years.) Then the actual radial velocity measurement for the i^{th} observation will be normally distributed with mean 0 and variance $s^2 + \sigma_i^2$. We'll here use $n = 30$ observations over a time window of 180 days, both of which are somewhat arbitrary. Figure 3.8 shows one such simulated set of data.

Now let's do as we did for the real data and examine what the periodogram looks like (Figure 3.9) and then look at what the data look like when plotted with the time

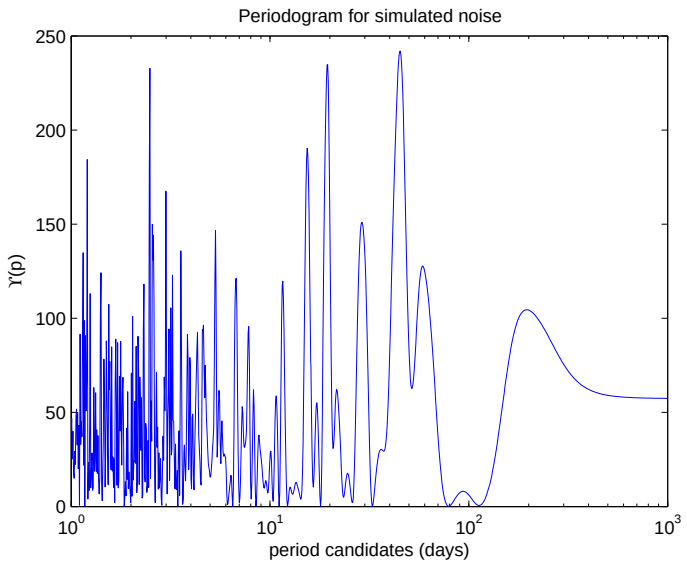

FIGURE 3.9: The tallest spike in the periodogram occurs at $P = 45.29$ days, though in this case it is misleading, since the data are in fact not periodic, but mere noise.

under the modulus suggested by the tallest spike in the periodogram (Figure 3.10).

In Figure 3.10 we see that, at least for this single simulated data set, the noise looks remarkably like a single period in a radial velocity function. It could be that this is a fluke or it could be that for this sample size, and when shown according the period suggested by the MPS, noise that looks this good is typical. In order to make that determination, we'll need to run many simulations, and find a way to summarize what we find. Running many simulations is (in this case) not very hard. In the next section we'll discuss what is the best way to summarize the information in each simulation.

3.2 The distribution of the MPS under M0

There are 11 sample sizes we will consider throughout this dissertation, and they're spaced evenly on a logarithmic scale: 4, 6, 10, 16, 25, 40, 63, 100, 159, 252, and 400. For each of those sample sizes, we will simulate data from a star that has no planet.

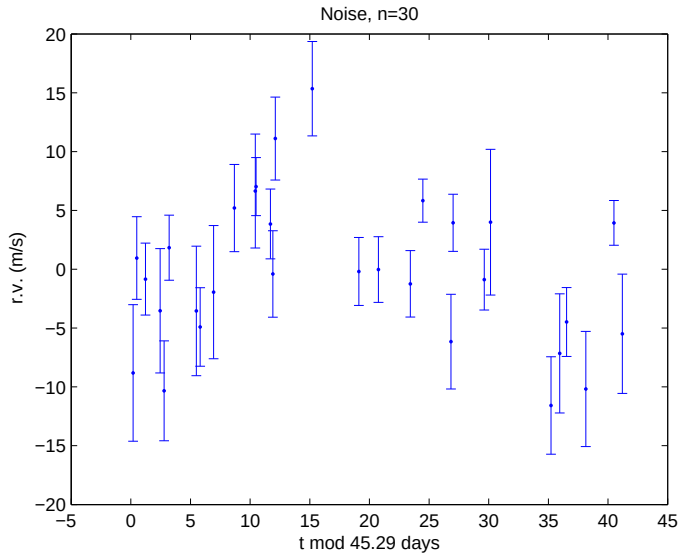

FIGURE 3.10: When shown "in their best light", the data look plausibly like a single period in a Keplerian radial velocity function. It remains to be seen whether this is a fluke or, for this sample size, noise will routinely look this good.

We will include stellar jitter by simulating for each star a value of $s \sim \mathsf{Gam}(3, 1)$, and we will include measurement errors by simulating for each observation a measurement error magnitude of $\sigma_i \sim \mathsf{Gam}(8, 0.5)$. (These two distributions reasonably approximate the actual distributions of stellar jitters and measurement errors recorded thus far.)

The observation times will be independent and uniformly distributed over a period of 42 days, an arbitrary choice. The radial velocities will be simulated from a Gaussian distribution having mean 0 and variance $s^2 + \sigma_i^2$. Then for each simulated data set, a periodogram will be computed and the height of the tallest spike recorded—i.e., $\Upsilon(P)$ will be recorded, not the P where it occurs. Although it is the height of tallest spike we are keeping track of, we'll simply call it the MPS for short, for "Maximal Periodogram Spike".

In Figure 3.11 we see these 11 distributions, one for each sample size under consideration. Interestingly, the distribution appears to change very little with increasing

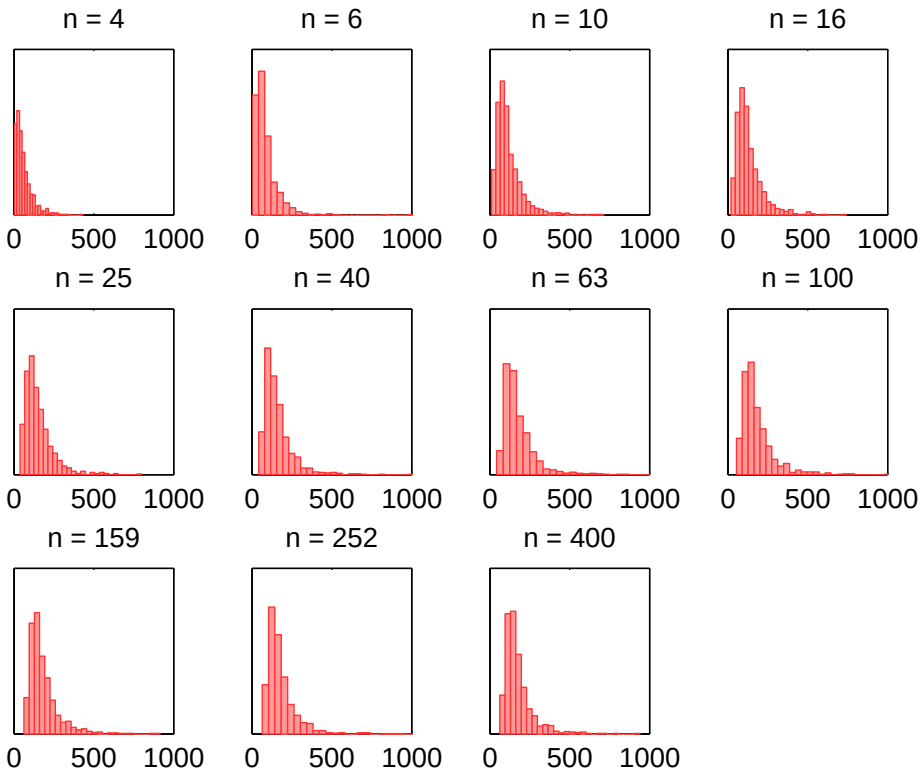

FIGURE 3.11: Beyond a fairly small sample size, the distribution of the MPS appears to change very little with increasing n.

n, beyond a sample size of about 10 or 16.

Because we will need it later when we begin considering the likelihood ratio as a means of evaluating the usefulness of a statistic, let us go ahead and use the non-parametric kernel density estimator (with a Gaussian kernel) to estimate these densities. The bandwidth of the kernel function will be chosen automatically based on the number of simulations (1000), using a formula internal to the software (Matlab). Figure 3.12 shows the eleven distributions again with the kernel density estimates superimposed. Figure 3.13 shows the cumulative distribution functions of the kernel density estimates superimposed over the ogives of the simulated values of the MPS. For no sample size are these at all distinguishable, indicating that the density

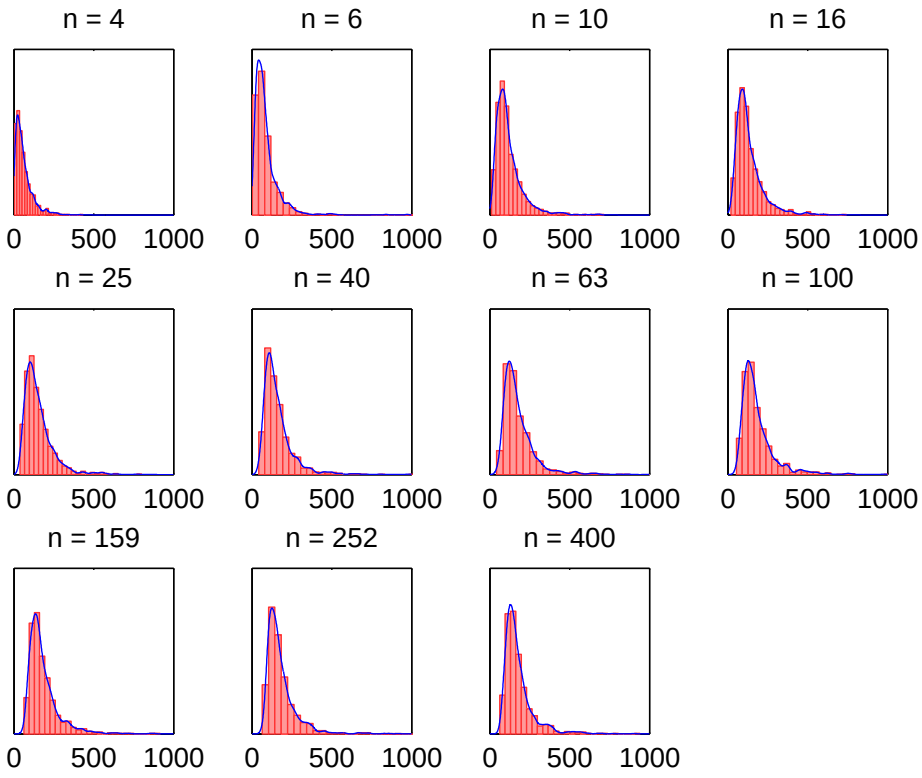

FIGURE 3.12: The kernel density estimates appear to do a good job estimating the densities of the MPS in all cases.

estimation is reasonably good.

3.3 Simulating the MPS under model M1

Now we will repeat the simulation process, only this time we will simulate data from a star having a planet. Throughout this dissertation we will use a period of $P = 7$ days for all of our planets. The period *does* affect (or, more precisely, is affected by) the mass of the orbiting planet; it also indicates the orbital radius of the planet. So it is hardly an uninteresting parameter! But the *detectability* of a planet is not affected by the period, assuming that observations have been made over at least several periods, because changing the period only changes the *scale* of the time

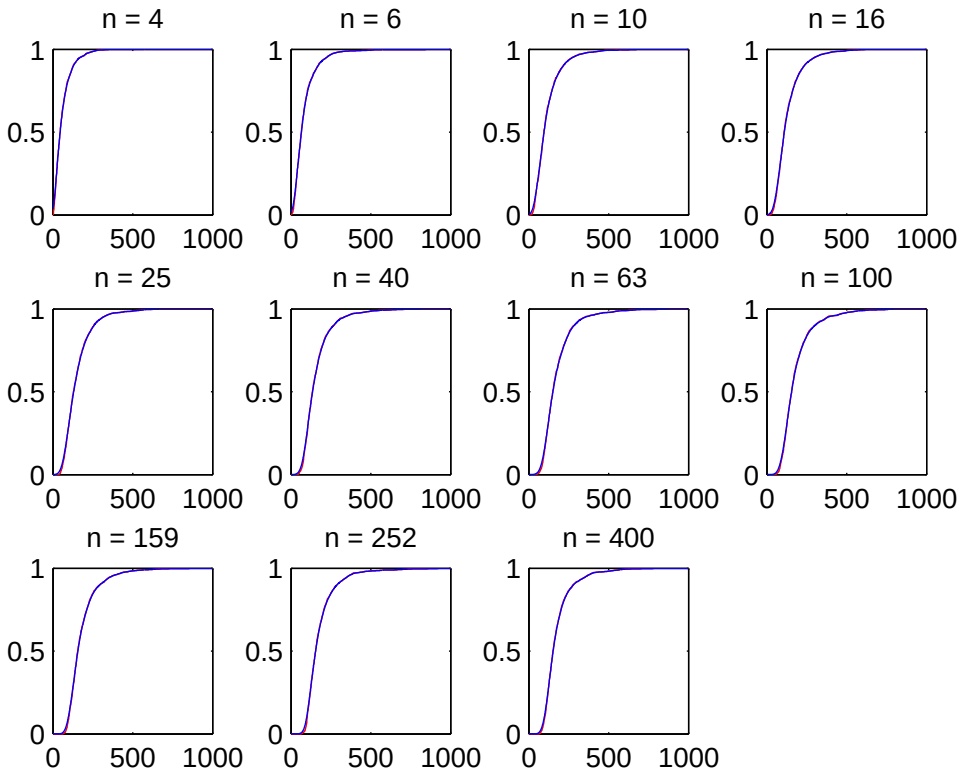

FIGURE 3.13: The cdf of the kernel density estimate is nowhere distinguishable from the ogive of the simulated values of the MPS.

variable. So the choice of $P = 7$ is arbitrary, but the conclusions we draw about planetary detectability apply to any period, supposing always that observations are made over at least several periods.[5]

So that observations may be made from at least six periods, we will draw observations at random times from a timespan of 42 days. This will also be done consistently throughout the dissertation.

The stellar jitter and measurement errors will be drawn from the same distributions as in the no-planet model, mimicking their actual distributions among stars

[5] The special difficulties involved in detecting long-period planets will not be discussed in this dissertation.

observed so far.

We will consistently use $C = 0$ in our simulations, the "background velocity" of a star, which has nothing to do with variations in radial velocity measurements. (This was also done in the no-planet model, when we used 0 as the mean of the normal distribution of each V_i.)

We will consider the following six types of planets. A brief explanation of each one's meaning is given.

1. $K = 50$ m s^{-1}, $e = 0$. The first exoplanet discovered was orbiting 51 Peg, whose semiamplitude is $K = 56$ m s^{-1}. This was a fairly big planet, having half the mass of Jupiter, and we now know that there are many similar planets out there which we call "hot Jupiters," because they're massive and close to their host stars. Although our planet will be farther from its host, having $P = 7$, and therefore it will be even more massive, its detectability over several periods should be about that of 51 Peg, which was very readily detected. Since we are focusing on detectability, not orbital parameters, then we need not look at K's much larger than 50 m s^{-1}. Such planets are screaming to be detected.

2. $K = 20$ m s^{-1}, $e = 0$. The long-period planet orbiting HD117207 induces a semiamplitude of $K = 26.6$ m s^{-1} on its host star. Because it is far away, we know it must be massive. (Its minimum mass is 1.88 times the mass of Jupiter.) But similarly detectable would be a planet with the same K and a period of 30 or 3 days, whose minimal masses would be (assuming a sun-sized host star) approximately 0.46 M_J and 0.21 M_J, respectively.

3. $K = 7$ m s^{-1}, $e = 0$. One of the smallest planets yet discovered that is not part of a larger planetary system is the companion planet to the star called HD219828, whose orbiting planet was discovered in 2007. This planet has a minimal mass about one-twentieth that of Jupiter, which is approximately

equal to the mass of Neptune, or about 20 times that of the earth. Its orbit is nearly circular, its period is $P = 3.8$ days and HD219828's radial velocity function's semiamplitude is $K = 7$ m s^{-1}, only slightly larger than the typical stellar jitter. The proximity of the star to its host is the reason that it could be so small and still detectable.

4. $K = 3$ m s^{-1}, $e = 0$. We will continue to use $P = 7$ days. If such a planet were found, its mass would be about 0.03 times the mass of Jupiter, or about 7 times the mass of the Earth—smaller than any stand-alone planet yet discovered (although slightly smaller planets have been found in planetary systems).

5. $K = 7$ m s^{-1}, $e = 0.3$. We anticipate seeing that when $K = 7$ m s^{-1} and $e = 0$, planets are fairly hard to detect. The nature of K is such that when e is greater than 0, even if K stays the same, the possible spread in radial velocities is greater, potentially making planets more easily detected. But there are two drawbacks to eccentricities greater than zero: first, there is some chance that the orbit's orientation will be such that long spells occur with relatively little variation in the radial velocities, and that only lucky observations during occasional "bursts" will pick up on planetary presence. And second, the periodogram is not specially designed to detect eccentric Keplerian patterns, but rather sinusoidal patterns. So here we will see whether an eccentricity of 0.3 makes a planet harder or easier to detect. Eccentricities of around 0.3 are not uncommon.

6. $K = 7$ m s^{-1}, $e = 0.6$. Relatively few discovered planets orbit stars whose orbital eccentricities exceed 0.6. Could that be because the periodogram is not a great tool for discovering them? Or could it be that there are relatively few planets with highly eccentric orbits, given that there are physical processes

that reduce eccentricity over time (such as tidal dissipation and planet-planet interactions)? Interest in understanding these effects makes it more important to determine the extent to which the observed eccentricity distribution reflects methodology effects.

We will also consider one more scenario. If we assume that the planets discovered so far approximate the actual distribution of planets in the galaxy (an unrealistic assumption due to known selection bias, but we have no better distribution to go by); and if we further assume that orbital parameters are more or less independent of one another (also somewhat unrealistic, but not terribly so given the present catalog of discovered planets); then we may draw stars from a random distribution with the following characteristics:

- We will take $\log K$ to have a normal distribution with $\mu = 4$ and $\sigma = 1.2$.
- 10% of the stars hosting planets will have $e = 0$. Among the rest, we will take the distribution of e to be triangular, with a linearly decreasing density from 0 to 1.
- We will take the distribution of $\log P$to be a mixture of two normals, one-third of the mass on one with $\mu_1 = 1.5$ and $\sigma_1 = 1$, the rest on one with $\mu_2 = 6.5$ and $\sigma_2 = 1.5$. This roughly approximates the periods of planets found so far, the smaller periods being mostly the hot Jupiters mentioned earlier, and the latter being more Earth-like in their distances from their host stars. The mixture is one-third on the smaller periods and two-thirds on the larger periods. (However, although we observe that here, we will not use a distribution of P's in our simulations. We will continue to use the constant and arbitrary $P = 7$ days, noting once again that so long as observations are made over a time period of at least several periods (here, we draw from a time span spanning

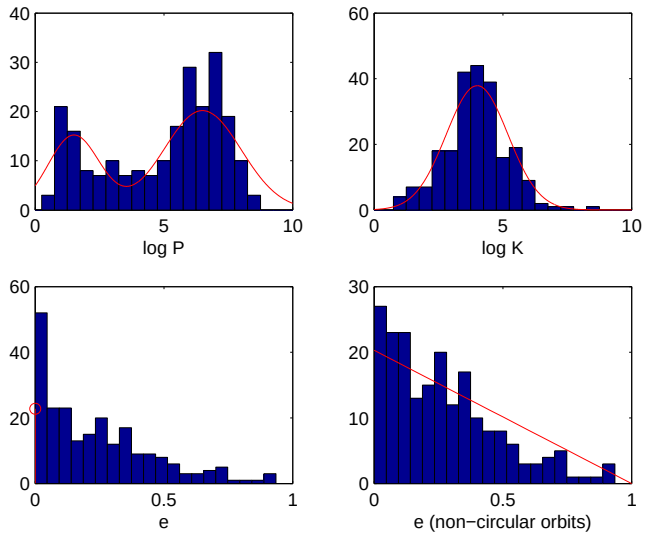

FIGURE 3.14: The distribution of planetary characteristics found so far, plotted against the (rescaled) distributions we will use to simulate stars in the last section of this chapter. Logarithms are natural logarithms.

six periods) then the actual period has no effect on the heights of spikes in the periodogram.)

Figure 3.14 shows histograms of these orbital characteristics among the first 228 planets found, and superimposed upon them are the distributions we will use to simulate neighboring stars in our galaxy.

In the last three cases, when the eccentricity is not zero, we will require values for the parameters ω and M_0. There is no reason to believe that the orientations of star systems is not random, and M_0 is an artifact of what time we happen to name $t = 0$. So we may treat both of these as random variables uniformly distributed over the interval $[0, 2\pi)$.

So now we will separately consider those seven cases. In each case, we will be interested in addressing the question: how many observations does it take before we can detect a planet using our invented statistic, the MPS? (And in the next two

chapters we will consider exactly the same seven scenarios, but the detection statistic will be different.)

The criterion or criteria for judging how well a statistic works, or how many observations are needed, has not yet been well-defined. We'll address that in the next two sections, while we also consider one of the star scenarios above: that of $K = 7\text{ms}^{-1}$, $e = 0$, which will be instructive.

3.4 The "ROC curve" and the "IMDA", applied to the $K = 7$ m s^{-1}, $e = 0$ scenario

So now let us consider one scenario in particular, that of $K = 7\text{ms}^{-1}$ and $e = 0$, as it will be instructive here. First let's look at the results of 1000 simulations of the MPS for each of the eleven sample sizes of interest. 3.15 shows these distributions, along with their kernel density estimates (to be used later). The comparisons of the cdfs and the ogives are not included here as they were for model M0; however, they are similarly indistinguishable, indicating quite good fits.

If you compare the distributions in Figure 3.12 to those in Figure 3.15, you'll notice that for small sample sizes, the distribution of the MPS is virtually the same for both the no-planet and the one-planet model. However, as the sample size increases, the distributions become more distinct, indicating that the planet is becoming easier to detect with the MPS (as you'd expect). What we need is a way to quantify that detectability. In this section and the next we'll consider two possible ways of quantifying it. In this section we'll develop the "IMDA", or the "Index of Model Distinguishing Ability".

A "Receiver Operating Characteristic" curve, or "ROC curve" (the name comes from signal detection theory) is a plot of true positive detection rates against false positive detection rates. In the language of hypothesis testing, if we posit H_0: There is no planet, and H_A: there is one planet, then the ROC curve is a plot of the *power*

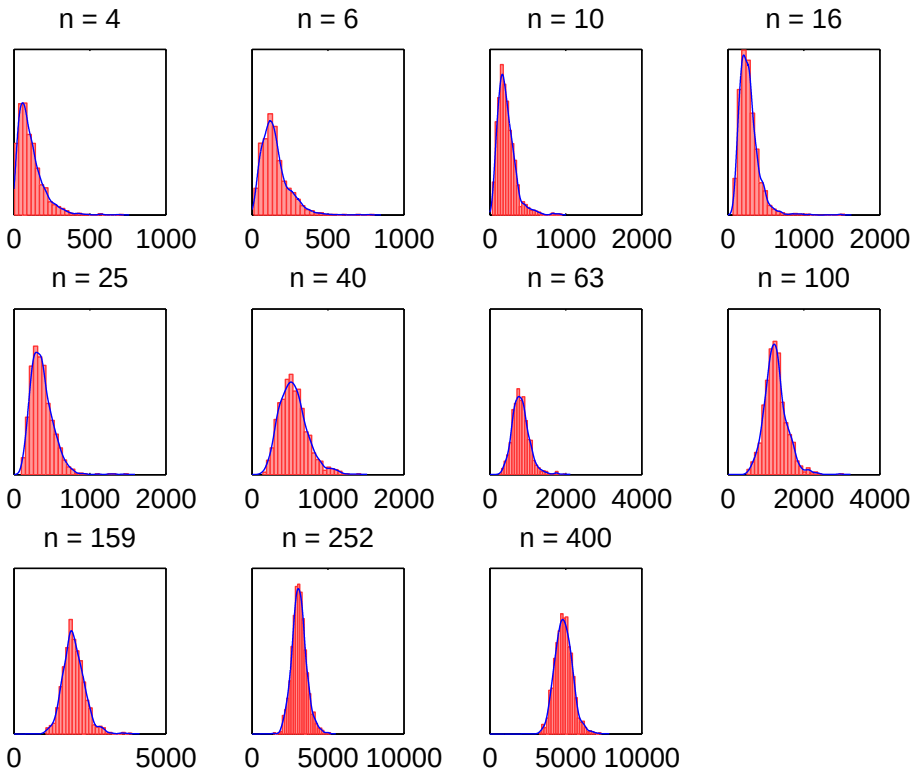

FIGURE 3.15: Distributions of the MPS when $K = 7\text{ms}^{-1}$. Unlike the distributions under model M0, these have a tendency to center on larger numbers when n is larger.

of the hypothesis test (vertical axis) associated with a significance level (horizontal axis).

Let's consider an extreme case of stupid decision-making. You could decide to ignore the data and make your decision about planetary presence using the roll of a 100-sided die. If the roll is less than or equal to a threshold value D, you'll decide there is a planet present. If your roll is greater than D, you'll decide there is no planet present. That is of course an utterly uninformed decision. And because it doesn't matter what the data say, then the chance of a false positive is the same as the chance of a true positive, both depending only on what D you happen to choose. A straight diagonal line from the point (0,0) up to the point (1,1) would be the ROC

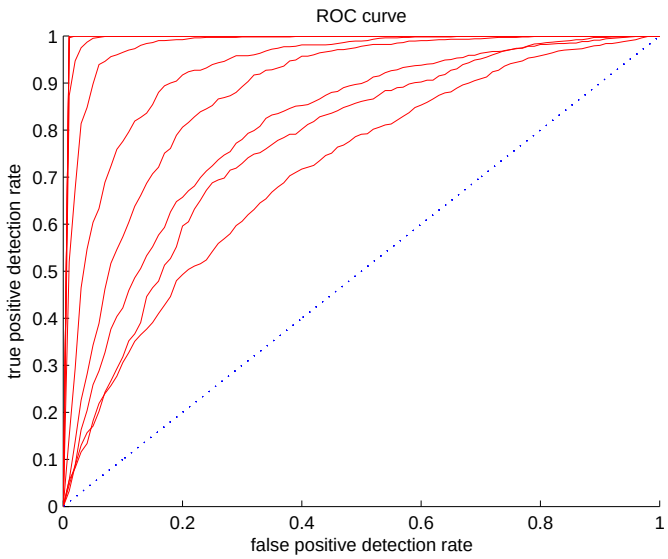

FIGURE 3.16: ROC curves for the statistic MPS when $K = 7\text{ms}^{-1}$ and $e = 0$. The dotted line is the line of maximal uninformativeness. The innermost curve is for the smallest sample size, $n = 4$, and the outermost curve is for the cases $n = 100$ or larger—beyond that sample size they're all essentially utterly informative.

curve for a maximally uninformative decision criterion.

On the other hand, let's consider the opposite extreme, in which the data may be boiled down to some statistic whose distributions under M0 and M1 are almost completely non-overlapping. In that case, no matter what threshold you are willing to allow for a false positive detection rate, the true positive detection rate will be very nearly 1, since the statistic will point so clearly towards the presence of a planet. So a horizontal line at 1 would be the ROC curve for a maximally informative decision criterion.

Actual ROC curves will lie somewhere between these two extremes, but the higher the better. In Figure 3.16 we see ROC curves for all the sample sizes being considered, superimposed on one plot together.

The lowest curve corresponds to the smallest sample size. By the time the sample size reaches $n = 100$, the ROC curve has become maximally informative, and so

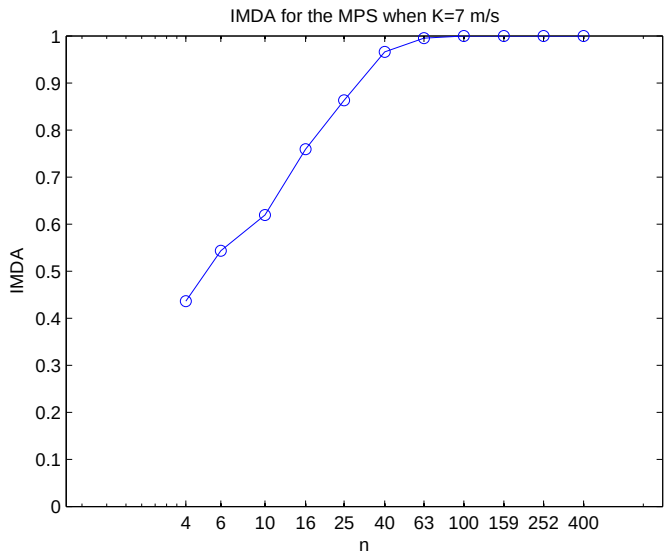

FIGURE 3.17: The IMDA plotted against the sample size shows that the usefulness of the MPS increases steadily with $\log n$ until about $n = 100$ when $K = 7\text{ms}^{-1}$, at which point it is so certain that the planet is there that no further observations can be of any further use for that purpose.

changes almost not at all.

It would be convenient to have a way to summarize the ROC curve with a single numerical measure of usefulness at distinguishing models, and this is where the "IMDA" comes in. Since the area between the ROC curve and the line of maximal uninformativeness ranges between 0 and 0.5, it seems reasonable to compute that area, double it, and let it be an index of model distinguishing ability, or IMDA.

A final graph that would be informative would be a plot of the IMDA against the sample size. That is shown in Figure 3.17.

We see in Figure 3.17 that the IMDA is very nearly 0.5 even when $n = 4$. (How good is that? We'll address that momentarily.) Then as n increases, the IMDA increases roughly linearly with $\log n$ until around $n = 100$, when we're at near perfect distinguishing ability, and so the IMDA can increase no further.

We're about to leave this section for now, but before we do, let's consider the

question of just how good is an IMDA of 0.5? By what standard should we judge it? Any standard will obviously be somewhat arbitrary, but it may be worth considering a scenario common in published research: a researcher performs a hypothesis test at a significance level of $\alpha = 0.05$ and would like to have enough power against the suspected or pertinent alternative hypothesis that rejection of H_0 will be likely; power of 0.8 is commonly sought. If you draw a triangle connecting the points (0,0), (1,1) and (0.05, 0.80), then it forms a sort-of ROC curve with an area of one-half of 0.75, so its IMDA would be 0.75. So perhaps that is as reasonable a criterion as any for deciding when a statistic has reasonably good model-distinguishing ability. By that standard, then if an astronomer committed in advance to making at least 16 observations on any new star—spread out over a reasonable amount of time—then he could be reasonably confident that if there were a planet orbiting the star in a circular orbit, such that the semiamplitude of the star's radial velocity was at least 7 meters per second, then the MPS would be able to pick up on the planet's presence. If no planet were detected by then, and if smaller semiamplitudes were not of interest, then at that point the researcher could quit making observations on that star.[6]

3.5 The likelihood ratio L applied to the $K = 7\text{ms}^{-1}$, $e = 0$ scenario

Now let's consider a second possible tool we might use to evaluate the usefulness of a statistic. As we noted before, the reason that the statistic MPS becomes more useful with larger sample sizes is that its distributions under M0 and M1 become more distinct from one another, with less overlap. (See Figures 3.12 and 3.15.) In the last section we quantified this lack-of-overlap using essentially frequentist tools. In this section we use likelihood tools. We have already seen that quite a good approximation

[6] He or she might return after a long period of time to make another observation in hopes of detecting a *long-period* planet. This is in fact done in practice. But as was noted earlier in this paper, that is not part of this dissertation.

of the density of the MPS may be made using a kernel density estimator, for many sample sizes and under both models M0 and M1 with $K = 7\text{ms}^{-1}$—and also, though we have not seen it yet—for all of the seven scenarios we plan to look at. Using these density estimations, we may compute the ratio of the probability density of the MPS under M1 to its density under M0. That ratio—the likelihood ratio—will be high when the evidence is strong favoring a planet and low when the evidence is weak.

The greater L is, the greater the weight of the evidence that there is a planet present. Richard Royall has suggested [14] that independently of the context, a likelihood ratio of $L > 8$ provides at least "pretty strong" evidence in favor of model M1, and $L > 32$ provides "very strong" evidence in favor of model M1. Similarly, $L < \frac{1}{8}$ provides "pretty strong" evidence in favor of model M0 and $L < \frac{1}{32}$ provides "very strong" evidence in favor of model M0. If $\frac{1}{8} < L < 8$, then we might say that the data are fairly ambiguous, providing at best weak evidence favoring one model, but not so as to be especially conclusive.

This new decision-making criterion has over the classical rejection-region method the advantage that it has a mechanism for supporting model M0, which classical hypothesis testing cannot do.

We will take the simulated data from our model M1 with $K = 7\text{ms}^{-1}$ and for each sample size, compute the likelihood ratio L. We will then categorize the data into one of five categories:

- Very strong ($L > 32$) correct evidence favoring M1;
- Pretty strong ($8 < L < 32$) correct evidence favoring M1;
- Ambiguous evidence ($\frac{1}{8} < L < 8$);
- Pretty strong ($\frac{1}{32} < L < \frac{1}{8}$) *misleading* evidence favoring M0; and
- Very strong ($L < \frac{1}{32}$) *misleading* evidence favoring M0.

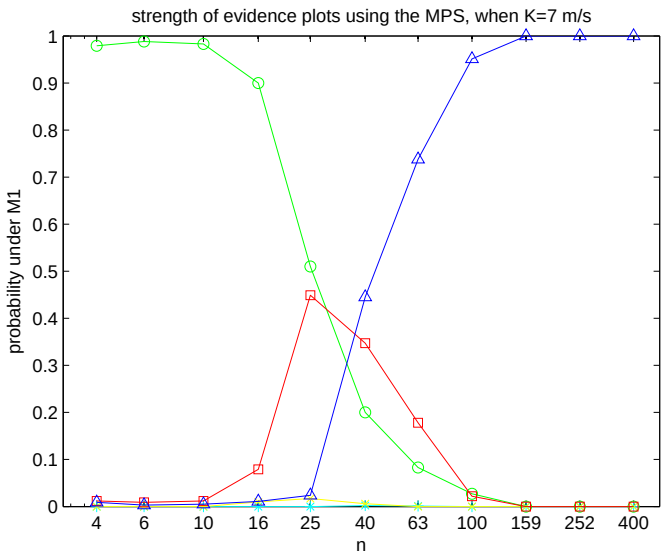

FIGURE 3.18: In this plot, green circles represent ambiguous information; red squares represent "pretty strong" correct evidence favoring M1; and blue triangles represent very strong correct evidence favoring M1. There are also yellow plusses and cyan stars representing, respectively, pretty strong and very strong misleading evidence favoring M0, but these are so improbable that they need only be mentioned briefly.

Those thresholds are obviously somewhat arbitrary, but if we use them consistently for all types of planets and all sample sizes, they will provide useful comparisons.

What we are interested in is the following: if a star really has an orbit with $K = 7\text{ms}^{-1}$ (and $e = 0$, still), then how will the strength of the evidence change with the increasing sample size? We have already seen the kernel density estimates of the MPS density for both models M0 and M1, so all that remains is to compute L for each of our sample sizes and see what we get. Figure 3.18 shows the results.

We see in the plots that with very few data, we are very likely to get ambiguous evidence. However, as n increases, the chance of us getting pretty strong evidence favoring M1 rises, leaving ambiguity behind. Then sometime around $n = 40$ or $n = 63$, the probability of getting very strong evidence favoring M1 increases, and

eventually becomes almost certainly the outcome.

We can see from the plots that with at least 40 or so observations, we are likely to get *at least* pretty strong evidence of the planet's presence (the sum of "pretty strong" and "very strong" probabilities is around 0.8), and with at least 100, it is nearly certain that the evidence will be very strong. This is generally in agreement with the conclusion we reached using the IMDA as a measure of the usefulness of the MPS, although this "strength-of-evidence" standard is a bit stricter than the IMDA standard, requiring more observations. This is important, since the question we're really addressing is, how many observations should an astronomer be prepared to commit to in advance if he wants to be reasonably confident that he'll detect this sort of planet, should it in fact be orbiting the star under observation? There's a big difference between 16 or 25 observations (suggested by the IMDA plot) and 40 observations (suggested by the strength-of-evidence plots). The minimum commitment level really depends on how willing you are to quit when you might be just on the verge of discovering a smallish planet. Using the $n = 16$ or $n = 25$ threshold, the chance for a "Type II error" (failure to detect a planet that is present) is around 10% or 20%. If that is unacceptably high, then the $n = 40$ criterion might be more desirable.

3.6 Other circular orbit scenarios

We need not consider separately all the different scenarios listed earlier in this chapter. Indeed, that would be less than helpful, since one of the things we want to do is compare them: how many more observations are needed to detected this type of planet versus that type of planet? So we will consider here all the four circular orbit scenarios together, and in the next section we will consider together all three scenarios having $K = 20\text{ms}^{-1}$ and varying eccentricities.

We have already established two graphical means for evaluating the usefulness of

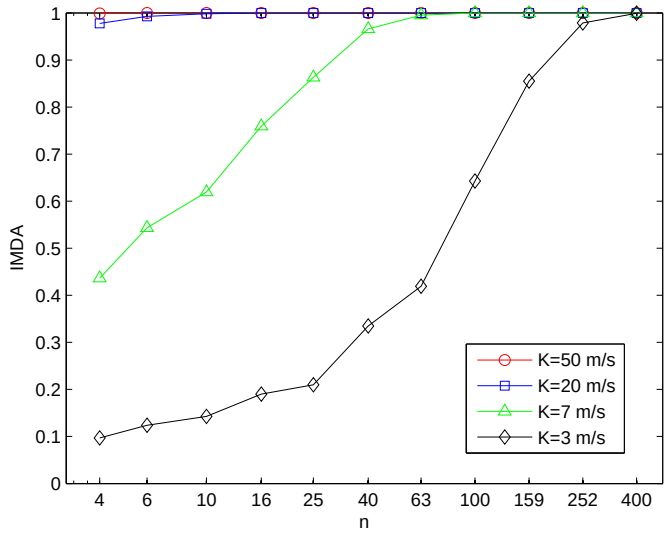

FIGURE 3.19: The IMDA plots for each of the four semiamplitudes under consideration. The reason the IMDA is so high for large K is that even with as few as $n = 4$ radial velocity observations, if the semiamplitude is large, then it is highly probable that the range of the observed radial velocities will be too large to attribute to mere stellar noise and measurement errors. On the other hand, when $K = 3\text{ms}^{-1}$, the semiamplitude is comparable in magnitude to the stellar jitter, and distinguishing periodic behavior from all the noise is more difficult.

a statistic at detecting the presence of a planet: the IMDA plot, and the "strength-of-evidence" plot. We will now repeat the analyses we did in the two preceding sections, only this time using the other semi-amplitudes for K: 50ms^{-1}, 20ms^{-1}, 7ms^{-1} and 3ms^{-1}.

Figure 3.19 shows the IMDA plots for each of those four circular-orbit scenarios, and Figure 3.20 shows the strength-of-evidence plots for the same. For clarity, the latter plot shows only the probability of "very strong evidence" correctly favoring model M1 over M0 (i.e., a likelihood ratio of at least 32).

Almost unawares, one of the most interesting findings in this dissertation has now appeared. A planet having a circular orbit with a period of 7 days and inducing a semiamplitude of $K = 3\text{ms}^{-1}$ on its host star could be smaller than any stand-

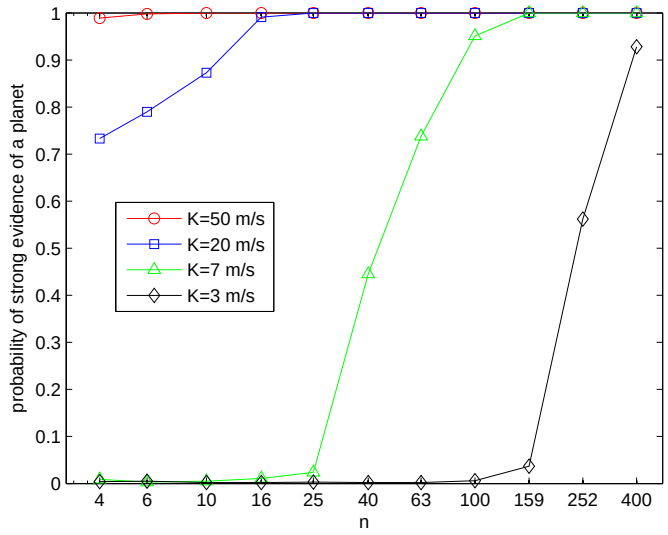

FIGURE 3.20: Strength-of-evidence plots for the MPS for the four circular orbits under consideration. As expected, larger K's are easier to detect than smaller ones. Even fairly small semiamplitudes are detectable with enough observations.

alone planet yet discovered, about 7 times the mass of the Earth. It would be a very hot planet, because the small period would indicate a close proximity to its host star. Whether such planetary systems are common or rare is completely unknown. However, if a commitment were made in advance to make 400 radial velocity observations on each of, say, 10 promising candidate stars—regardless of whether they showed evidence of a planet as the observations accumulated, then in each case, if there were such a small planet present, there would be a very high probability of detection. The commonness or rarity of such systems could then be roughly estimated based on how many of the 10 stars under observation did in fact show a planet. 4000 radial velocity observations is of course a lot; but at present more than that are made annually on hundreds and hundreds of different stars, and generally speaking, if they don't show signs of a planet after a dozen or so observations, they are shelved for a long time. Perhaps it is at least worth considering committing some radial velocity measurements to a small number of stars for several

years, until around 400 or so observations are made on each one. The detection of no planets after that period would be disappointing, but it would still be informative; if they were there, they would have been found.[7]

3.7 Eccentric orbit scenarios

Now we repeat the analysis from the last section, only this time we apply it to the three scenarios of varying eccentricity. In all cases, we let $K = 20\text{ms}^{-1}$; we consider $e = 0$ (which is actually a repeat of one of the scenarios from the last section), $e = 0.3$, and $e = 0.6$. Figures 3.21 and 3.22 show, respectively, the IMDA plots and the strength-of-evidence plots for these three types of orbits.

Recall that it was not obvious whether eccentricity would make an orbit with a given semiamplitude easier or harder to detect with the MPS. On the one hand, greater eccentricity meant greater potential variations in the radial velocity, should the observations happen to include any from that spell of time when the planet is swinging close to its host star and moving more rapidly. On the other hand, greater eccentricity can also mean long lulls when little seems to be happening at all. Additionally, the periodogram is not designed to detect Keplerian motion, but sinusoidal patterns.

These two plots answer the question: greater eccentricity makes a planet harder to detect. The effect is apparently—very roughly—that up to modest eccentricities of about 0.3 (the real cutoff cannot be seen without simulations of more eccentricities), the eccentricity affects the orbit so little that the planet is not much hard to detect

[7] Since the writing of this paper, I have learned that there are at least two reasons for being more circumspect with regard to these findings and recommendations. Firstly, expending such large resources on a single object is not feasible given limited telescope resources. Second, for $n = 400$, one is assuming that the model is accurately describing the noise and jitter to an unrealistic level—the normal assumptions are essentially letting the noise average out to a level of $\frac{1}{\sqrt{n}}$, or one twentieth of its nominal single-sample value, which is of the order of 1 meter per second. It is likely that systematic error is larger than this and will come to dominate the results well before $n = 400$.

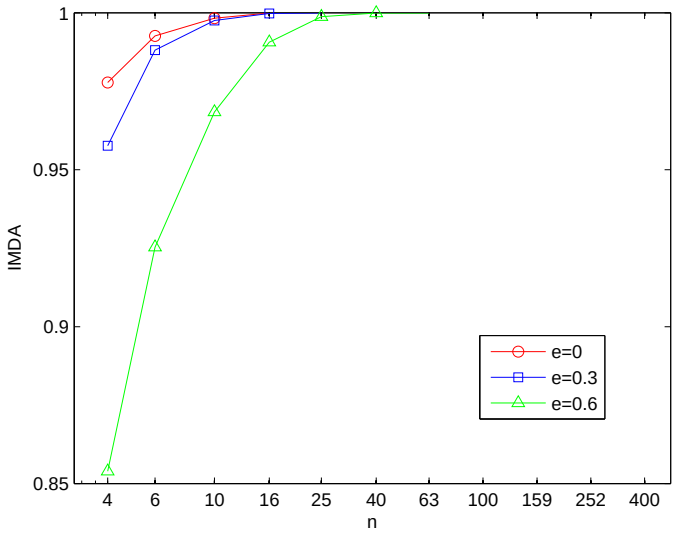

FIGURE 3.21: IMDA plots for MPS under the three scenarios with varying eccentricity. Note that the vertical axis begins at an IMDA of 0.85. Even with $n = 4$, highly eccentric orbits having this radial velocity semiamplitude should be fairly easy to detect.

than would be a similar planet with a circular orbit. But when the eccentricity reaches 0.6, about three times as many observations are required (relative to a circular orbit) to be similarly certain of detecting a planet.

3.8 Already detected planets

Finally, we suppose that the distribution of orbital parameters of stars that do host planets is similar to those that have already been found to host planets. This is not a reasonable assumption, but it will lead us to one useful bit of information. We run simulations similar to those already described in this chapter, only this time we draw orbital parameters from distributions modeling those of known planets. The resulting IMDA plot and strength-of-evidence plots are shown in Figures 3.23 and 3.24 respectively.

What we learn from these two plots is essentially this: if all we cared about

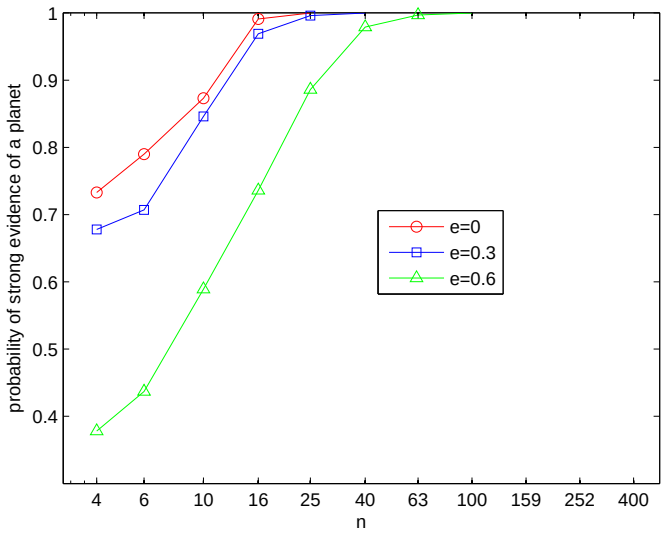

FIGURE 3.22: Strength-of-evidence plots for the MPS under the three scenarios with varying eccentricity. The $e = 0.6$ plot appears, very roughly, to be a horizontal shift (on the log scale) of the the other plots, indicating that perhaps three times as many observations should be expected to detect a planet with a highly eccentric orbit, compared to one with a circular orbit. That is for $K = 20\text{ms}^{-1}$—it is not known whether the same would be true for larger or smaller semiamplitudes.

was detecting lots more planets of the sorts we've already found, then relatively few observations should be committed to each star. Indeed, a mere four observations on each star would be the most efficient use of resources. (That's because the expected number of detected planets per radial velocity measurement is highest at $n = 4$; indeed, it might be higher still at $n = 2$.)

3.9 MPS summary

Thus we see that if the maximal periodogram spike is used as a planet-detection statistic, then we may expect the following:

- Larger planets require fewer observations to detect than do smaller ones;
- More eccentric orbits—especially those with $e >\approx 0.6$—require about three

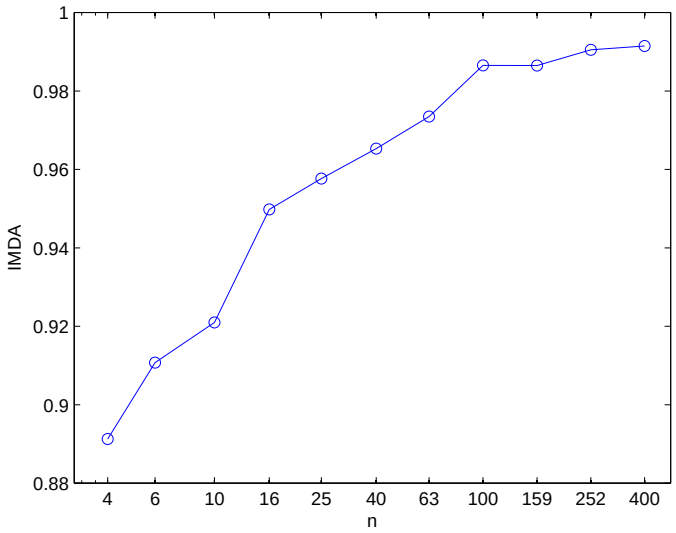

FIGURE 3.23: IMDA plot for the MPS, using stars mimicking those known to host planets. Note that the vertical scale begins at 0.88.

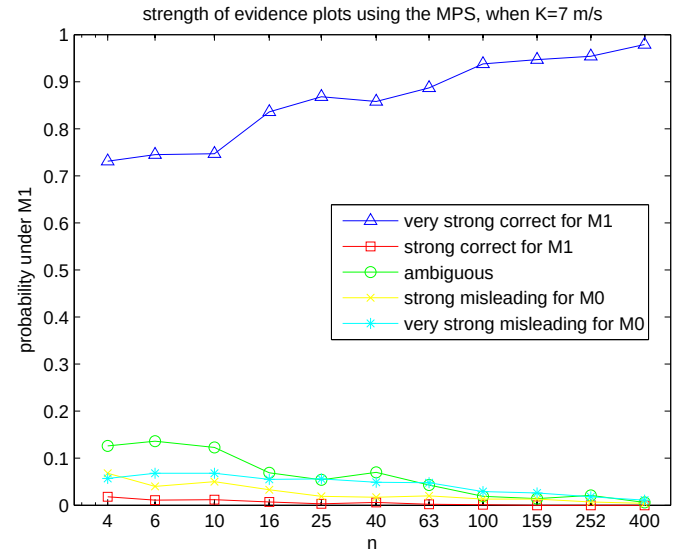

FIGURE 3.24: Strength-of-evidence plots for the MPS, using stars mimicking those known to host planets.

times as many observations to detect (when $K = 20\text{ms}^{-1}$) than do comparable circular orbits;

- Even relatively small planets are detectable with enough observations, provided that the sources of systematic error can be controlled to a small fraction of a meter per second (this may be unrealistic).

- It is possible to know in advance how many observations one should be prepared to make if one is hoping to detect orbits of a certain type;

- It is similarly possible, after making some observations, to say, "if a planet of this type were there, we would have detected it by now";

- If your goal is to detect lots of planets, make fewer observations per star;

- If your goal is to estimate the commonness of systems with fairly small largest planets, then commit to hundreds of observations per star for a small number of stars.

In the next two chapters we will perform similar analysis on two other statistics that might be used to detect planets: the Maximum Likelihood Ratio, and the Bayes Factor.

4

The Maximum Likelihood Ratio

In the last chapter we examined the "maximum periodogram spike" in part to see how many observations it would take, using this statistic, before we could reasonably expect a planet of any of several types to become discernable. But we also examined it so that we could see whether there might be another statistic—one that can be computed with reasonable efficiency—that might do better; i.e., one that might require fewer observations to discern a planet.

In this chapter we will consider the Maximum Likelihood Ratio, or MLR, as a tool for distinguishing model M0 from model M1. Under certain assumptions, its distribution approaches a limiting distribution as n increases. If Λ is the maximum likelihood ratio for model M1 (which has seven parameters) to model M0 (which has only two), then theoretically, for large n, the distribution of $-2\log\Lambda$ should be approximately χ^2_5—the five degrees of freedom being the difference in the numbers of parameters between the two models.

In this chapter we will see how well that holds when:

- the sample size is small;

- the MLE of model M1 is somewhat hard to find, and so the resulting maximum likelihood estimate may in fact be too small; and
- measurement errors are taken into account. (The theory behind the limiting distribution permits additional noise like this in both models, but it slows down the rate at which the distribution approaches its limiting distribution.)

If the theory holds up well, then we would expect, even when there is no planet present, to find a maximum likelihood ratio of as much as 1000 to 1 favoring model M1.[1] Values of Λ beyond 1000 would be more unusual and may indicate a planet. That's just the theory, but it gives us something to be looking out for. We'll once again use simulations to see what the actual distribution of Λ looks like when model M0 is true, and then when various versions of model M1 are true. (We'll consider the same versions—i.e., the same types of orbits—that we looked at in the last chapter, so that we may compare that chapter's statistic MPS to this one's Λ.

4.1 Finding the maximum likelihood under M0

Recall that under the no-planet model, we have:

$$V_i \overset{\text{iid}}{\sim} N(C, \sigma_i^2 + s^2), \tag{4.1}$$

where C and s are the only two parameters that require estimation. (The measurement errors σ_i^2 are known.) If we let L be the likelihood under this model, and $\ell \equiv \log L$ be the log likelihood, then we have:

$$\ell = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum \log(\sigma_i^2 + s^2) - \frac{1}{2}\sum \frac{(C - V_i)^2}{\sigma_i^2 + s^2} \tag{4.2}$$

[1] Note that likelihoods are not probabilities per se, and the meaning of a particular likelihood ratio is not constant between different problems, nor even between the same problem with different sample sizes.

This is quadratic in C, which means that its conditional maximum (conditional upon s) is quite easily found.

$$\frac{d\ell}{dc} = \frac{-1}{2} \sum \frac{2(C - V_i)}{\sigma_i^2 + s^2}.$$

Setting this equal to zero yields:

$$C = \frac{\sum \frac{V_i}{\sigma_i^2 + s^2}}{\sum \frac{1}{\sigma_i^2 + s^2}}.$$

That makes intuitive sense: C is simply a weighted average of the observations V_i, the weights being proportional to the precisions.

It is a little trickier maximizing ℓ in s, given its peculiar attachment to σ_i^2 in each term. Setting the first derivative $\frac{d\ell}{ds}$ equal to zero yields:

$$s \left[\sum \frac{-1}{\sigma_i^2 + s^2} + \sum \frac{(C - V_i)^2}{(\sigma_i^2 + s^2)^2} \right] = 0 \tag{4.3}$$

$s = 0$ is obviously not an interesting solution. Writing the remaining factor with a common denominator gives:

$$\sum \left[-(\sigma_i^2 + s^2) + (C - V_i)^2 \right] = 0 \tag{4.4}$$

This is quadratic in both C and s; we have an earlier equation relating C and s as well. It is possible to solve the system for both C and s simultaneously, although it is sufficiently messy algebraically that a numerical solution is appropriate, and that is what we will use throughout this chapter for finding the MLE, and consequently the maximum likelihood, for any data set under model M0.

4.2 Finding the maximum likelihood under M1

Finding the maximum likelihood under M1 is not so easy. Not only is an analytical solution impossible, but a numerical search is also quite challenging, for the same

reasons that were brought up in Chapter 2: the likelihood function is multidimensional and multimodal. A strictly numerical search for a maximum that is completely uninformed by any properties of the likelihood function has virtually no chance of finding the true MLE.

Fortunately, we do have information that can guide the search: the periodogram. In the last chapter, we used the MPS—the maximal periodogram spike's height—as a statistic unto itself. Here we will use the periodogram in a different way. We will look not only at the tallest spike, but at all spikes of comparable height to the tallest, and use those as starting points to help guide us to *local* likelihood maxima in the full seven-dimensional parameter space. There will still be no guarantee that we'll have found the true global MLE in the end. In particular, since the periodogram is based on sinusoidal patterns and some Keplerian patterns are not very sinusoidal, there is a chance that the greatest likelihood will occur at a period that doesn't stand out in the periodogram. We will have to take somewhat on faith that the global MLE—which may not, of course, correspond to the actual parameter values in the model that generated the data—will in fact correspond to a period that stands out in the periodogram. However, we need not take this entirely on faith. In the very first section we hope to show that it is a reasonable tenet.

After that we will discuss an algorithm intended to find the global MLE, and consequently the maximum likelihood, associated with model M1. Then we will have the tools we need to compute the maximum likelihood ratio, Λ, for any data set, and we may study its usefulness in exactly the same way we studied the usefulness of the MPS in the previous chapter.

4.2.1 Can the MLE's period hide easily in the periodogram?

Of some concern is whether an MLE for a non-sinusoidal Keplerian function can have a period that does not stand out in the periodogram, due not to noise, but to

the fact that the periodogram wasn't made to look at sawtooth-wave functions and their ilk. In this section we hope to show that that is a rarity.

It's tempting to simulate a lot of data from an eccentric-orbit model, and look for the model's period P in the resulting periodograms. But that wouldn't address the real question, which isn't whether the actual generating P must stand out in the periodogram, but whether the MLE's period P must stand out. And of course, the MLE need not correspond to the actual parameter values.

But if we instead simulate lots of data and then try to look for their MLE's periods in their periodograms, then we'll be up against the very problem we have to solve: how to find the MLE. Fortunately, there is a way out of this Gordian knot. We can simulate data from stars with eccentric orbits—*but not incorporate any noise into the simulated radial velocities*. Then the MLE will have to correspond to the model that produced the data, since you can't do any better with a set of normally distributed data than to have each one take the value of its distribution's mean. However, the periodogram will not know this; to the periodogram, deviations from pure sinusoidal behavior will be taken as noise, and will potentially lower the spike in the periodogram corresponding to the true P. We wish to see whether this is common.

Let's first consider a single example. Figure 4.1 shows 10 simulated observations from a star with $P = 7$ days, $K = 20\text{ms}^{-1}$, $e = 0.5$, and $M_0 = 1$. Since there is no noise, the MLE of the parameters must be the model that generated the data, which includes $P = 7$ days. Will the periodogram show a spike at $P = 7$? Figure 4.2 shows that it does indeed.

Figure 4.3 shows another example, this time even more non-sinusoidal, and yet the period $P = 7$ does show up as a spike in the periodogram, or very nearly so. Only three spikes are of comparable height, and none are considerably taller.

We should consider an example in which the eccentricity is large, ω is near zero,

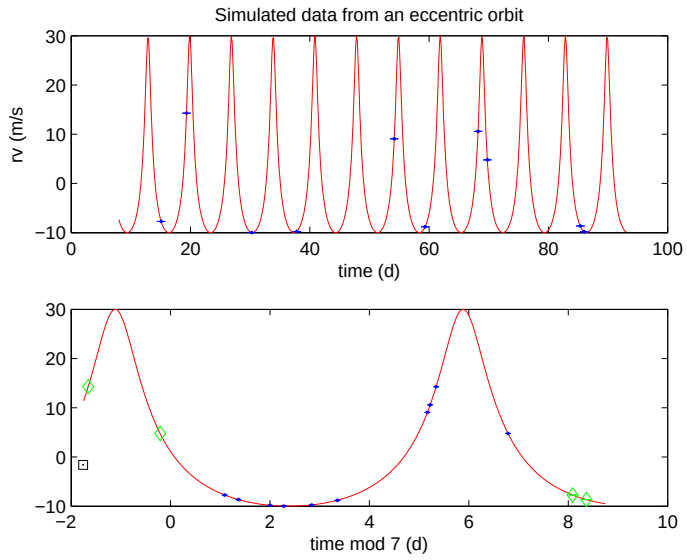

FIGURE 4.1: Ten simulated radial velocities from a star with an eccentric orbit, with no noise at all. The upper plot shows the data over true time, the lower plot shows them over time modulo 7 days.

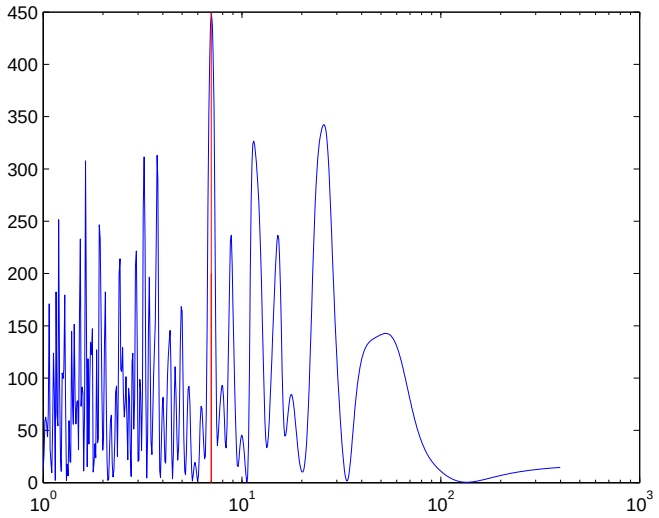

FIGURE 4.2: The periodogram for the simulated data shows that the tallest spike is indeed at $P = 7$ days, even though the radial velocity function was not sinusoidal.

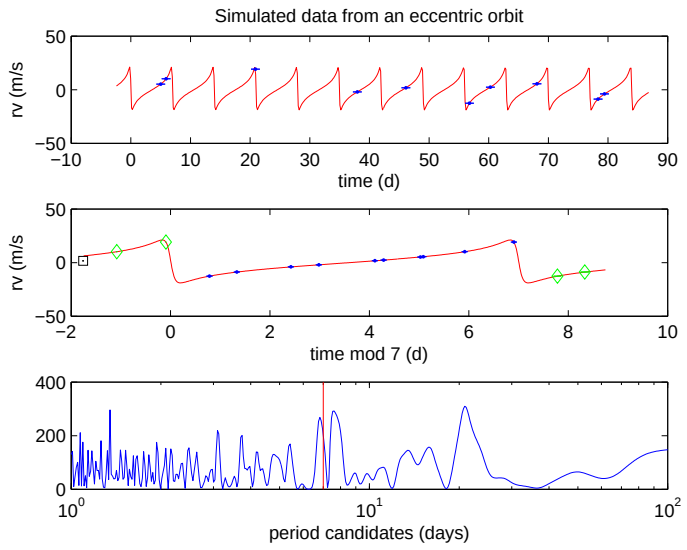

FIGURE 4.3: Here, $\omega = 1.5$, nearly $\frac{1}{2}\pi$, and $e = 0.8$. Yet still $P = 7$ stands out in the periodogram.

and the observations don't happen to include any that occur during the brief periods when the radial velocity changes rapidly. Figure 4.4 shows such an example.

In this example, no data occur during the period of periastron passage, when the planet whips around the star at its closest point. Even so, the periodogram picks out the P of the MLE, because of the gentle curvature of the radial velocity function at other times.

These three examples do not prove anything, nor is that my intent here. I only wanted to provide evidence that periodic Keplerian motion that is not sinusoidal can produce data for which the MLE's period does still show up in the periodogram. It may not be the tallest spike—and we must keep that in mind shortly as we develop an algorithm for finding the global MLE—but it appears to produce a spike nonetheless.

In the next several sections we'll develop an algorithm meant to find the global MLE under model M1 so that we may use its likelihood in the likelihood ratio Λ. Let's begin by considering once again a particular simulated data set—one that will

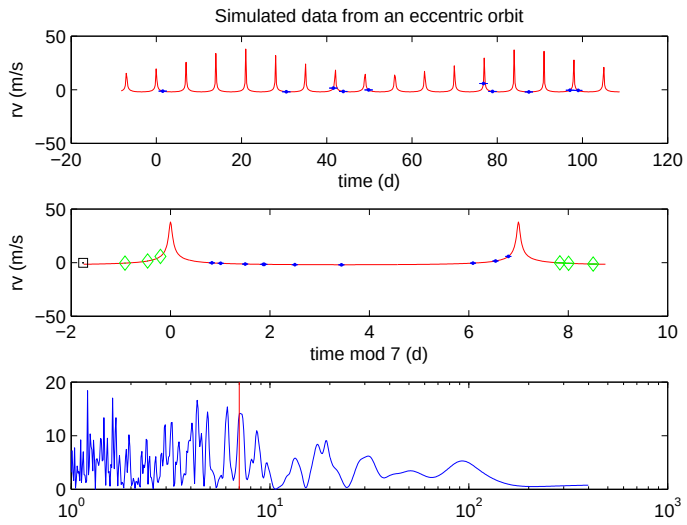

FIGURE 4.4: Here, $\omega = 0$, and $e = 0.9$. Yet rather remarkably, $P = 7$ still stands out in the periodogram.

serve as a good example for motivating the steps of our algorithm as we develop it. We'll keep some of the parameters from the simulations in the last section ($C = 0$ ms^{-1}, $K = 50$ ms^{-1}, $M_0 = 0$, and $P = 7$ days) but we'll now use $e = 0.6$, $\omega = 1$ (radian), and $s = 4.5$ ms^{-1} (a fairly typical stellar jitter value). Figure 4.5 shows these 10 points (sampled at random from a uniform distribution over a 30-day period) and the function matching the parameter values that generated them.

Since we're looking for the global MLE, it is very reasonable right now to ask whether other orbital parameters would also fit these data—perhaps local MLE's. We begin by considering a periodogram of the data, which is shown in figure 4.6.

In this periodogram there are four maxima of comparable height. They are located (in order from tallest to shortest periodogram spike) at $P = 7.148$ days, $P = 3.482$ days, $P = 1.162$ days, and 2.100 days. Figure 4.7 shows the simulated data shown modulo each of these four periods.

It would be good to look at these four plots with radial velocity models superim-

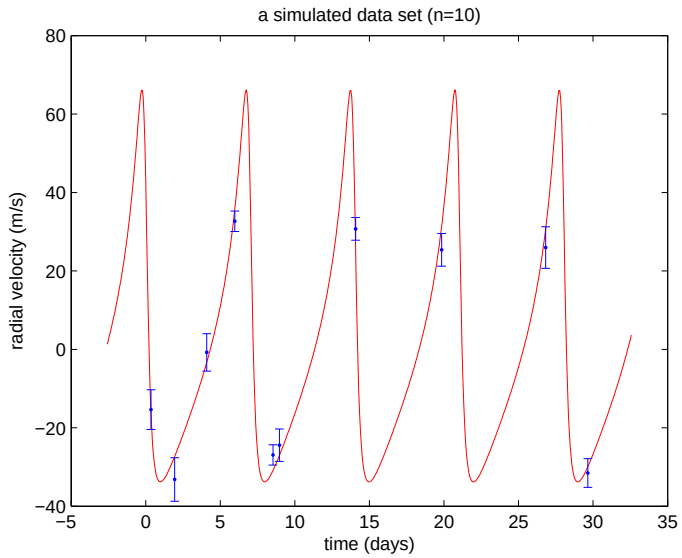

FIGURE 4.5: Ten simulated radial velocities from a star with an eccentric orbit. The curve shows the radial velocity function corresponding to the orbital parameters from which the data were generated.

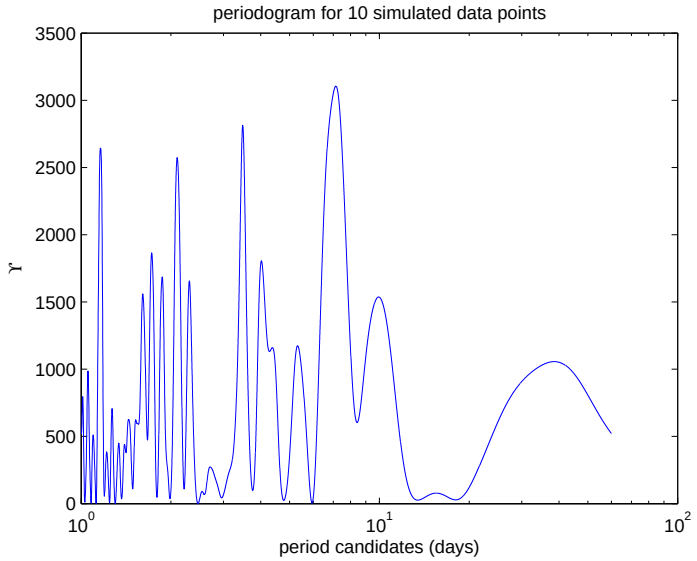

FIGURE 4.6: The periodogram for the data suggest that there are at least two periods that are equally plausible .

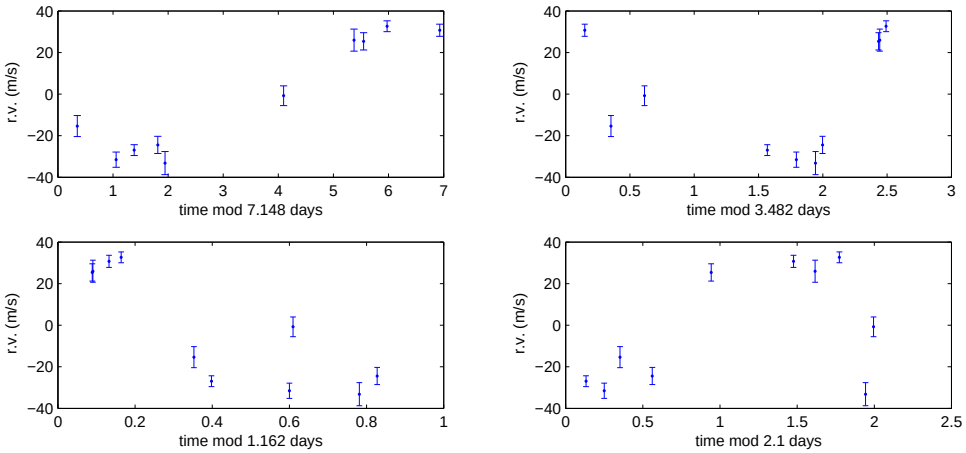

FIGURE 4.7: The simulated data are shown under four moduli, corresponding to the four tallest local maxima in the periodogram. Given the size of the error bars, it would appear that there is considerably less likelihood attached to the last of these than the first, but one cannot be sure without seeing superimposed upon the data its corresponding best possible orbital velocity function—especially since these illustrations do not include the stellar jitter s, which can account for some of the deviation from the orbital velocity function.

posed, as they were in figure 4.5. But we no longer want to compare the data with the function that generated the data; now we want to compare the data with the model that best fits the data under the given period, or a comparable period. To do that we need a method for finding local maxima in the likelihood function, which we'll address next. We'll then return to the four plots in figure 4.7 and superimpose their local MLE radial velocity functions.

4.2.2 Finding local maxima in the likelihood function, and an algorithm for finding the global maximum likelihood

Now let's find local maxima in the full 7-parameter likelihood function[2] that "correspond" to each of these three solutions, and superimpose them on the plots. What do we mean by "correspond"? We know that in the full 7-parameter likelihood func-

[2] There are only six orbital parameters; the seventh parameter, the stellar jitter s does, however, appear in the likelihood function.

tion, there are many hundreds of local modes. We would like to find the modes that are "nearest" to the solutions we've found. In particular, we want to find, for each spike in the periodogram, a mode in the likelihood function whose period is very close to the location of the periodogram's spike. If there happen to be multiple distinct modes corresponding to approximately the same period (this could happen, for example, if the eccentricity e is large but the data do not nail down the argument of periastron M_0) we would like to find all of them. So how are we going to find these modes?

Let us look once again at the likelihood function, first introduced in chapter 1 of this paper and then revisited in chapter 2 when it was shown that the function is highly multimodal in the parameter P. Recall that the data have the following distribution:

$$V_i \sim \mathsf{N}\left(C + \Delta V(t_i|\theta), \sigma_i^2 + s^2\right),$$

where

$$\theta = (K, P, e, \omega, M_0),$$

and

$$\Delta V(t|\theta) = K\left[\cos(\omega + T(t|e, P, M_0)) + e\cos\omega\right].$$

The likelihood function is therefore given by

$$L(\vec{t}, \vec{V}, \vec{\sigma}, \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi(\sigma_i^2 + s^2)}} \cdot \exp\left(-\frac{1}{2}\frac{\left[V_i - (C + \Delta V(t_i|\theta))\right]^2}{\sigma_i^2 + s^2}\right) \tag{4.5}$$

And since we are at present interested only in finding maxima, we may more easily consider the log of the likelihood function instead:

$$\ell(\vec{t},\vec{V},\vec{\sigma},\theta) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\left[\sum \log(\sigma_i^2+s^2) + \sum\left(\frac{[V_i-(C+\Delta V(t_i|\theta))]^2}{\sigma_i^2+s^2}\right)\right] \tag{4.6}$$

It is not obvious how the function ℓ behaves with respect to the different parameters. A few observations may be made, however:

- Obviously the constant $-\frac{n}{2}\log 2\pi$ is irrelevant to the maximization process.
- ℓ is quadratic in C and therefore unimodal in C; and holding all other parameters constant, ℓ may be easily maximized in C.
- K enters the function ΔV only as a single factor, and therefore ℓ is also quadratic (and unimodal) in K; and again, holding all other parameters constant, ℓ may be easily maximized in K.
- Looking carefully, we can see that the function ΔV is actually sinusoidal in ω, and (holding all other parameters constant) does not vary with t_i. Therefore, ℓ is unimodal in ω and may be fairly easily maximized in this parameter as well.
- s is somewhat more problematic. If for convenience we briefly define $\xi_i \equiv [V_i-(C+\Delta V(t_i|\theta))]$ (which is the actual, though unknown, difference between the measured radial velocity and the actual radial velocity), then $\frac{\partial \ell}{\partial s} = s\sum \frac{\xi_i^2-(\sigma_i^2+s^2)}{(\sigma_i^2+s^2)^2}$. We may note that $\lim_{s\to\infty}\frac{\partial \ell}{\partial s} = 0^-$, so the function ℓ must eventually decrease forever. There is also a root at $s=0$, though it may correspond to an endpoint maximum or minimum, depending on the relative magnitudes of the ξ_i's and the σ_i's. If the σ_i's were all relatively tiny, then $\frac{\partial^2 \ell}{\partial s^2}|_{s=0}$ would be positive, favoring somewhat larger values of s over $s=0$. That makes sense: if the measurement errors are small, then the deviations must be due to non-zero

stellar jitter. Similarly, if the σ_i's are all relatively huge, then $\frac{\partial^2 \ell}{\partial s^2}|_{s=0}$ would be negative, favoring $s = 0$. However, in such a case, we would be likely to question the presence of a planet at all, unless perhaps there were a huge number of data. In either case, we posit here without proof (though with a graphical justification to follow shortly) that the behavior of ℓ with respect to s is at any rate not very violent. Common sense dictates that as s increases away from zero, the likelihood function must increase to explain radial velocity deviations that are not sufficiently explained by measurement errors. Then eventually as s continues to increase, we have already seen that the likelihood will eventually decrease to infinity.

- P is very much a problem, but since the periodogram suggests values of P that are good candidates, we will try to devise a search that fixes P at each credible periodogram peak to get "close", and then finally frees P up to move to its true value in the MLE.

- The remaining problem parameters are e and M_0. These, like s, we will simply study informally and posit that the behavior of ℓ with respect to each of them is not terribly violent.

I have stated without proof (or indeed any evidence yet) that the behavior of ℓ with respect to the parameters s, e, and M_0 is not necessarily unimodal, but is also not very violent. A proof of this claim is not forthcoming. We have seen that the derivative $\frac{\partial \ell}{\partial s}$ involves a messy sum and so is not easily set equal to zero and solved for s. Also recall that the function $\Delta V(t_i|\theta)$ is transcendental, making it impossible to differentiate. However, ΔV is not utterly incomprehensible. We know that when $e = 0$, ΔV becomes sinusoidal in t_i, and also incidentally in M_0 and in $\frac{1}{P}$. What's more, as e deviates from zero, the function ΔV changes only rather gradually from being sinusoidal (in each t_i) to being more sharply pointed. (Figure

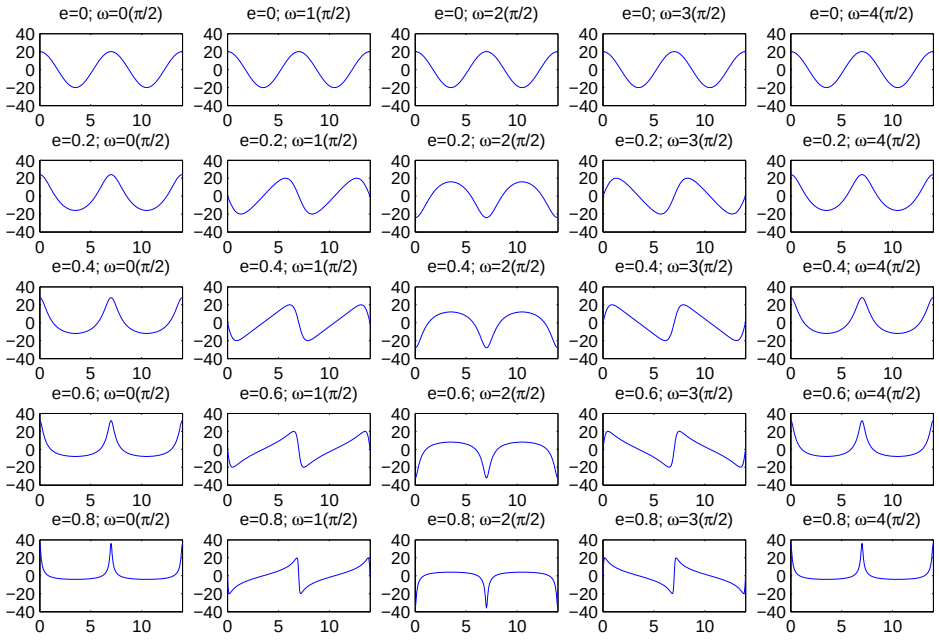

FIGURE 4.8: The radial velocity function reduces to a sinusoidal function when $e = 0$. But even with $e > 0$, the function does not exhibit violent behavior in t_i; and therefore one might reasonably expect the behavior with respect to e and M_0 to be fairly non-violent as well.

4.8 is recalled from chapter 1, illustrating this.) Therefore, one might reasonably expect the behavior of ℓ with respect to M_0 and e to be fairly non-violent as well.

Let us pause to look at some "slices" of the likelihood function (i.e., conditional upon fixed values of the other parameters) ℓ across each of the parameters s, e, and M_0 for the simulated data set that we saw earlier, which was shown in figures 4.5, 4.6 (the periodogram), and 4.7. (There is no need to look at slices with respect to C or K, as they would be mere parabolas; or ω, which we've seen must be sinusoidal; or indeed P, which we know would be spiky like the periodogram.) Let us first suppose that we could somehow know what the true parameter values were that generated the data. (They were $C = 0$, $K = 50$, $P = 7$, $e = 0.6$, $\omega = 1$, $M_0 = 0$, and

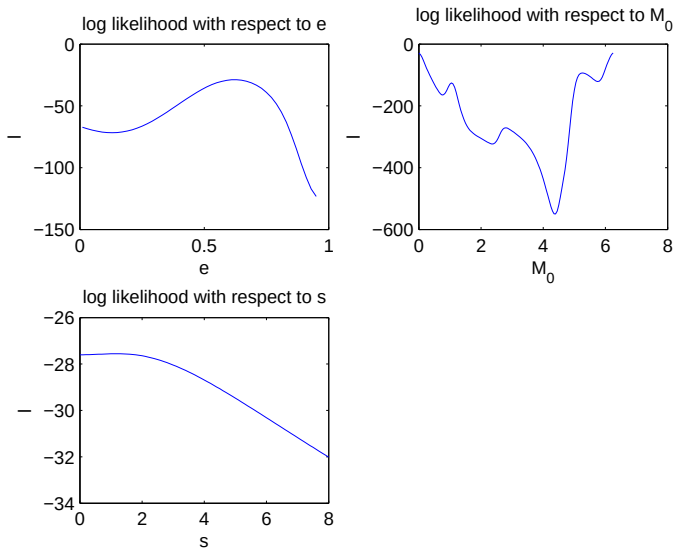

FIGURE 4.9: The log likelihood function shown here in three cross-sections—i.e., conditional upon fixed values of the other parameters—all passing through the point in the parameter space that generated the data. The function is not very violently fluctuating in the parameters e, s, and M_0 as it is in P.

$s = 4.5$.) Let's look at one- and two-dimensional cross sections of the function ℓ that pass through this location in the parameter space, focusing on s, e, and ω. Figure 4.9 shows the three one-dimensional cross-sections passing through the point that generated the data; figures 4.10, 4.11, and 4.12 show two-dimensional cross sections of the same function passing through that point.

These images suggest that the likelihood function, while highly multimodal in P—especially in the range of small values of P—it is not so extremely multimodal in the other parameters. M_0 behaves rather like a phase-shift (indeed, it is a phase shift in the case when $e = 0$); if all the other orbital parameters are nearly correct, but the phase shift is quite wrong, then of course the likelihood will be abysmally low. That is why the graphs in these figures show such a dramatic drop in ℓ between the correct value of M_0 (which was $M_0 = 0$ for these simulated data) and other values—notice the scale on the graphs.

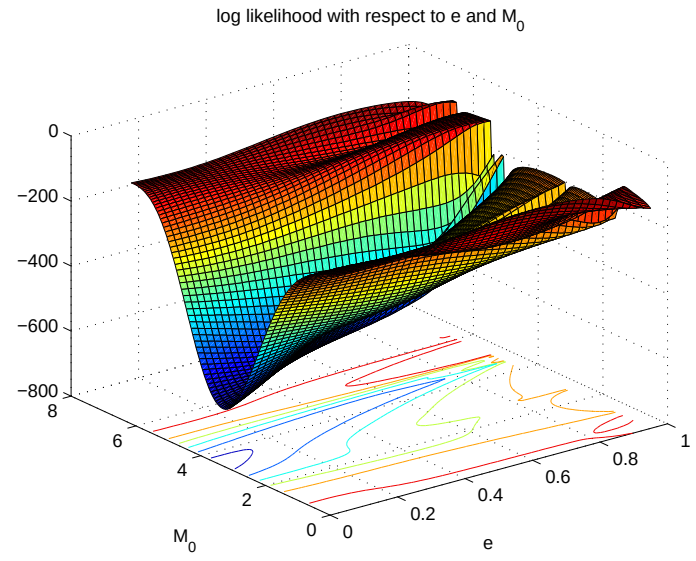

FIGURE 4.10:

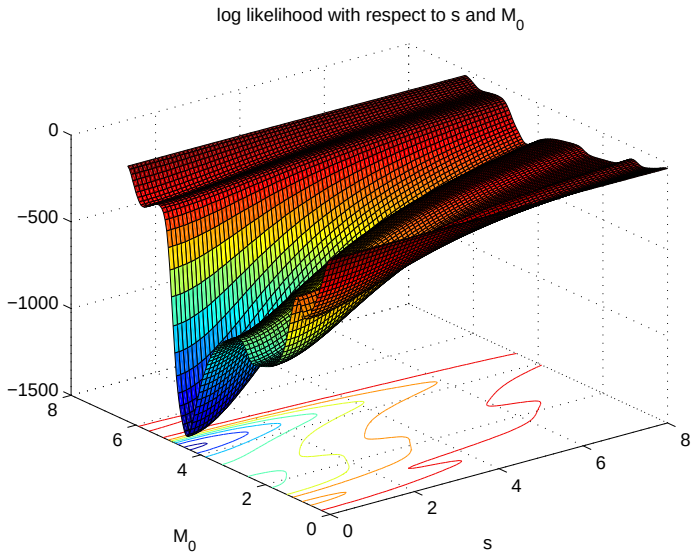

FIGURE 4.11:

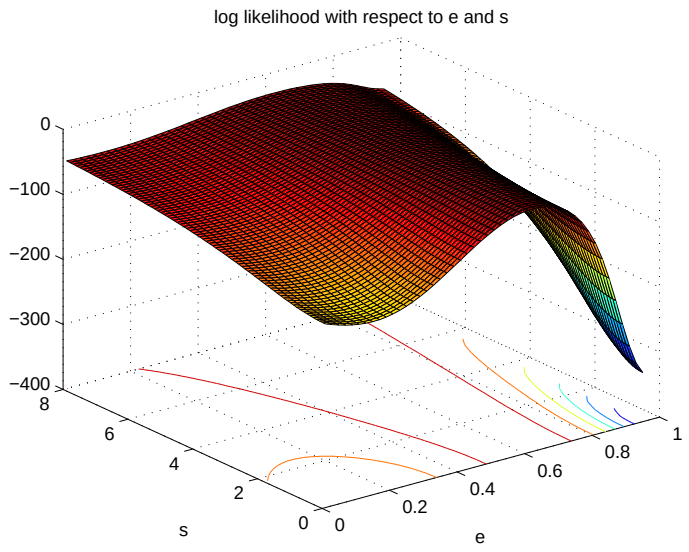

FIGURE 4.12:

These images are merely meant to suggest an approach to searching for modes in the likelihood function. Further, we have only looked at cross-sections of the log likelihood function at the value of θ that generated the data. And we have done so for only one simulated data set, with sample size $n = 10$. Does ℓ behave similarly in other regions of the parameter space? For larger or smaller data sets? For stars of greater or lesser eccentricity?

I am going to simply posit that the function ℓ is somewhat multimodal in M_0, but not so much so that ℓ will be as "spiky" in M_0 as it is in P. It must get more spiky as n increases, since the increase of M_0 would result in the radial velocity function fluctuating around more points. If there really is a planet present and the data are sufficient to detect it, then spikiness shouldn't be a big problem, as the likelihood will be so much higher in the "correct" region of the parameter space than everywhere else. The problem is when there is no planet present; and most especially when there is no planet present and n is large. This we will look at in the next section.

But before going there, let us use what we've shown or posited here to devise a

method of finding local maxima in ℓ. This is what we will do:

1. First, we will order the maxima of the periodogram according to Υ. One by one we will take the following steps on the maxima we've found, until we decide that we've found all the maxima in ℓ that matter. (A stopping rule will be discussed later.)

2. We will fix P at the value suggested by the periodogram spike. Then we will "grid on" s, e, and M_0, dividing each of their domains uniformly into segments and thereby creating a lattice in their three dimensions. The grid will be fairly coarse on s and e, while finer on M_0. Here we will divide s and e into 5 parts each, and M_0 into 25 parts. (In fact, as we shall see later, the coarseness of the division of M_0 depends upon n; larger n will require a greater division of M_0. But for now, we will divide M_0's domain of $[0, 2\pi]$ into 25 parts.

3. At each lattice point, we will fix the three parameters of the lattice and use the Nelder-Mead optimization algorithm to find the maximum of ℓ relative to all the other parameters, including P. (The starting point of the algorithm will be the P from the periodogram spike).

4. The resulting maximum for every point in the lattice will now be used as a *new* starting point for the Nelder-Mead algorithm, but this time freeing up all the parameters, including those that were in the lattice to begin with. A new set of local maxima will be found, this time maximal in the full 7-dimensional parameter space.

5. Almost certainly, many of the points in the lattice will have led to the same local MLE. Indeed, it's possible that every single point in the lattice will lead to the same local MLE. (This is a good thing!) If multiple local MLE's are found, then since we're looking here for the global MLE, we only need to keep the local

maximum with the highest likelihood. (But note: in the next chapter of this dissertation, we will want to recycle this algorithm, except rather than keeping only the local MLE with the highest likelihood, we will want to temporarily hold onto all local MLE's that are within a certain factor—say, one percent—of the greatest. The reader will be reminded of that at the time. I mention it here to remark that although right now we're only looking for the global MLE, later those other local maxima in ℓ will come in handy—and we have a way to find them.)

6. These steps will be repeated for each spike in the periodogram until it is determined that all relevant local MLE's have been found.

Recall that figure 4.7 showed our simulated data set of 10 points according to four different moduli in time, one for each of the four tallest spikes in the periodogram. However, we were unable at that time to superimpose the best-fitting radial velocity models for each graph because we had not yet estimated the best values of the parameters. Now we'll apply the algorithm above and see what we get.

4.2.3 Returning to the simulated data

We earlier saw in figure 4.6 a periodogram corresponding to 10 simulated data points, having four spikes in it that appeared to correspond to candidate periods worth considering. We then looked at the data modulo each of the four periods, but in order to judge whether they remain viable candidates worth considering, it would be helpful to superimpose on each plot the radial velocity function corresponding to the local maximum in the likelihood function having a similar period, as obtained by the algorithm described in the previous section. We now will do that. Here were the results for each of the four periods.

- When we seeded the algorithm with the period $P = 7.148$ days, it resulted in

a single local MLE, having a period of $P = 7.0631$ days, slightly different from that suggested by the periodogram. The log likelihood at this local MLE is -24.5612.

- When we seeded the algorithm with the period 3.482 days, it resulted in a single local MLE, having a period of $P = 3.4799$ days, again slightly different from that suggested by the periodogram. The log likelihood at this local MLE is -34.812 and does not beat the other to become the global MLE. (Indeed, it is sufficiently unlikely relative to the first one found that it won't really remain a contender even in the last chapter of this dissertation, when we are integrating to find Bayes's factors, rather than just MLE's.)

- The third seed having period $P = 1.162$ days led to *56* local MLEs, all with periods roughly equal to 1.162 days, but dramatically different in character. Figure 4.13 shows radial velocity functions for four of these 56 solutions.

- Finally, the fourth seed having period $P = 2.100$ days led to 46 local MLEs, again with periods all about the same. Four of these solutions are shown in figure 4.14.

Now we compare the four periods once again. Figure 4.15 shows the data from the earlier section—ten simulated data points from an eccentric orbit—under these four moduli, and with superimposed curves corresponding to the local MLEs. In the case of the multiple solutions, the highest MLE is shown.

The one of these local MLE's that has the greatest likelihood—the first, with $P = 7.0631$ days—is so much more likely than the others, that even if we were attempting to integrate the likelihood function to find a marginal likelihood of M1 rather than a maximum likelihood, we wouldn't have to worry about those other solutions. And since in this chapter we're only concerned with the global MLE

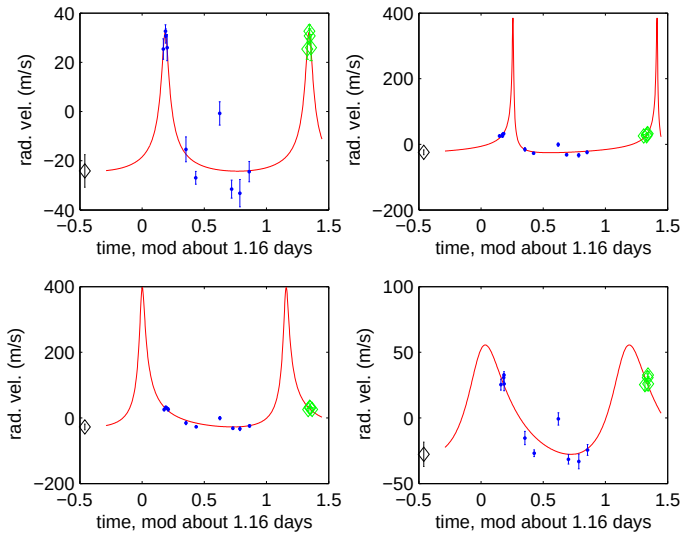

FIGURE 4.13: These four radial velocity functions—all shown with the same simulated data set—correspond to four of 56 different local MLEs that were found by the search algorithm when seeded with the period of $P = 1.162$ days, which was suggested by the third-tallest spike in the periodogram. The vertical bar towards the left of each plot, with a black diamond in the center, indicates the magnitude of the stellar jitter s estimated by the MLE, otherwise not visible in the plot. The four log likelihoods of these shown are, from top to bottom and left to right, -35.1413, -35.9722, -37.1419, and -37.5659. Thus, the most and least likely of these four differ by a likelihood ratio of only about 11.3. One would perhaps therefore consider all of roughly comparable plausibility, from a likelihood standpoint. The multiplicity of the solutions is not a grave problem when they all are so much less likely than the one whose period is about 7 days. However, it is not hard to see why the multiplicity of solutions will be a problem when the data are less conclusive—and in particular when there is no planet present at all and the data really are nothing more than noise. That will be addressed in the next section of this chapter.

anyway, then we can certainly toss those others out. But it is still useful—and satisfying—to compare the appearances of the best fits that the four periodogram spikes led us to. Although the four periodogram spikes looked to be of comparable height, the one that yielded the global MLE was in fact considerably more likely than all the others, by orders of magnitude.

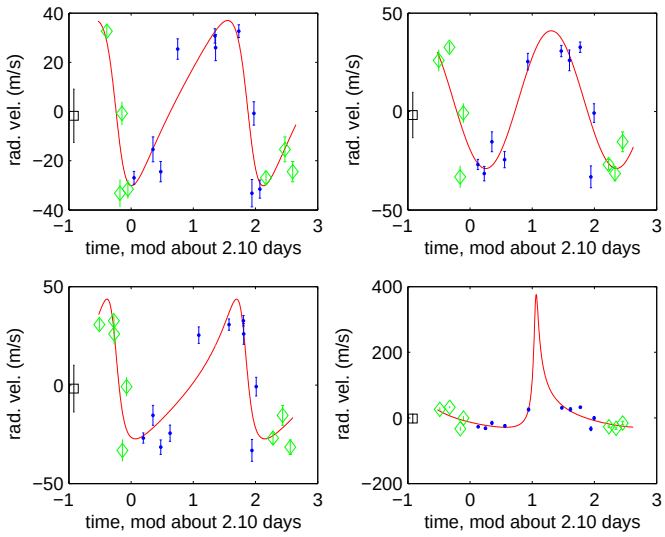

FIGURE 4.14: As in figure 4.13, we see here four dramatically different solutions, all corresponding to roughly the same period. These suggest that integrating the likelihood function over the parameter space will be challenging if the data do not single out one solution or a very few as serious contenders. In the case when there truly is no planet, this will of course not occur.

That suggests that we may not need to look at very many periodogram spikes before stopping, which will save time when implementing our algorithm. We'll proceed from the tallest to the next-tallest, and so on. But if we ever find a local MLE whose likelihood is considerably less than the greatest we've found so far—say, smaller by a factor of 1000, then we may quit searching, confident that nothing better will turn up later. What would have been worrisome would have been if lots of periodogram spikes produced local MLE's with comparable likelihoods. Then we would have to deal with the fact that our global MLE search algorithm was apt to take a long time, considering spike after spike after spike in the periodogram. But it seems that that may not actually occur, at least not when the data do support planetary presence.

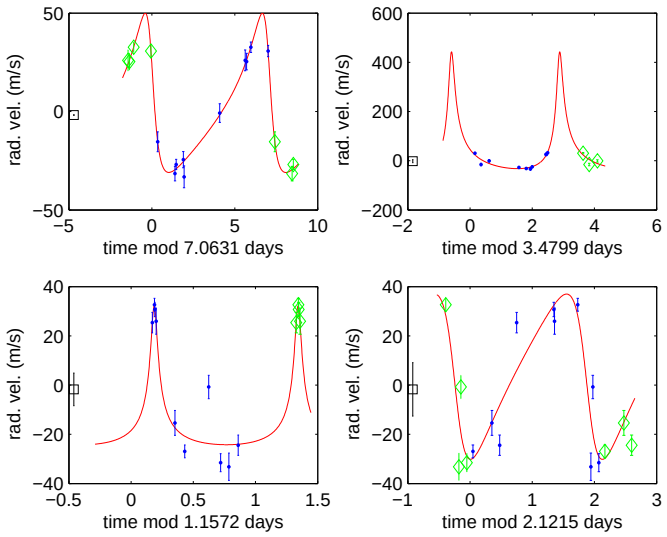

FIGURE 4.15: The data from the earlier section are now shown under the moduli of the four local MLE's, along with (superimposed) the radial velocity function corresponding to the MLE. Since that function does not itself display the MLE estimate of s, the stellar jitter, in any way, that is indicated in each plot by a vertical line segment at the left of the plot surrounding a black square. The 3.48 and 7.06 day peridos are very likely harmonics, reflecting eccentricity.

4.2.4 Finding the global maximum likelihood under M1 when the data provide no compelling evidence for M1

The search algorithm described in the previous section would appear to be reasonably good at finding the global MLE fairly quickly, so long as the data do in fact suggest the presence of a planet. For in that case, one period will stand out and the search won't have to go on very long. But of course, since we're trying to see how well the likelihood ratio Λ will perform even when the data do not greatly support model M1—perhaps because there is not really a planet there at all—then we must be prepared to use our algorithm in that non-compelling case as well. What is different then?

Two things are different: one for the worse, one for the better. The change that is for the worse is that we may expect spikes in the periodogram to correspond to

local MLE's that are of comparable likelihood, rather than one standing out. But the change for the better is that since the MLE's are of comparable likelihood, then if we settle for one that isn't actually the global MLE, we won't be too far off.

Let's put it another way. We are considering an algorithm that begins with the best candidate for global MLE and goes on to consider other candidates for global MLE repeatedly until it becomes clear that none will beat out the present winner. If the candidates become considerably less likely very quickly, then our algorithm won't have to run very long, and we may be reasonably confident that we've found the global MLE. On the other hand, if the candidates do not become considerably less likely very quickly, then our algorithm will have to run longer, but the local maximal likelihoods it turns up will be sufficiently similar to one another that if we terminate the algorithm before having actually found the global MLE, we will nevertheless have an estimate that's pretty close.

This leads us to a modification of our stopping rule. We said in the previous section that we would stop looking at periodogram spikes whenever the local maximal likelihood from one of them is many times (say, 1000 times) less likely than the current maximum. That may take too long, so an alternate stopping rule may be: if we look at a certain number of periodogram spikes of decreasing height (say, 10), and none of them produces a local MLE whose likelihood is higher than what we have presently found to be the greatest, then we may stop, reasonably confident that nothing much better is going to turn up. The justification for the 10 is this: if the periodogram spikes lead to local MLE's with very similar likelihoods, then if we have one that is not the global MLE but just an "ordinary" local MLE, then we may think of the others as each having a 0.5 chance of being greater than the one we have. The probability of 10 in a row going by that do not exceed the one we have is therefore 1 in 2^{10}, or about 1 in 1000—sufficiently improbable that should it occur, we may be confident that the one we have really is the global MLE—or at least

awfully close. And although the algorithm may occasionally lead us to an estimate of the global MLE that is incorrect, it should occur infrequently, so that eventual probability estimates we make regarding the strength of evidence of Λ should be pretty accurate.

4.3 Putting the two together: what is the distribution of Λ under M0?

So now we have a very fast algorithm for finding the maximum likelihood under M0 with certainty, and a less fast and slightly less certain algorithm for guessing at the maximum likelihood under M1—but one that should be acceptable. We'll now simulate data under M0, as we did in the previous Chapter, and for each data set we'll compute those two maximal likelihoods and compute Λ, so that we may get a picture of what its distribution looks like under M0. As in the previous chapter, we'll look at it for different sample sizes, to see how this makes a difference.

We recall from the beginning of this chapter that we anticipate a certain behavior. For large n, we think the distribution of $-2\log\Lambda$ should be approximately χ^2_5. If that is so, then we should find most of our Λ's to be between 1 and 1000 under M0, and our $\log\Lambda$'s to be between 0 and around 6 or 7. Figure 4.16 shows, for each of the 11 sample sizes we're considering, a histogram of 1,000 simulated log likelihood ratios. Superimposed on each one is the pdf $\chi^2_5(2x)$, rescaled vertically so that its area matches that of the histogram.

Also for comparison, we can look at the mean and variance of the simulated values, compared with what the mean and variance would be under the theoretical distribution $\chi^2_5(2x)$. Such a comparison is shown in Figure 4.17.

There are two things we may notice in the distributions that require some comment. One is that the distribution may be heading towards the theoretical limit, but so slowly that it would imprudent, even with $n = 400$ to treat the actual dis-

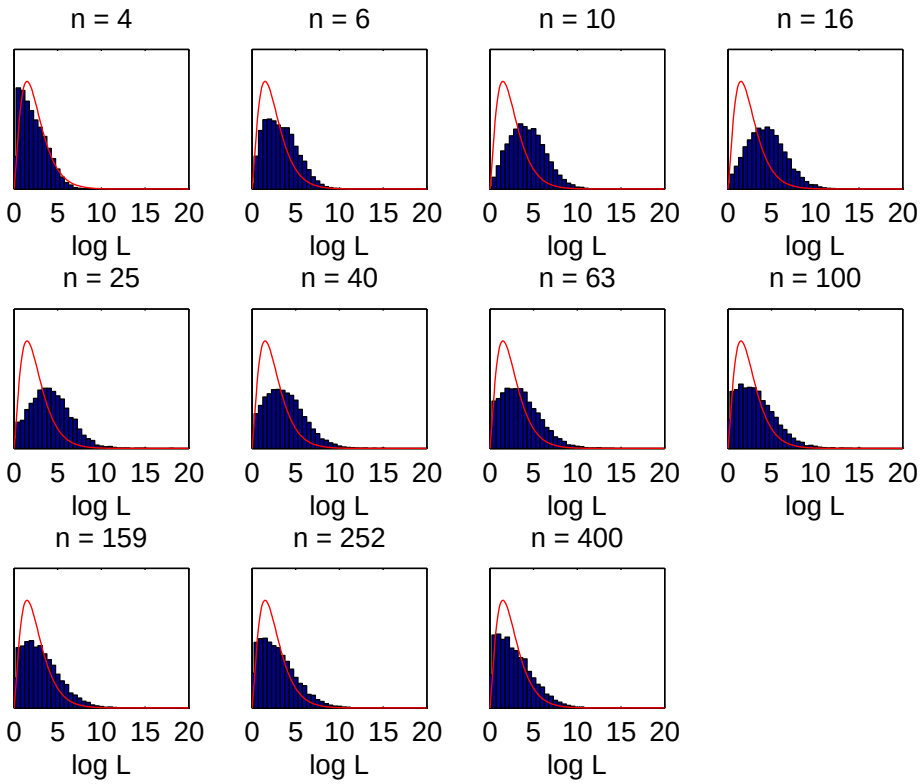

FIGURE 4.16: For each of 11 sample sizes, we look at the distribution of 10,000 simulated values of $\log \Lambda$ under model M0 (blue histogram). Superimposed on each histogram is the (vertically rescaled) function $\chi^2_5(2x)$ (red curve), which should in theory match the distribution of $\log \Lambda$ in the limit as $n \to \infty$.

tribution of $\log \Lambda$ as if it were $\chi^2_5(2x)$. The second is that the distribution of $\log \Lambda$ tends towards greater means and variances (and skew) for the first four sample sizes, and then "retreats" towards smaller values, apparently beginning to approach the anticipated limiting behavior.

There are two possible explanations for this that seem most plausible. One is simply that that's the way the likelihood ratio behaves. The other is that the search for the MLE in model M1 didn't find it. I would like to posit that the first is actually a reasonable explanation; here's why. The model M1 contains seven parameters. If

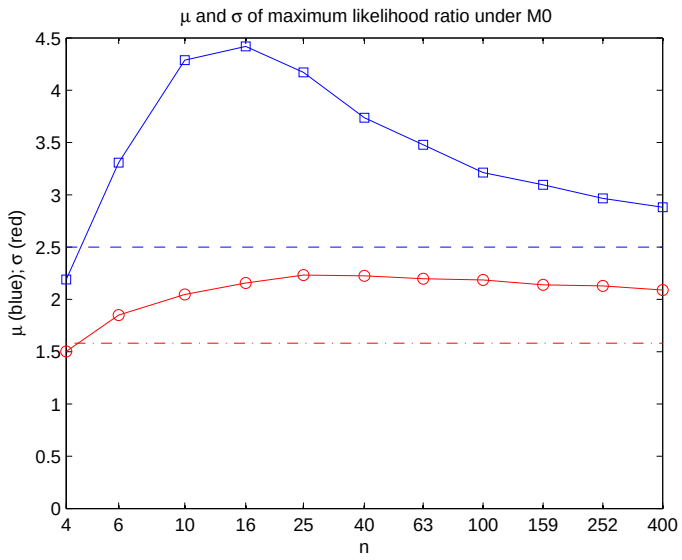

FIGURE 4.17: The mean and standard deviation of the simulated $\log \Lambda$'s under model M0 are plotted against sample size. The mean is represented by blue squares, the standard deviation by red circles. Also on the graph are horizontal lines at 2.5 and $\sqrt{2.5}$, the theoretical limiting values of μ and σ as $n \to \infty$.

you have nearly any four points or six points (t_i, V_i), it should be possible to find a radial velocity function that fits them exactly. But the two-parameter model M0 cannot be fit quite so well, and we would expect its fit to diminish with added data points, even when n is as small as 4. As n increases, however, even model M1 begins to have difficulty finding a good fit. So after 10 or 20 data points, the distribution of $\log \Lambda$ begins to "retreat" back towards smaller numbers. And those numbers are indeed reasonably consistent with the theoretical limit—enough to be satisfied that our result is believable, but not so much so that we may use $\chi^2_5(2x)$ as the distribution of $\log \Lambda$.

Since we do not feel satisfied that $2 \log \Lambda \sim \chi^2_5$ is a sufficiently good model, then how shall we model the distribution of $\log \Lambda$? The answer is that we don't need a parametric model at all. We just wanted to see how good the theoretical limiting distribution was, almost out of curiosity. The two analyses we used in the last

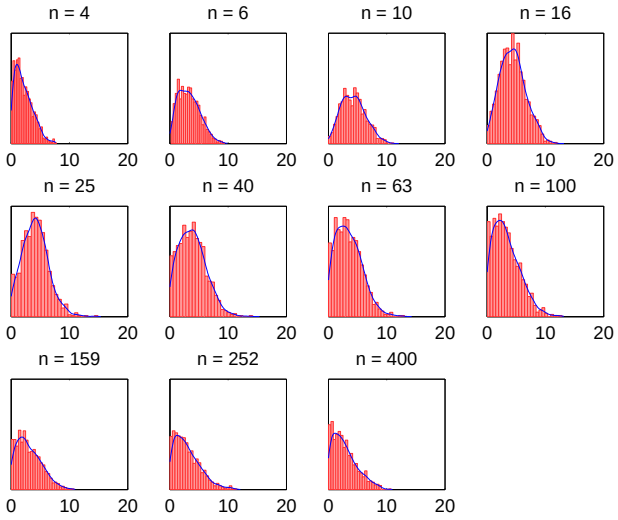

FIGURE 4.18: Distributions of $\log \Lambda$ under M0.

chapter—the IMDA plots and the "strength of evidence" plots—will work perfectly well here, and for the latter, we may again use the kernel density estimator as a non-parametric tool for estimating the density of $\log \Lambda$ under M0 and M1.

It would be nice, perhaps, to see how well the kernel density estimator performs at estimating these densities. Figure 4.18 shows the histograms of the simulated values of $\log \Lambda$ along with the superimposed kernel density function for each sample size under model M0. For comparison, we'll also look the scenario under M1 when $e = 0$ and $K = 7\text{ms}^{-1}$. Figure 4.19 shows those distributions.

Comparing the histograms in the two figures, we can see that, as was the case in the last chapter with the MPS, the distributions under M0 and M1 are similar for small n but very distinct for larger n. Thus, the statistic $\log \Lambda$ will have not be good at detecting a planet like this with few observations, but will do well when there are a large number of observations.

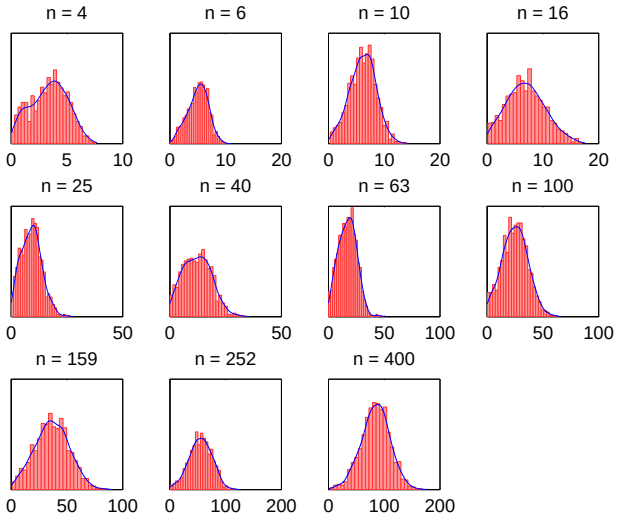

FIGURE 4.19: Distributions of $\log \Lambda$ under M1, with $e = 0$ and $K = 7\text{ms}^{-1}$.

4.4 Λ applied to the circular orbits

Now let's see what the IMDA plots for our circular orbits under model M1 look like when our decision tool is Λ, the maximum likelihood ratio. Recall first what they looked like using the MPS. That was at the end of the previous chapter and is shown again in Figure 4.20. Figure 4.21 shows how the IMDA plots look for Λ.

Something that immediately calls attention to itself is the fact that the ability to distinguish a planet appears to grow weaker with increasing n when n is small. How can more data provide less model-distinguishing ability than less data? In the case of Λ, it is entirely possible that Λ is more likely to mislead (falsely announcing non-planets) when the sample size is 6 or 10 than when it is 4. In all of these cases, a non-planet-hosting star can look somewhat like a planet-hosting star, simply because some period may be found under which the data appear to have a pattern. That grows less likely with much larger n. But with relatively small n's, it is not only possible, but indeed quite *likely*. It doesn't mean that less data is better than more

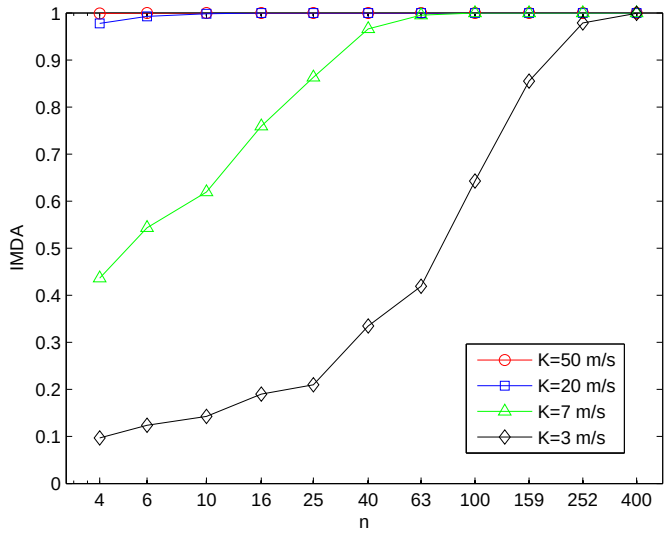

FIGURE 4.20: IMDA plots for the statistic MPS, discussed in the preceding chapter. The four plots correspond to—from top to bottom—$K = 50$ ms^{-1}, $K = 20$ ms^{-1}, $K = 7$ ms^{-1}, and $K = 3$ ms^{-1}.

data, only that the maximum likelihood ratio Λ should not be greatly trusted when n is relatively small.

Comparing these two plots reveals that the maximum likelihood ratio is not—by the IMDA standard—superior to the maximal periodogram spike (MPS) when it comes to discerning planets. It is at best comparably good for middle-sized values of K, but is worse for larger K's (where the decision is relatively easy regardless of what statistic you use–or even if you use none at all) and is also, somewhat surprisingly, worse for smaller K's as well. It would appear that when the stellar wobble is little more than a mere wiggle, comparable in magnitude to the measurement errors and the stellar jitter, the maximum likelihood ratio statistic simply does not excel at discerning the planetary presence. Again, that is by the present standard of IMDA, which is more or less a measure of the overlap between the distributions of Λ under M0 and M1. Many common Λs from a star hosting a small planet are apparently also fairly typical from a star hosting no planet at all.

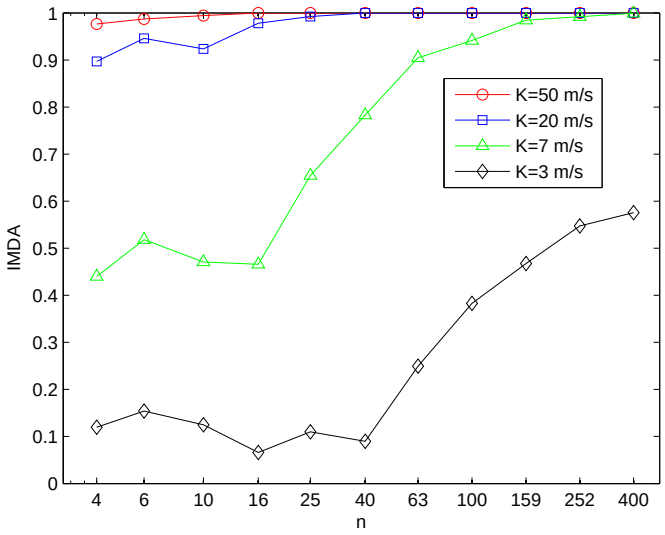

FIGURE 4.21: IMDA plots for the maximum likelihood ratio Λ, discussed in the present chapter. The four plots correspond to—from top to bottom—$K = 50$ ms^{-1}, $K = 20$ ms^{-1}, $K = 7$ ms^{-1}, and $K = 3$ ms^{-1}. Although is a peculiar rise in the IMDA for $K = 3$ms^{-1} at $n = 25$, this may be an anomaly, due to having done only 1000 simulations in each case. However, it cannot be a mere fluke chance that three of these IMDA plots decline before they increase again.

One possible explanation for this is that the algorithm for finding the MLE didn't actually find it when the planet was very small. Certainly the algorithm has a harder time in such a case. But as we have seen, the likelihood function is so very multimodal that any search algorithm needs a helping hand in finding local maxima that are reasonable candidates for the global maximum. And when K is very small, that search becomes exceedingly difficult, even when periods are suggested by the periodogram. There is so much noise drowning out the signal that the signal is very, very, hard to find. Or perhaps the problem is simply that the maximum likelihood ratio is truly not a superior tool for finding planets than is the maximal periodogram spike.

Either way, we have learned something: the present practice of using the periodogram to look for planets is not improved upon easily by using the maximum

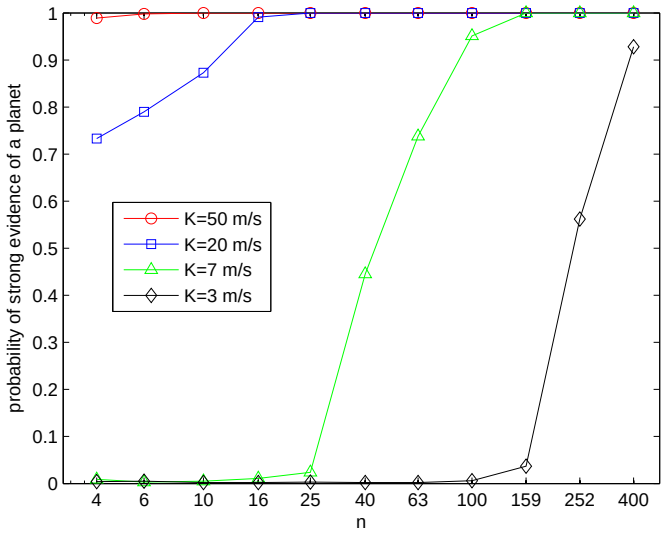

FIGURE 4.22: Strength-of-evidence plots for the MPS. Recall that these plots show the probability of obtaining a probability density ratio greater than 32 favoring model M1.

likelihood ratio instead.

What about the "strength-of-evidence" plots? Recall from the end of the previous chapter what these looked like using the MPS. They're shown in Figure 4.22.

Figure 4.23 shows the strength-of-evidence plot for the same planets when the statistic being used is Λ.

They're plots are pretty, and oddly, don't decline and rise again as did the IMDA plots. But they agree with the IMDA plots in one regard: for no sample size is Λ a better planet-detecting tool than the MPS.

4.5 Λ applied to the eccentric orbits

In Figure 4.24 we recall the IMDA plots for the three cases we considered in the last chapter, with $e = 0$, $e = 0.3$, and $e = 0.6$.

Now in Figure 4.25 we look at what the IMDA plots look like for Λ under the same three scenarios.

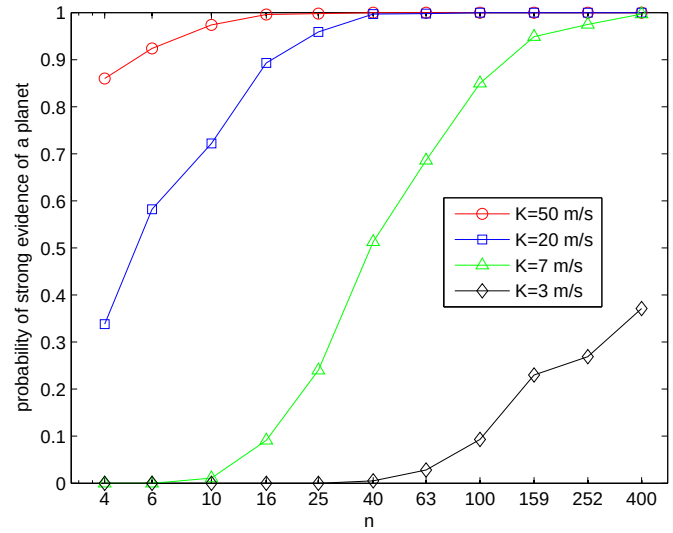

FIGURE 4.23: Strength-of-evidence plots for Λ. Interestingly, these do *not* show the decline and then increase that was apparent in the IMDA plots.

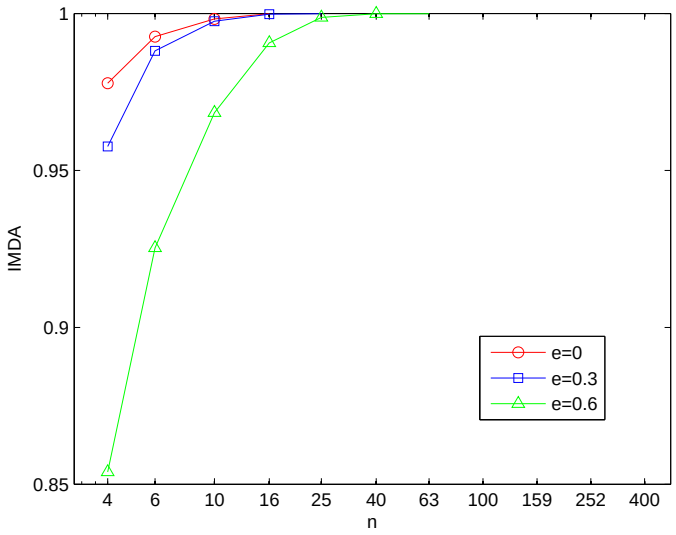

FIGURE 4.24: IMDA plots for the MPS, eccentric orbits.

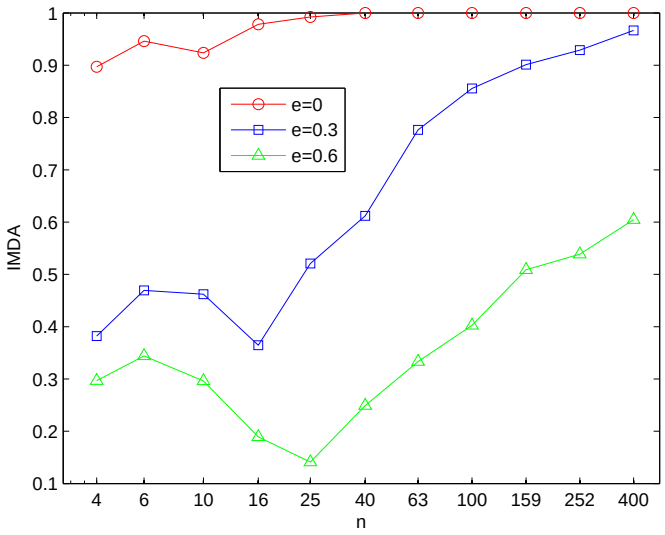

FIGURE 4.25: IMDA plots for Λ, eccentric orbits.

One of the remarkable features of these plots is the pronounced initial decline as n increases. Once again, this could be due to the fact that until there are a sufficient number of data points, it's just not that hard to fit a seven-parameter function to a set of data that are actually mere noise. The flexibility that all those parameters allows makes it so much easier to fit a good model under M1 than under M0 that the additional data observations serve merely to punish model M0 while rewarding model M1.

Another remarkable feature of the plots is how much lower they are than the IMDA plots for the MPS. As with circular orbits, the MPS is a superior planet-detecting tool over Λ. Figures 4.26 and 4.27 show the strength-of-evidence plots for the MPS and Λ, and we see the same conclusion: the MPS is considerably more likely to provide strong evidence for the presence of a planet than is Λ.

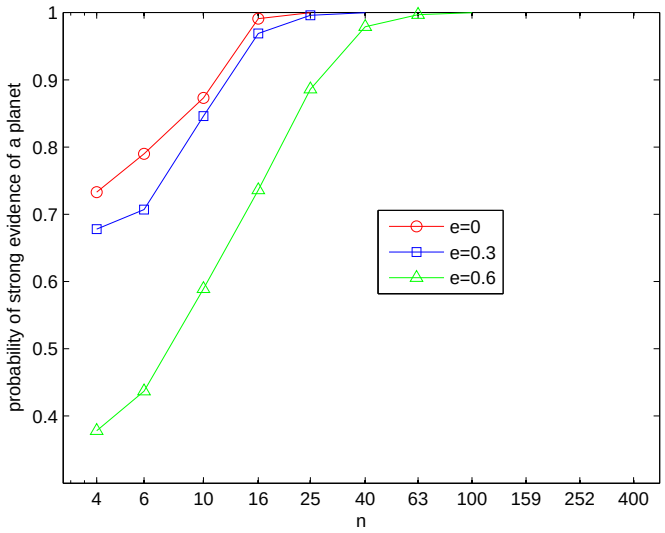

FIGURE 4.26: Strength-of-evidence for MPS, eccentric orbits.

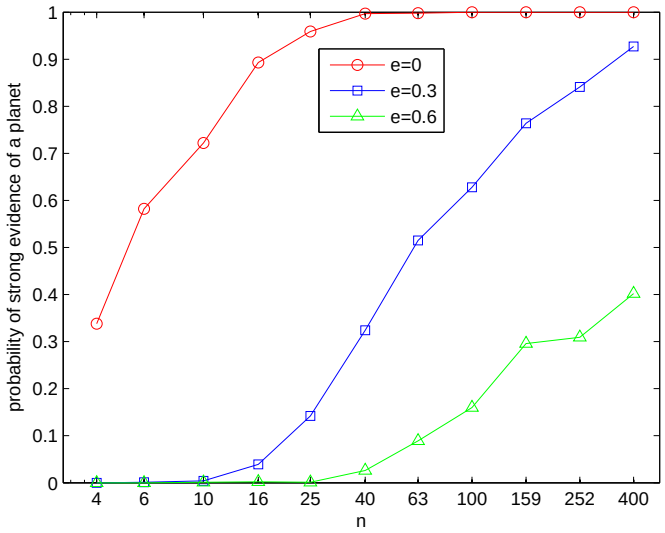

FIGURE 4.27: strength-of-evidence plots for Λ, eccentric orbits.

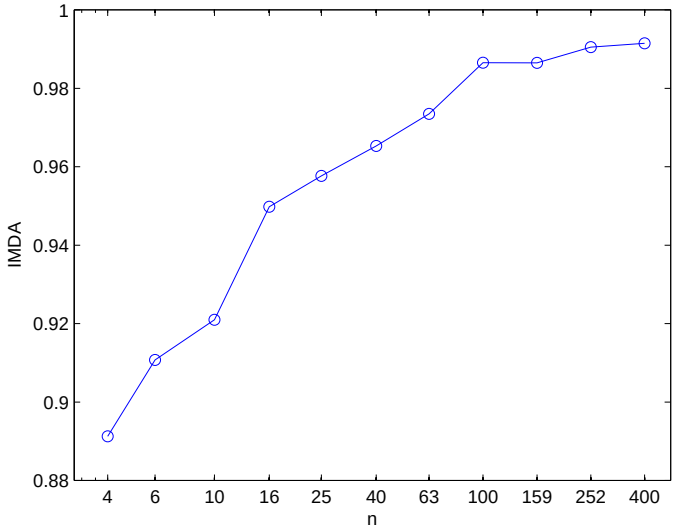

FIGURE 4.28: IMDA plot for MPS, stars mimicking known planets

4.6 Already detected planets

As in the last section, we'll now compare the performance of the MPS to Λ when faced with stars mimicking those hosting known planets. Figures 4.28 and 4.29 compare the IMDA under this scenario; Figures 4.30 and 4.31 compare the strength-of-evidence plots.

These plots seem to indicate that Λ is a very poor tool for planet detection. However, they imply such improbably poor performance of Λ that a more reasonable explanation for the apparent poor performance of Λ throughout this chapter—especially with eccentric orbits—is not that Λ is really that poor at detecting planets, but rather, that the algorithm described in this chapter for finding the MLE for the one-planet model is not always, or perhaps even often, finding the actual MLE. It is hard to say where the problem lies, but I suspect that the most likely culprit is the step where a three-dimensional lattice was created on the three parameters s, M_0, and e. If that lattice is too coarse, then it may miss the MLE. I used 25 divisions of the interval $[0, 2\pi)$ for the parameter M_0. However, the likelihood function is quite

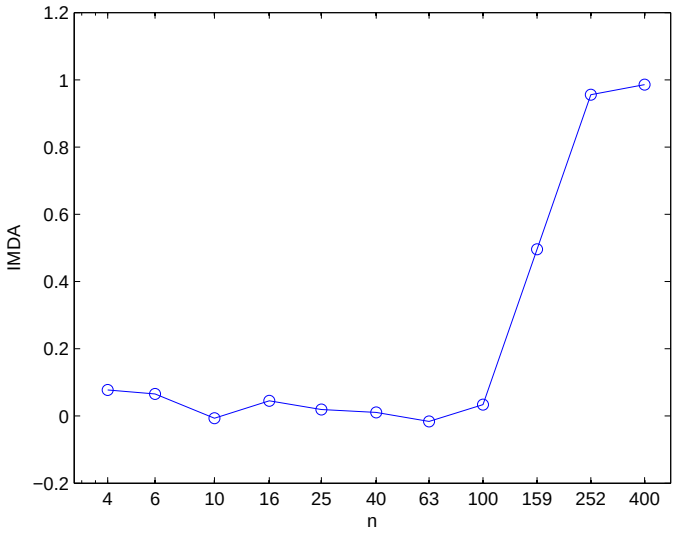

FIGURE 4.29: IMDA plot for Λ, stars mimicking known planets

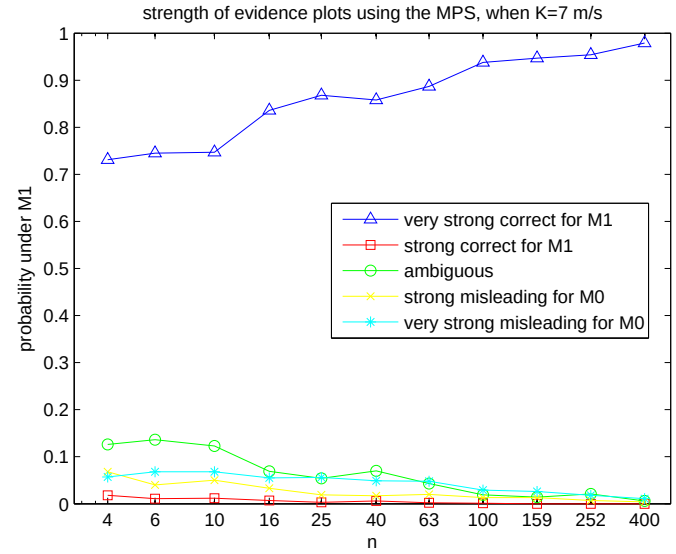

FIGURE 4.30: strength-of-evidence plot for MPS, stars mimicking known planets

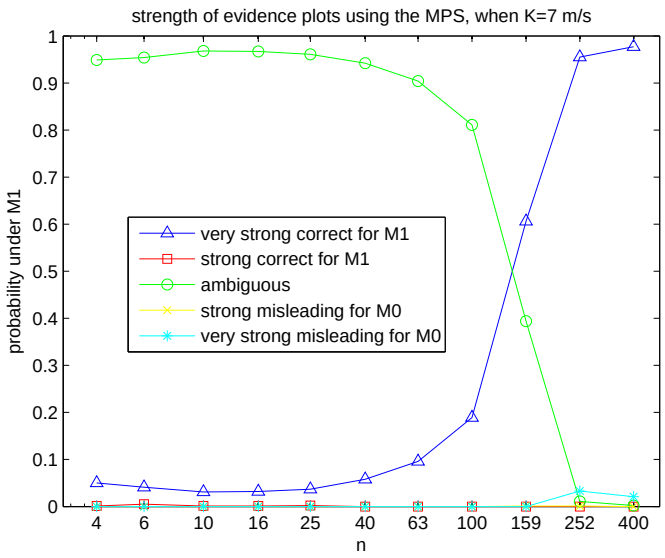

FIGURE 4.31: strength-of-evidence plot for Λ, stars mimicking known planets

sensitive to changes in M_0, especially when $e > 0$, so it's very possible that the actual MLE of the parameters is routinely "slipping through the cracks" of the lattice structure, so that even the Nelder-Mead search algorithm that began at every point in the lattice was still failing to make its way to the true local MLE. If the MLE estimate of M_0 is off by even a fairly small fraction of a radian, then the resulting likelihood estimate would be fairly arbitrary, because a miss is as good as a mile. The resulting likelihood ratio Λ would therefore be comparable to that under the model M0, which would explain its apparently poor planet-detection ability.

I would be lying if I said this was not disappointing; of course it is. But it has successfully addressed one of the primary research questions in this dissertation. We wanted to see if there was some statistic that was at least as good at planet-detection as the height of the maximal periodogram spike, *and that was also relatively easy to estimate*. Here we have shown that while the maximum likelihood ratio may be have comparable planet-detection power to the MPS (it appears to be comparable for circular orbits, when finding the MLE in model M1 is easier), it is not easy to

estimate.

Of course, in practice one wouldn't need to estimate Λ thousands of times, but only once for each data set. However, interpreting the resulting observed Λ requires having some idea of what its distribution is like under different models. That remains unknown.

5

The Bayes Factor

We have considered the maximal periodogram spike (MPS) as a tool for planetary detection (Chapter 3), and also the maximum likelihood ratio (Chapter 4). In this chapter, we consider the Bayes factor as a tool for addressing the same questions. (Is a planet present? How many observations will it take to tell whether a planet of this type is present?)

However, we have a problem. We saw in the previous chapter that the MLE for the one-planet model was sometimes very hard to find—indeed, an algorithm that pushed the limits of what was computationally possible was still apparently unable to find the MLE for the full seven-parameter one-planet model consistently. That was probably what led to the MLE being so poor at planet detection. If the MLE is hard to find, then the Bayes Factor will be harder still.

Estimating the Bayes Factor requires integrating out the parameters in model M0 and model M1. The former is not hard to do. It cannot be done easily analytically due to the way s appears in the likelihood function, but the likelihood function, and hence the posterior density as well, are both unimodal and defined

over a two-dimensional space. So it is not hard to estimate the integral using importance sampling.

However, integrating out the parameters of the one-planet model is exceedingly difficult, and I will tell the reader now that this chapter will not succeed in estimating the Bayes Factor sufficiently well to compare it properly to the MPS and the maximum likelihood ratio. In a later chapter in the appendix, Chapter B, we will spend great effort trying to see whether a creative approach called "nested sampling" will work. (It won't.) In this present chapter we will discuss efforts to estimate the integral using importance sampling. However, in the end, we will reach the unhappy conclusion that the Bayes Factor is so difficult to estimate even once for one real problem, and so much more difficult to estimate multiple times in a simulation setting, that it is not the tool we were searching for. Thus, the essential conclusion of this dissertation may be stated here and now: the maximal periodogram spike, or MPS, is far, far easier to compute than either the maximum likelihood ratio or the Bayes factor, and it appears to do sufficiently well at detecting planets, even when the orbits are eccentric, that small planets may be discernable with that tool given enough observations.

The remainder of this chapter will therefore be spent explaining the problems that I encountered when trying to estimate the marginal likelihood of model M1.

5.1 Integrating the parameters out of M0

Let us consider model M0, which only has two unknown parameters: s and C. Recall that the prior on these parameters is given by

$$\pi_0(C, s) = \kappa_0 \cdot \frac{1}{s_0 + s},$$

where

$$\kappa_0 = \frac{1}{C_{max} - C_{min}} \cdot \frac{1}{\log\left(1 + \frac{s_{max}}{s_0}\right)}$$

and where $C_{min} \leq C \leq C_{max}$, $0 < s \leq s_{max}$. Also require that the data are Gaussian with mean C and with variance $s^2 + \sigma_i^2$. Because C_{min} and C_{max} are so large in magnitude relative to the measured radial velocities, the posterior distribution of the parameters is approximately Gaussian in C. However, although it is unimodal in s, a closed form is not apparent, due to the peculiar location of s in the likelihood function. So importance sampling would appear to be a good tool for integrating out the parameters. Furthermore, since the precision of importance sampling diminishes relatively little with increasing dimensionality (assuming that the importance sampling function is a reasonably good one) then I will use an importance sampler that draws from both dimensions, C and s.

I will use a product of univariate t distributions with four degrees of freedom (overly heavy tails are far preferable to overly light tails) for the importance sampler and ignore possible correlations between C and s. (See below.) I will use the Nelder-Mead algorithm to find the location of the maximum value of the posterior density function, and use that as the center of each t distribution. C and s must have some correlation in their joint posterior distribution, because if C escapes the range of the data, becoming particularly large in magnitude in either the positive or the negative direction, then the deviations in the data from C must be explained with a larger s. However, despite that anticipated correlation, it is worth trying doing the importance sampler with no correlation in t, at least to see if it works reasonably well, since that is so much easier than trying to get the correlation right. If the sampler works poorly, we can return to re-examine the correlation structure. But if it works reasonably well, then we can press on.

Because s can be very close to zero, it makes more sense to work in the space of

$\log s$ than s, and the prior is transformed appropriately. (See the transformation in Chapter 1.)

Even ignoring correlation structure, we must still estimate the scale of each t distribution. Here is the algorithm I propose:

1. Beginning at the point of maximal posterior density, move along the C axis in both directions until the density becomes, in either direction, less than 1% of its value at the maximum. Note the distance traveled.

2. The $t_{\nu=2}$ distribution reaches 1% of its maximum height at a distance of about 2.7 from the point (0,0). Setting that distance equal to the distance traveled in step 1 gives a "stretch factor" to be used for the parameter C.

3. Repeat the same two steps for $\log s$.

Using the product of these two univariate $t_{\nu=4}$ distributions for the importance sampler, we will make 1000 draws, using each one to obtain an unbiased estimate of the desired integral. The "effective sample size" (ESS) is defined to be the number of estimates divided by one plus the variance of the standardized estimates, where the standardized estimates are the estimates divided by their common mean.

$$ESS = \frac{n}{\sum 1 + Var(\frac{Y_j}{\bar{Y}})}, \tag{5.1}$$

where Y_i are the integral estimates. If the ESS is close to 1000, or even if it is greater than 200 or so, then the efficiency is sufficient to estimate the integral with reasonably good precision. On the other hand, if thee ESS is quite low, then it suggests revisiting the chosen sampling function, perhaps taking into account correlations between the two parameters.

Happily, this algorithm works fairly well and routinely estimates the integrals with an efficiency of around 50%. Figure 5.1 shows the distributions of the marginal

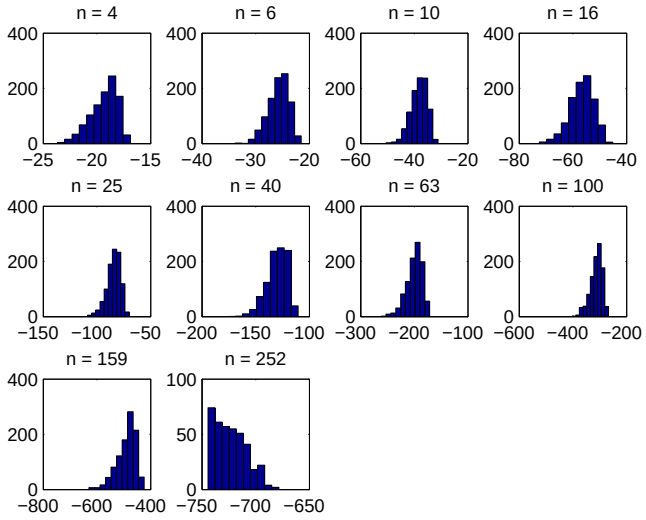

FIGURE 5.1: Distributions of log marginal likelihoods of M0, based on 1000 simulations of noise (no planet present). At the sample size of $n = 252$, only 385 of 1000 simulations produced integral estimates greater than zero, and at $n = 400$, no integral estimates were greater than zero. This is because the importance sampler requires averaging together integral estimates, not log integral estimates, and the integral estimates were themselves tiny beyond the capacity of the computer to handle them.

likelihood of M0 when stellar noise is simulated with the different sample sizes under consideration in this paper.

Figure 5.2 shows the distributions of the effective sample sizes for the same simulations.

The ESS for all of these importance computations was around, or was greater than, 500, indicating an efficiency of at least 50%. This is quite acceptable for an importance sampler. It is interesting to notice the bimodality of the efficiency, with one group of simulations running at about 50% efficiency for all sample sizes, and another running at an efficiency that improves with increasing n. It is not clear to me what the explanation for this is.

These integrals were all performed on data that were mere noise. It is worth

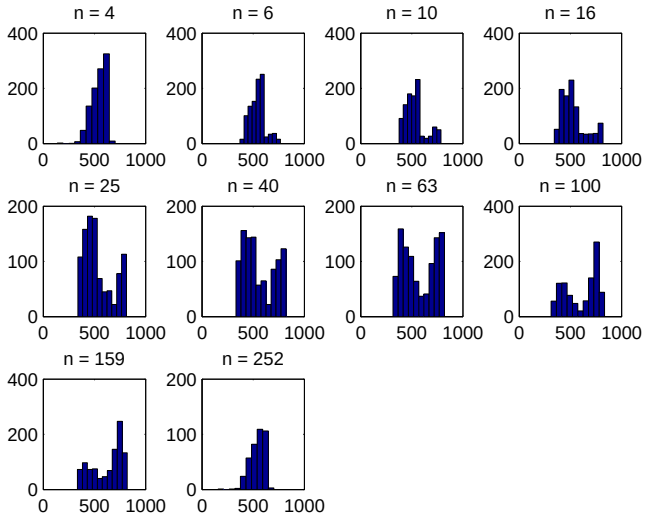

FIGURE 5.2: Distributions of ESS, based on 1000 simulations of noise (no planet present). The effective sample size is consistently around or above 500, indicating an efficiency of 50%, which is quite acceptable.

asking whether this integral-estimating algorithm would work equally well when the data were from a star hosting a planet. In principle, that should be the case, since this algorithm pays no attention to patterns or even the time-ordered nature of the data at all. However, when there is a planet present, especially if K is large, then there is greater variability in the radial velocities than when there is no planet present. That would make the posterior distribution of $\log s$ tend towards larger numbers, so it is worth seeing whether the algorithm would still work.

Figures 5.3 and 5.4 show the same distributions as before, only this time the data were simulated as coming from a star with a circular orbit and $K = 20\text{ms}^{-1}$.

We see by comparing these distributions to those generated from the no-planet star that even without computing the Bayes factor, just computing the marginal likelihood of M0, a planet like this would be very detectable after $n = 16$ or so observations, as was the case using the MPS.

But we are not interested in that statistic. We are interested in the Bayes factor.

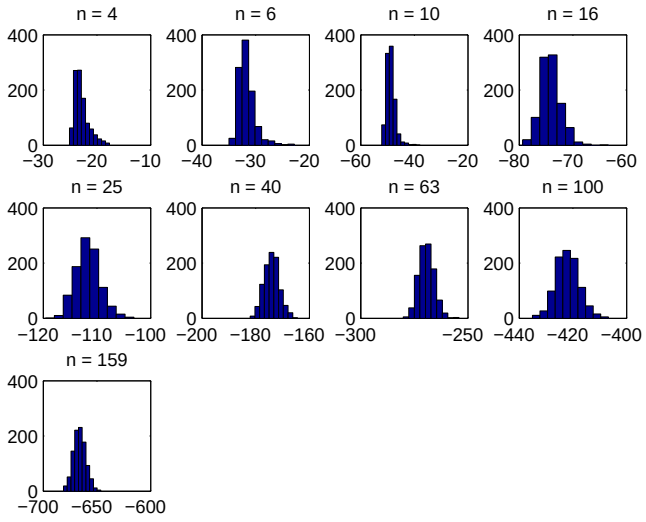

FIGURE 5.3: As with the simulated noise, these estimates are believable and acceptable.

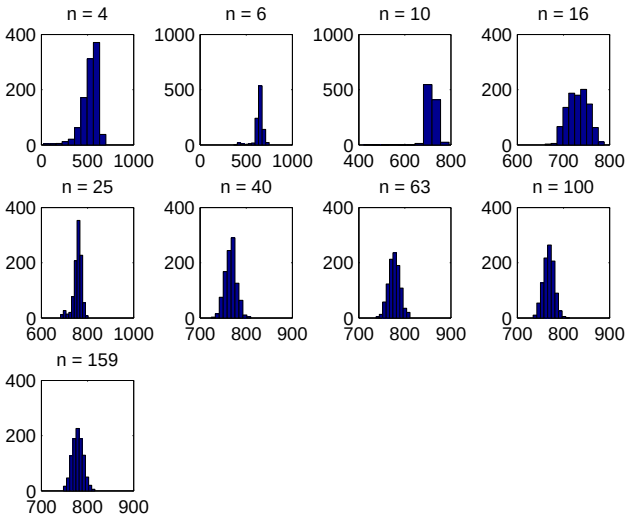

FIGURE 5.4: These distributions are not bimodal, but simply grow more efficient as the sample size increases.

And to compute that, we need an estimate of the marginal likelihood under model M1.

5.2 Integrating the parameters out of M1

We saw in the last section that integrating the parameters out of model M0 worked fairly well for sample sizes up to $n = 159$ using importance sampling. And beyond that, it apparently wasn't the algorithm that failed, but the technology's capacity to store numbers whose logs were less than -750 or so.

So we will try the same approach for model M1. First, we will try to find the location of the maximum posterior density. (This was easy under model M0, since the density function was unimodal—not so easy here!) Then we will move away from that point laterally along each parameter's axis separately until a point is found where the density is 1% of the maximum. That distance will correspond to a distance of 2.7 units in the $t_{\nu=4}$ distribution. Correlations will be ignored.

(Incidentally, lest the reader think that the same covariance matrix is being used across the board, that is not the case. These "stretch factors"—essentially variance estimates—are estimated individually for each parameter and separately for every data set, customizing the importance sampler to the data set.)

Now there are numerous problems we may run into trying to integrate parameters out of the multimodal space of M1 that we did not encounter in the last section when the space was unimodal. These include the following:

- The more parameters there are, the greater chance there is of correlations between some of them in the posterior density function. However, we are using the transformed parameters—and appropriately transformed prior—which greatly reduces those correlations.

- Since the space is multimodal, it's possible that when using this method for

estimating the scale needed on one or more of the parameters, we may "walk away" from the local maximum density point, but begin climbing another local hill in the density function before we've fallen to 1% of the maximum density where we began. This is a problem that is not tackled in this dissertation, though one can imagine using a greater threshold of 5% or 10% to reduce the possibility of this happening.

- This algorithm requires knowing where the point of maximal posterior density, or at least the MLE, is. Yet we saw in the last chapter that this point can be extremely difficult to find, especially when an orbit is eccentric, and especially when K is small—in other words, especially in the most interesting cases.

- It is very possible–especially in the case when there is not really a planet present, or when there is only a small planet present—that apart from one dominant mode in the posterior, there will be other pockets of non-negligible probability. I attempted to use mixtures of t distributions as an importance sampling function to address this possibility, but was not successful. The code I wrote to incorporate multiple modes churned forever without ever stopping. Recall from page 85 that even when a planet is present, a single promising candidate period P may have in its nearby vicinity dozens of very distinct locally optimal solutions, each with comparable likelihood. That is almost certainly even more likely when a planet is not actually present. I also was not able to solve this problem in preparing this dissertation.

The last of these problems is the one I conjecture is the most serious. I originally planned to simulate at least a thousand Bayes factors for each of the planet types under consideration in this dissertation, so as to see whether it could detect planets using fewer observations than could the maximal periodogram spike. However, if the prior distributions are reasonable, then of course a single Bayes factor can alone be

an inference tool, with no distributional properties needed. A Bayes factor of, say, 200, would provide very convincing evidence of the presence of a planet.

Yet this problem of multiple modes only goes away when a single pocket of probability dominates the posterior density function, and when that happens, then planetary detection isn't an issue any more. It's a Catch-22: if only there were a stand-out pocket of probability density in the posterior, then we'd be able to estimate the Bayes factor. But if that were the case, then we wouldn't need the Bayes factor at all. It's most useful when it's least obtainable.

So. I have applied the above algorithm, and when there is no planet present, the typical effective sample sizes are less than 2, indicating that there is one point drawn from the importance sampler that carries nearly all the probability mass estimation power among the 1000 draws. This happens not because some fluke point was drawn, but because the importance sampling function simply doesn't have the same shape as the posterior density function. The distribution of the estimates will always be strongly skewed because so many of the draws come from regions where the density of the importance sampler is low but the density of the posterior is (relatively) high. The solution might seem to stretch the importance sampler out, but if the posterior is highly multimodal (and it is) then there will always be little pockets of probability further out... and further out... and further out. Importance sampling with a unimodal sampler simply cannot effectively estimate the integral when the integrand is so highly multimodal.

5.3 Unhappy ending

And so at this point, gentle reader, we reach the unhappy conclusion of this chapter: It is only easy to estimate the Bayes factor for cases when the presence of planet is not only obvious, but even its orbital characteristics are well-known. This is less useful than the MPS, which is able to detect a planet's presence with as few as three

or four data points, simply by virtue of the magnitude of K, even when the orbital characteristics are completely unknown.

6

No measurement errors

Although the conclusion of the last chapter was an unhappy one, at least from the point of view of this dissertation writer, here is a happy addition to the dissertation: a bonus, if you will.

We have up to this point found that the maximal periodogram spike is easy to compute and is a reasonably good tool for detecting planets. We have also found how many observations are required to detect planets of different types using this tool.

Since that tool has turned out to be useful and relatively easy to compute, we now use it to address a question that may be of interest to astronomers that has not been addressed yet in this paper. So far we have run all simulations assuming that measurement errors have a distribution that is $\mathsf{Gam}(8, 0.5)$, which approximates reasonably well the magnitudes of the measurement errors published so far. However, they are declining as technology improves. And one often hears in the media that smaller and smaller planets will become detectable as the measurement errors continue to decrease. However, this raises the question: is there a limit to how detectable a planet will be if the measurement errors continue to decline? And the

answer is both Yes and No.

The first answer is: Yes, there must be a limit, because there will always be stellar jitter masking stellar wobble due to an orbiting planet. And in this section we'll look briefly at the question of how many observations would be required to detect planets of the types discussed in this paper, if there were no measurement errors at all. That will be easy to do: we'll simply repeat the simulations that we did earlier, only this time using $\sigma_i = 0$ for all observations.

The other answer is: No, there is no limit to how small the detectable planets could be, at least supposing that the measurement errors could be made arbitrarily small. That is because stellar jitter, while a source of noise in radial velocity measurements, is not of equal magnitude for all stars. Some stars have almost no stellar jitter. So if several radial velocity measurements were made on a star, for example, and they were all observed to be *remarkably* close together, then this would likely be a star with very small inherent stellar jitter, and as such, it might be a good place to look for a particularly small planet. There are also stellar properties that are associated with larger and smaller stellar jitters, so these could also indicate promising hunting grounds.

But that answer supposes that you've happened upon a star with very small stellar jitter. If you don't know whether the perturbations you see are due to stellar jitter or perhaps to a planet, then it still becomes worthwhile to ask, "How many observations would we require before we could be pretty convinced there was a planet here, using the periodogram spike as a detection tool, and assuming that our measurement errors are negligibly small?"

We will not address that question in this section for all the types of orbits we saw before, nor using both analytical tools we used before. We will simply consider, for the sake of curiosity, the four circular orbits of different semi-amplitudes K, and we'll compare them using the strength-of-evidence plots. Recall from the very first

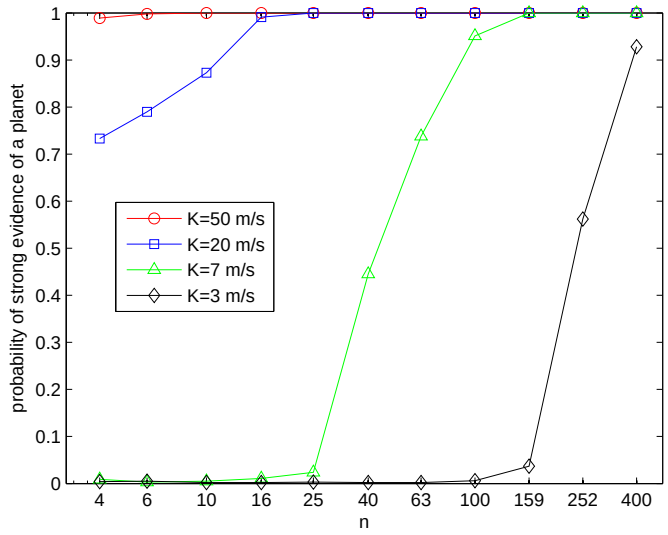

FIGURE 6.1: Strength-of-evidence plots for the MPS for the four circular orbits under consideration.

chapter what the plots looked like using the MPS when there were measurement errors as well as stellar jitter (Figure 6.1).

Now in Figure 6.2 we see what these look like when there is stellar jitter (each star having a stellar jitter drawn separately from a $\mathsf{Gam}(3, 1)$ distribution).

From these figures we see that small planets would require about half as many observations to detect if measurement errors could be reduced to zero, but stellar jitter was still possibly present.

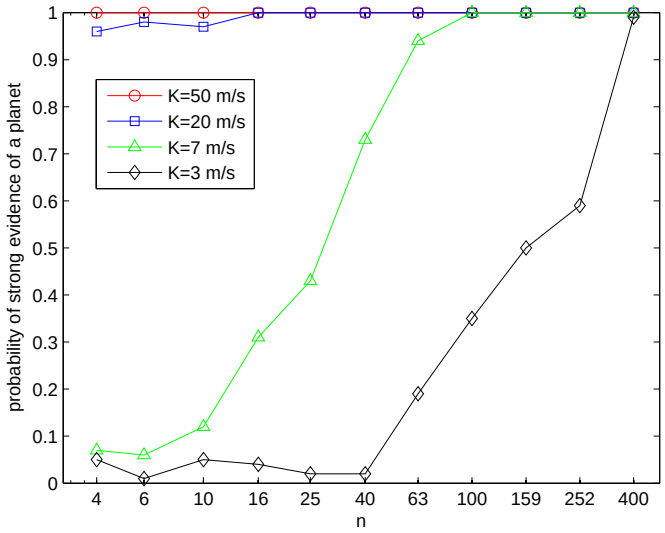

FIGURE 6.2: Strength-of-evidence plots for the MPS, no measurement errors.

7

Conclusion

This concludes our excursion into exoplanet detection. The more challenging statistics to estimate—the maximum likelihood ratio and the Bayes factor—proved to be sufficiently hard to estimate due to the multimodality of the likelihood function under model M1, and insufficiently superior to the maximal periodogram spike in planet-detection power (indeed, not superior at all, as estimated here, in the case of the maximum likelihood ratio) that the present practice of using the maximum periodogram spike as a tool for determining when planetary presence is promising, may be not only the simplest, but also the best tool there is.

Additionally, we found that if a very small planet is present—for example, one inducing an orbit on its host star with a semiamplitude of $K = 3\text{ms}^{-1}$—then it should be detectable using the maximal periodogram spike by the time 400 observations are made, and probably sooner.[1] This makes it possible to know in advance how many observations one must be prepared to make in order to detect a planet of a certain size. One can definitively say after a certain number of observations that if it were

[1] But see the footnote on page 62 regarding systematic error. Assuming that errors decline at a rate of $\frac{1}{\sqrt{n}}$ and that there is no systematic error is not realistic.

there, we'd have seen it by now. One can also be sure that if one quits prior to making a certain number of observations, then one is not likely to detect a planet of a certain type, even if it is there.

As measurement errors continue to decrease, smaller and smaller planets will be detectable. And if one is lucky enough to make observations on a star that has a very small stellar jitter, then it may eventually be therefore possible to detect a very small planet indeed, perhaps even Earth-sized. However, if measurement errors are reduced to near zero, one can still not say definitively that a planet of a certain size is *not* present, until one has made a certain (estimable) number of observations and still seen nothing; stellar jitter may be drowning out the planet's signal.

The Maximum Likelihood Ratio and Bayes Factor statistics are hard to estimate because the likelihood function is so multimodal, and they are hardest to estimate when the planetary signal is weakest—in other words, when we are most interested in knowing what the statistics have to say, they're the most difficult to estimate, and they may even be impossible to estimate. But that of course doesn't make them useless altogether. It just means that they may not be the best tool to use for model selection. There are many other questions of interest in exoplanet research for which they may be well-suited, such as orbital parameter estimation or adaptive design. And of course, there remains the possibility that someone clever will develop a very good means of estimating these statistics, even when a planet is so quiet relative to the noise of stellar jitter and measurement errors that there are still many, many orbital parameter candidates in the hopper.

As was stated in the introduction to this dissertation, the Appendix contains two non-trivial chapters that simply did not fit into the flow of the other chapters in this paper. The first is an attempt to improve upon the MPS by conditioning upon the estimated stellar jitter. The second is an attempt to integrate out the parameters of the one-planet model using a novel approach called "nested sampling". Unfortu-

nately, neither of these worked very well. The first suffered from over zealous use of parametric models by me, and the second was probably just a case of applying the wrong tool to the problem. Still, they remain in this paper to document research that, while never reaching a satisfying conclusion, is at least informative in unintended ways.

I hope that before I die I get to see exoplanetary images sent back to the earth from a deep-space probe sent to another planet at nearly light speed. It sounds far-fetched today, but just fifty years ago, it would have seemed quite far-fetched that we would be able to detect planets by observing minuscule Doppler shifts in their star's light, indicating radial velocities measurable with precisions of as little as 3ms^{-1}. Yet in just the last fifteen years or so we've detected several hundreds of exoplanets, and they keep on pouring in, and they keep getting smaller.

In my acknowledgements at the start of this paper I dedicated it to my parents. I also dedicate it now to posterity, to those future generations who may someday get to see up close those planets of which today we merely catch the faintest whiff.

Appendix A

The Maximal Periodogram Spike Height, Conditioned Upon the Estimated Stellar Jitter

A.1 Summarizing the MPS while taking into account a stellar jitter estimate

Stellar jitter varies from star to star. Although certain stellar properties such as metallicity and age can affect stellar jitter and therefore aid in estimating it, we will ignore that fact in this paper. Stellar jitter over many stars has a distribution that is centered on around 3 m s^{-1} and that has very little mass beyond around 10 m s^{-1}. A reasonable approximation of the estimate of stellar jitters is therefore $\mathsf{Gam}(3, 1)$, and for the rest of this paper, we will use that distribution when we are simulating stellar jitters or estimating their distribution among observed real stars. Stellar jitter is not known to be any different for stars that host planets and those that don't.

We are interested at present in the distribution of the MPS under the no-planet model, M0, because if an observed value is unusually large under that distribution, then it suggests the presence of a planet. It is tempting to simulate that distribution by simulating lots of samples of noise from lots of stars—each star having a stellar

jitter drawn independently from the $\mathsf{Gam}(3,1)$ distribution—and recording the MPS for each simulated sample. But that may be a mistake. When deciding whether, for a real data set, the MPS is "unusually large" under the M0 model, there is more information in the data than just the MPS. It is also possible to estimate the stellar jitter under the M0 model. And the conditional distribution of the MPS given that jitter estimate may be considerably different from the distribution of the MPS over many jitters. If that is in fact the case, then judging the performance of the MPS based on a distribution that ignores the jitter estimate would be judging a straw man.

In a similar way, when we perform a hypothesis test of $H_0 : \mu = \mu_0$, our decision whether to reject H_0 is based not only on $\bar{X}$ but also on s, the sample standard deviation. If s was ignored and the distribution of $\bar{X}$ was based on many possible values of σ, then it would certainly have greater spread than the distribution of $\bar{X}$ conditional upon s. So it would be harder to tell whether an observation of $\bar{x}$ came from its distribution under $H_0 : \mu = \mu_0$, or from a distribution based on some other μ. In a similar way, it is possible (though not obvious, nor necessarily true) that ignoring the stellar jitter will render the distribution of the MPS much more variable than it would be conditional upon an estimate of the jitter. So just as in the mean-testing setting, we transform $\bar{X}$ into a test statistic t that takes the sample standard deviation into account, here we need to transform the MPS into a statistic that takes the estimated stellar jitter into account. But how?

What we want to know is: "Given this particular estimate $\hat{s}$ of stellar jitter under the no-planet model M0, how large is this particular MPS under the no-planet model?" One way to address that is to estimate $\mathsf{Pr}(MPS < MPS_{observed}|M0, \hat{s})$, the probability of observing an MPS up to the magnitude of the one we did observe, given the stellar jitter estimate. If that probability is "unusually large" under the M0

model, then we will reject that model in favor of the M1 model.[1] The probability we just described is therefore a one-dimensional statistic that we can treat as decision-making tool. Let's call that statistic $P_{MPS\hat{s}}$. We need to determine its distribution under M0 and we need a good way to compute it for a given set of radial velocity measurements.

The first of these goals—determining the distribution of $P_{MPS\hat{s}}$ under M0—may be done using simulations, though we will see that it is a bit tricky, since we are conditioning not on the actual s (which we cannot know) but on the estimate $\hat{s}$. If we were able to condition on s_0, then the distribution of $P_{MPS,s}$ would be uniform over the interval $[0, 1]$. However, since we are conditioning on $\hat{s}$ instead, then the distribution may deviate somewhat from the uniform distribution. We will see in a simulation just how far from a uniform distribution it is.

Our second toal is finding a good way to compute the statistic for a given set of radial velocity measurements. Let's address that first and then return to the distribution question.

Let's consider the following algorithm:

- First choose a sample size n to study.[2]
- Fix the stellar jitter s at some particular plausible value s_0. [3]
- Generate the measurement errors σ_i that will be associated with our data.

[1] This is not exactly what astronomers really do with the periodogram; what they actually do is compute a chi-square statistic from the periodogram whose distribution is known to be $\mathsf{Chi-square}$ under the assumption of equal variances at every observation. By not making that assumption, we're actually permitting the MPS statistic to perform even better than perhaps it is already doing in practice. Thus, if our later analysis shows that one of the other two statistics—the Maximum Likelihood Ratio or our Bayes Factor Estimate—performs better than the MPS, then we can be nearly certain that it must perform better than the MPS as it is used in practice.

[2] The distribution of the MPS depends greatly on the sample size, so each one must be considered separately.

[3] The plausible range of stellar jitters is from 0 m s^{-1} to about 10 m s^{-1}. In fact, the distribution of stellar jitters among all stars might reasonably be approximated by a $\mathsf{Gam}(3, 1)$ distribution.

These will be independent draws from a $\mathsf{Gam}(8.0.5)$ distribution, which approximates reasonably well the distribution of actual measurement errors reported for radial velocity measurements taken over the last decade.[4]

- Simulate radial velocity measurements as independent draws from a $\mathsf{N}(0, s^2 + \sigma_i^2)$ distribution; that is, each radial velocity is normally distributed with mean 0 and variance equal to the sum of the variance due to stellar jitter and the variance due to the measurement errors.

- Using that data set, estimate the mean and the stellar jitter s. (Recall that the measurement errors are known.) There are different ways to estimate these two quantities. We'll use the two-dimensional MLE, which is easy to compute. We'll call the stellar jitter estimate $\hat{s}$ and the mean estimate $\hat{C}$. (Recall that the true mean radial velocity is the parameter C in the model M0.)

- There is no guarantee that our estimate $\hat{C}$ will equal zero, nor that our estimate $\hat{s}$ will equal the s_0 that we used to generate the data. Indeed, it would be highly unlikely for them to be identical. Now $P_{MPS\hat{s}}$ is not computed using s_0, but $\hat{s}$. So let's compute the MPS for our simulated data and then store it in a "library" on a "shelf" corresponding to the value $\hat{s}$. Since $\hat{s}$ is actually continuous, we'll round $\hat{s}$ off to the nearest 0.1, which should be sufficiently small that the actual distribution of $MPS|\hat{s}$ won't vary much among values of $\hat{s}$ that are on the same "shelf".

- We want to estimate the distribution of $MPS|\hat{s}$ for values of $\hat{s}$ ranging from 0 to about 10. That's 100 shelves in our library—not an impossibly large number of shelves to fill. So let's let s_0 vary from 0 to 10 (we could do this systematically

[4] Improvements in technology have reduced measurement errors somewhat, and there is hope that they will eventually be as small as 1 m s^{-1}; we will not assume that here, however.

or by drawing s at random from the $\mathsf{U}(0, 10)$ distribution) and repeat all the above steps, gradually filling the shelves of the library with estimates of $MPS|\hat{s}$.

- When all the shelves are sufficiently "full"—i.e., when they contain enough simulations of $MPS|\hat{s}$ for us to be comfortable that the distribution on every shelf is well-represented, we'll construct a conversion from the pair $(MPS, \hat{s})$ to the univariate statistic $P_{MPS\hat{s}}$ simply by referring to the shelf corresponding to $\hat{s}$ and computing the fraction of simulated MPS values that are less than the particular one under consideration.

Our "library" has 100 shelves on it, and each shelf will contain thousands or (ideally, if computational time is not too heavy) millions of simulations of $MPS|\hat{s}$. But all of that was for only one fixed sample size n. Sample sizes up into the hundreds (though probably not the thousands) will be of interest to us, and that multiplies by a considerably amount the number of simulations we will need in order to construct our library. But perhaps we'll be lucky, and the shelves in our library will show predictable patterns when full. Each is characterized by only two variables: $\hat{s}$ and n; maybe there will turn out to be a reasonable functional approximation of the distribution on every shelf, defined only by $\hat{s}$ and n.

To find out whether that is the case, we'll first fix n at some value and construct an abridged version of the library above, letting s_0 take only integer values from 0 to 10, and only shelving values of MPS for which $\hat{s}$ rounds to an integer when rounded to the nearest 0.1. Then we'll see whether there is a pattern to the distributions on all the shelves. If there is, we'll then fix $\hat{s}$ (as best as we're able—we'll actually be fixing s_0, but that should generate a lot of data sets in which $\hat{s}$ rounds to s_0) and systematically vary n from something rather small (2 or 4, perhaps) up to something quite large (around 500). Then we can see once again whether there is a discernable pattern dependent upon n. Finally, if we can conjecture a functional distribution

based solely upon $\hat{s}$ and n, then we can see whether our conjecture holds for several various combinations of $\hat{s}$ and n that we have not considered. Once we do that—assuming we can in fact do so—then we'll be in a very happy position. We'll be able to quickly simulate values of $P_{MPS\hat{s}}$ for any sample size, not just under model M0 (which is not so useful), but also M1. Then we will be in a very good position for evaluating how well the MPS (which we now see is really the pair MPS,$\hat{s}$) performs under various conditions.

A.2 The library containing $MPS|\hat{s}$ on shelves defined by $\hat{s}$

In the last section I described an algorithm for constructing a "library" that could be used to convert the pair $(MPS, \hat{s})$ to a single univariate statistic $P_{MPS\hat{s}}$. Then I expressed an optimistic hope that such a library might not need to be constructed in full—100 shelves full of simulations for each of several hundreds of sample sizes. Now let us see whether that hope was not in vain.

A.2.1 A set of 102 shelves

We'll first fix our sample size at (somewhat arbitrarily) $n = 30$[5], and we'll draw stellar jitters randomly from the $\mathsf{U}(0, 1)$ distribution. Then we'll follow the library-construction algorithm described in section A.1, doing a total of 500,000 simulations. We'll use shelves for values of $\hat{s}$ ranging from 0.0 to 10.1—the last shelf will collect all the MPS values corresponding to estimates of $\hat{s}$ that exceed 10.0, and we won't consider them. If the $\hat{s}$'s are approximately evenly distributed over the 101 other shelves, then each shelf will contain around 5,000 simulations, which should be sufficient for our purposes.

Although the radial velocities do not depend upon the observation time in model

[5] This is a footnote, not an exponent! Recall from section 3.1.3 that 30 radial velocity measurements that are mere noise can sometimes look like a planet is present

M0, the periodogram is affected by the range of the time observations. We'll arbitrarily let them be randomly drawn from a time window of 180 days. It is not likely that this time window will affect the eventual distribution of our statistic $P_{MPS\hat{s}}$, but it would be prudent not to take that for granted, and we will verify it at the appropriate time.

After running the simulations, there are two shelves that contain more simulations than the others. These are the shelf for $\hat{s} = 0$ and $\hat{s} = 10.1$, the two "end" shelves. The latter is easy to explain: $\hat{s}$ will not equal s_0 exactly, though it should always be in the same vicinity, perhaps $\pm 2\text{m s}^{-1}$ or so. Thus, whenever we draw an s_0 greater than around 8 m s^{-1}, there's a chance that $\hat{s}$ will exceed 10.0. Since that's the highest shelf we permitted, then that shelf will collect a lot of MPS's that correspond to values of $\hat{s}$ that do not in fact round to 10.1, and we therefore won't consider the items on that shelf.

It's more difficult to understand why the first shelf, $\hat{s} = 0$ should contain an unusually large number of results, though it's likely that the explanation is similar: s_0 must be positive, as must $\hat{s}$ be. If s_0 is drawn close to 0, then the distribution of $\hat{s}$'s will not be so spread out as they are for larger values of s_0, but will tend to "bunch up" near zero. So it's not surprising that the first shelf contains more simulations than the others. Figure A.1 shows the distribution of the values of $\hat{s}$ resulting from the simulation.

A.2.2 One shelf: can it be estimated with a functional form?

Now let's look at the values of MPS on one particular shelf—arbitrarily, let's choose $\hat{s} = 3.0$. It happens to have 4692 simulated MPS's on it. Their distribution is shown in Figure A.2.

We said earlier that we were hopeful that we would be able to predict the distribution of MPS's on a shelf based solely on the $\hat{s}$ label on the shelf and the sample size

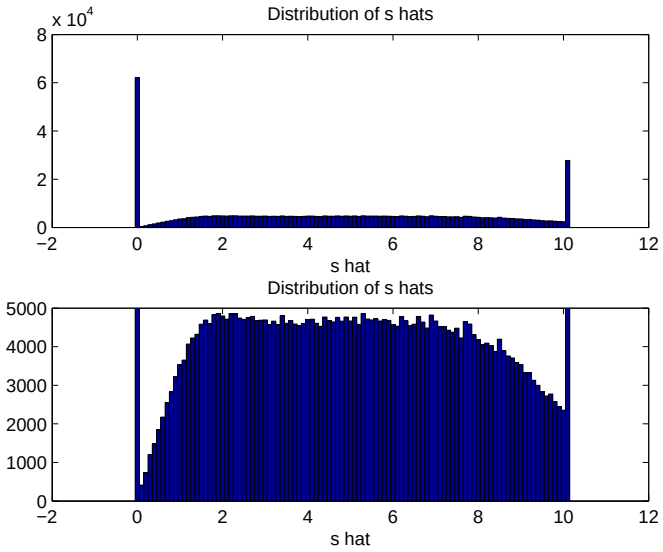

FIGURE A.1: The upper histogram shows the distribution of $\hat{s}$'s, including the abundance of 0's and 10.1's, the latter of which are actually just an accumulation of everything greater than 10.0, and it will be ignored. The lower histogram shows the same thing on a scale that permits us to see how many simulations of MPS sit on each shelf—approximately 5000, as expected.

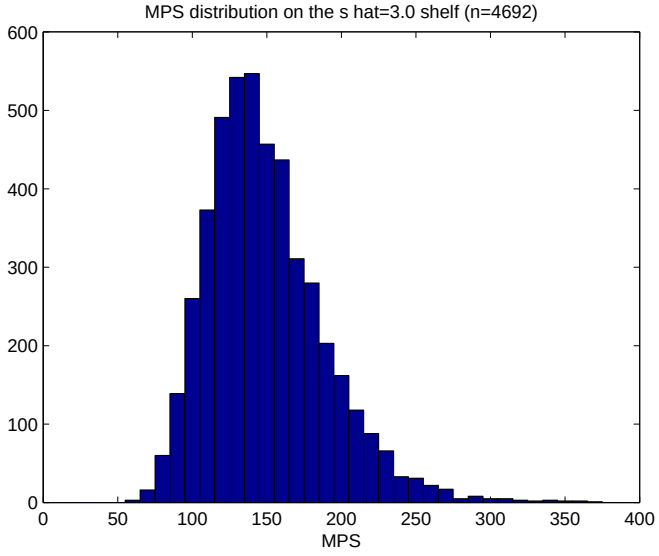

FIGURE A.2: The distribution of MPS estimates on the $\hat{s} = 3.0$ shelf looks promisingly Gamma.

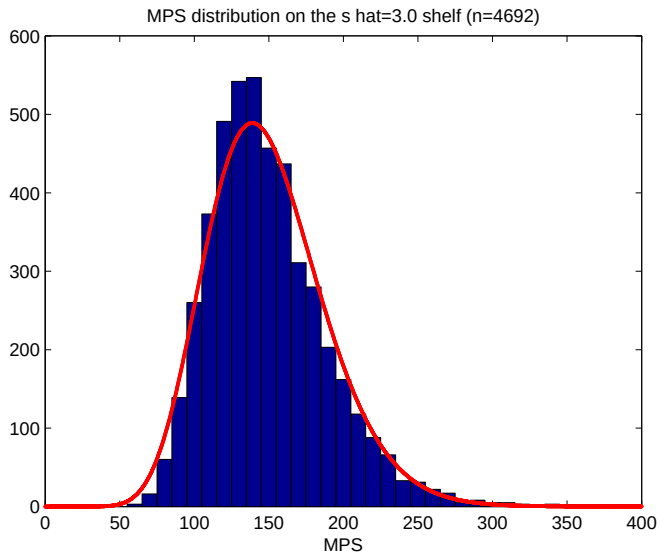

FIGURE A.3: The distribution of MPS's on the $\hat{s} = 3.0$ shelf, with the $\mathsf{Gam}(14.34, 10.42)$ distribution superimposed. It's not perfect, but it may be satisfactory for transforming MPS into a univariate statistic $P_{MPS\hat{s}}$ that takes into account the estimated jitter.

n. Let's now pursue that goal. Since the distribution in Figure A.2 looks promisingly Gamma, let's approximate it with a $\mathsf{Gam}(\alpha, \beta)$ distribution whose parameters α and β are chosen so that $\alpha\beta$ equals the mean of the 4692 MPS's, and so that $\alpha\beta^2$ equals their variance.[6] That happens to give us $\alpha = 14.34$ and $\beta = 10.42$. In figure A.3 we see the histogram of simulations once again, this time with the $\mathsf{Gam}(14.34, 10.42)$ distribution—rescaled to match the area of the histogram—superimposed.

We see from the figure that the fit is good, though not perfect. But our purpose doesn't require that the fit be perfect. We want to transform MPS into a statistic $P_{MPS\hat{s}}$ that takes into account the jitter estimate, and it would be nice for that transformation to have a functional form rather than requiring a reference library. If we come up with a functional form that isn't perfect, it may still be adequate.

Let's consider two other ways we can compare the simulated MPS's on our shelf

[6] The mean and variance, respectively, of the $\mathsf{Gam}(\alpha, \beta)$ distribution are $\alpha\beta$ and $\alpha\beta^2$.

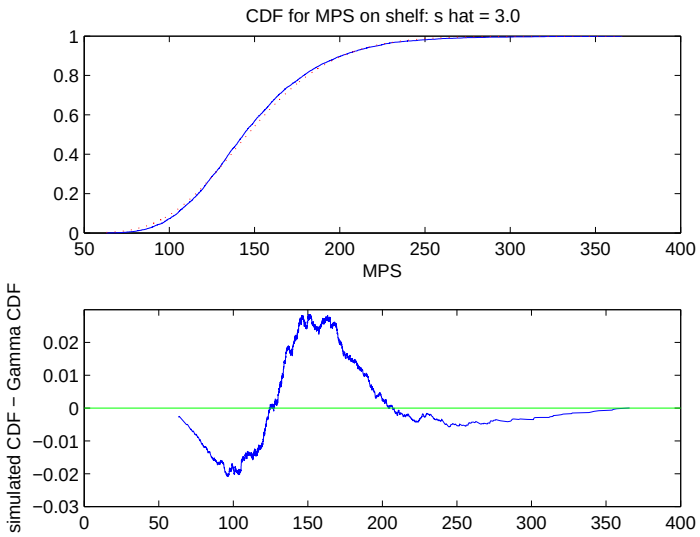

FIGURE A.4: In the upper plot, the solid blue curve is the CDF of the simulated MPS values. The dotted red curve is the CDF of the Gam$(14.34, 10.42)$ distribution. The lower curve shows by how much the two differ in probability, which never exceeds 0.03 in magnitude.

to the given Gamma distribution. One is to superimpose their CDF's, and that is shown in Figure A.4.

Another way we may compare these two distributions is with a quantile plot, which is shown in Figure **??**.

This plot shows some cause for concern. Were we to use the Gamma distribution to make estimates of the MPS, we would most of the time have a reasonably good approximation. However, about 2% of the time we would require an MPS estimate in excess of 250, and the Gamma distribution would tend to underestimate it. For example, we see in the figure that if the true MPS were around 350, our Gamma estimate would be only around 300. We need to consider just what we would be using the Gamma approximation *for*, and whether the figures we see here indicate trouble.

In a later section we will be examining the statistic we've called $P_{MPS\hat{s}}$, looking at

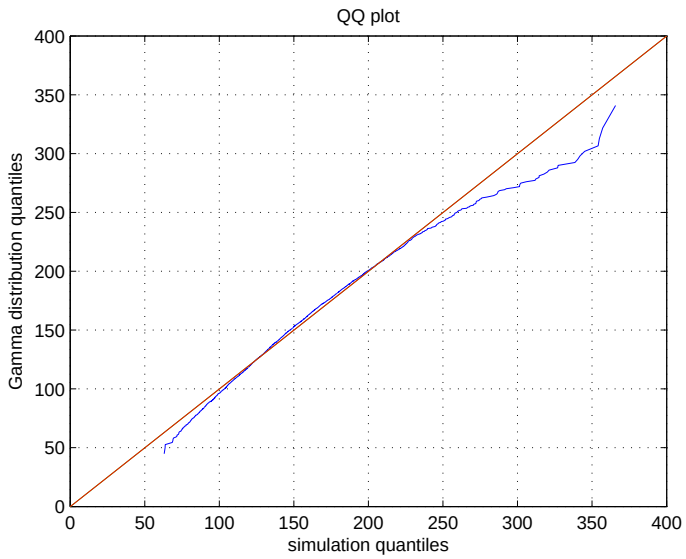

FIGURE A.5: The QQ plot shows that over most of the distribution, the Gamma distribution could be used to estimate an MPS value with relative accuracy. However, at the upper end of the distribution, the Gamma estimation would tend to underestimate the actual quantile. However, this happens only very rarely. The probability of getting an MPS value in excess of 250 is only 0.019, as we see in Figure **??**, and in this figure we see that only MPS values bigger than 250 are likely to result in considerable underestimation by the Gamma approximation.

its distribution under model M0 (which we earlier asserted ought to be approximately $\mathsf{U}(0,1)$); and also under M1, with the distribution of planets approximating what we think is actually out there in the universe; and also under M1 with the planet distribution tailored to particular types of planets. With this set of distributions, we will be equipped to evaluate how well the statistic performs. For example, if we want to take a frequentist approach, then the distributions will permit us to plot ROC curves, showing true planet-detection rates versus false detection rates. Or if we want to take a likelihood approach, then the distributions will permit us to estimate the Bayes Factor of the statistic under different models, and thereby estimate how likely the statistic is to provide strong, ambiguous, or misleading evidence.

But both of those statistic-evaluation criteria involve comparing the distribution

of a statistic under two different models. Even if our statistic $P_{MPS\hat{s}}$ doesn't have exactly a uniform distribution under model M0 as we had asserted earlier, that's okay: we can compute it for each simulation using the Gamma approximation to the library shelf, *but simulate its distribution under M0, and use that distribution to evaluate its effectiveness.* We thus will have achieved our goal of permitting the MPS to look its best by taking into account the estimate of stellar jitter. Even if we do so imperfectly, the errors will occur at the stage of estimating the statistic $P_{MPS\hat{s}}$, not at evaluating the effectiveness of the statistic. $P_{MPS\hat{s}}$ can hardly be called a "straw man". Our whole purpose for the library and the conversion of MPS into something else was to take into account the stellar jitter estimate. That we will have done with this statistic. If it performs as well as another statistic T, we can be reasonably confident that the MPS (with the stellar jitter estimate) is as good a planet-detection tool as is T. If T proves superior to $P_{MPS\hat{s}}$, then we can be reasonably confident that it would also be superior to any statistic that was a function only of MPS and $\hat{s}$, including the "true" value of $P_{MPS\hat{s}}$ that were aiming for.

But although we may be reasonably confident that *this* Gamma distribution does a sufficiently good job of approximating the distribution of MPS values on *this* shelf, we are far from being ready to study the performance of the statistic $P_{MPS\hat{s}}$. We still need to consider: first, whether the MPS values on other shelves may also be approximated by a Gamma distribution; and second, what the relationship is between the parameters of the Gamma distribution and the values of $\hat{s}$ and n (need these be a library themselves, or will a function present itself?). And we should be prepared to treat as a special case the shelf $\hat{s} = 0$, since its accumulation of so many more estimates than the other shelves suggests that the behavior of MPS on that shelf might be qualitatively different from the others. Or we may get lucky and it may look just like all the others.

A.2.3 Other shelves

We'll eventually need to look at other sample sizes, but first let's consider the 99 shelves from $\hat{s} = 0.1$ m s^{-1} to $\hat{s} = 10.0$ m s^{-1} that we've already simulated. Since we've already seen that one of them looks reasonably Gamma, we might speculate that the others will, too. And we can use the same method of moments to fit a "best" Gamma distribution to each one. Let's do that, and then make two plots. One of them showing the maximal difference between the cumulative probability estimated by the simulation and the cumulative probability estimated by the corresponding Gamma distribution. For the shelf we already looked at, $\hat{s} = 3.0$, that was 0.0285. And let's make another plot showing the values of α and β that the method of moments produces for each estimate. In both cases, these values of interest may be plotted against $\hat{s}$.

So in Figure A.6 we see the differences between the cumulative probability estimates plotted against $\hat{s}$, and in Figure A.7 we see the *alpha* and β estimates, also plotted against $\hat{s}$.

These two plots are very promising. Figure A.6 suggests that a Gamma distribution is for all values of $\hat{s}$ a reasonably good approximation of the distribution of $MPS|\hat{s}$. And Figure A.7 indicates that there is a pattern to the Gamma distributions' α and β values that looks like it could be reasonably well approximated—over this range, at least, and larger values of $\hat{s}$ are highly improbable—using a simple regression, such as quadratic regression.

But we won't do that right away, since any results we obtained would be local to this sample size n. Instead, we'll first consider a single $\hat{s}$ and let the sample size vary, to see if the Gamma approximation will work for other sample sizes as well. And if it will, then we won't need an entire library full of shelves, we'll just need a way to estimate α and β based on n and $\hat{s}$. And that we should be able to do using

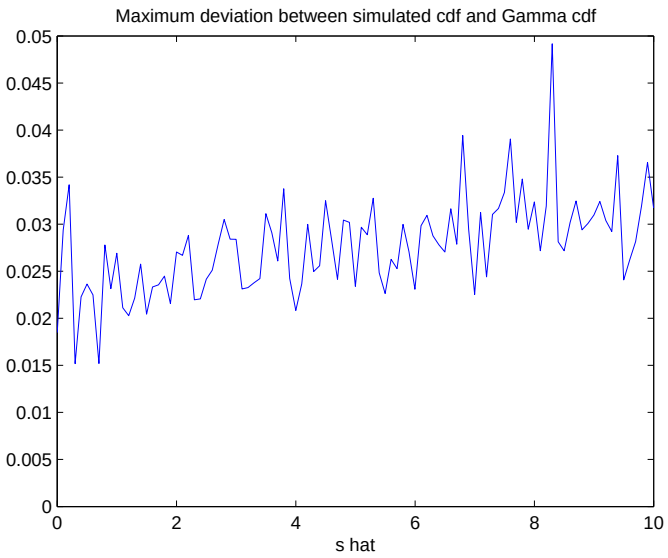

FIGURE A.6: The maximal errors in using a Gamma distribution to approximate the distribution of $MPS|\hat{s}$—happily, even when $\hat{s} = 0$—are all around 0.03, and none more than 0.05. This suggests that the Gamma approximation is probably all right to use, at least for this sample size.

regression on these two variables, assuming that the sample size n is as kind to us as was $\hat{s}$.

A.2.4 Other sample sizes

Let's consider only $\hat{s} = 3.0$. We'll look at sample sizes ranging from quite small ($n = 4$) to quite large ($n = 400$), and in each case we'll simulate MPS values under model M0. There is an inconvenient element here that we can't easily avoid, which is that our simulations are dictated by s_0, not $\hat{s}$. Happily, at eight combinations of $(\hat{s}, n)$, I found the following to be remarkably consistent: simulating MPS values using s_0 and keeping all values; simulating them using s_0 and only keeping those MPS values for which $\hat{s}$ rounded to s_0 when rounding to the nearest integer; and simulating them using s_0 and only keeping those MPS values for which $\hat{s}$ founded to s_0 when rounding to the nearest 0.1—these three simulations always resulted in extremely

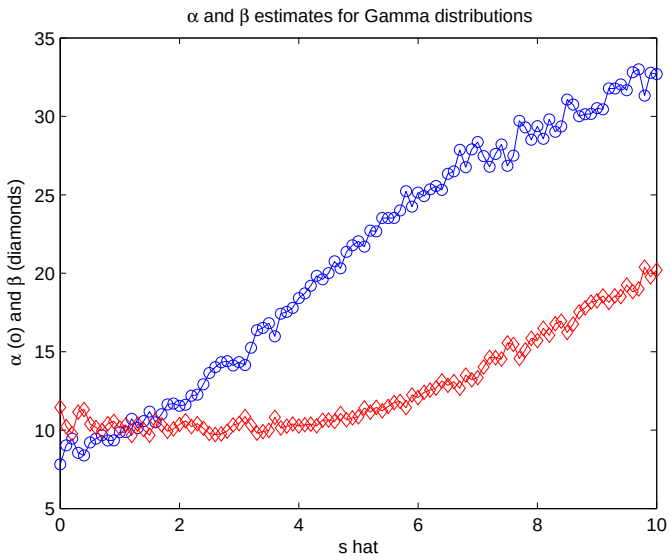

FIGURE A.7: The blue circles show estimates of α and the red diamonds, estimates of β. Although these are not exactly linear—especially β, they appear to be well-behaved, and within this range of values of $\hat{s}$ (others are highly improbable) it may be reasonable to approximate these using quadratic regression.

similar distributions. The means were nearly identical over 10,000 simulations, and the standard deviations had a consistent relationship, which was that the greatest spread resulted from the first simulation, in which all values of MPS were kept; the standard deviation of the MPS's from the second simulation was about 84% of those from the first; and the standard deviation of the MPS's from the third simulation was about 76% of those from the first. This suggested to me a reasonably efficient way of simulating $MPS|\hat{s}$: I would simply simulate $MPS|s_0$, and then compress the results towards their mean by 24%, and treat the results as if they were a simulation of $MPS|\hat{s}$.

So using that simulation technique, we'll continue until we have 5000 values of MPS for each sample size under consideration, and then we'll be ready to again evaluate how well a Gamma distribution will approximate each one.

First let's look at Figure A.8, which shows the size of the differences between

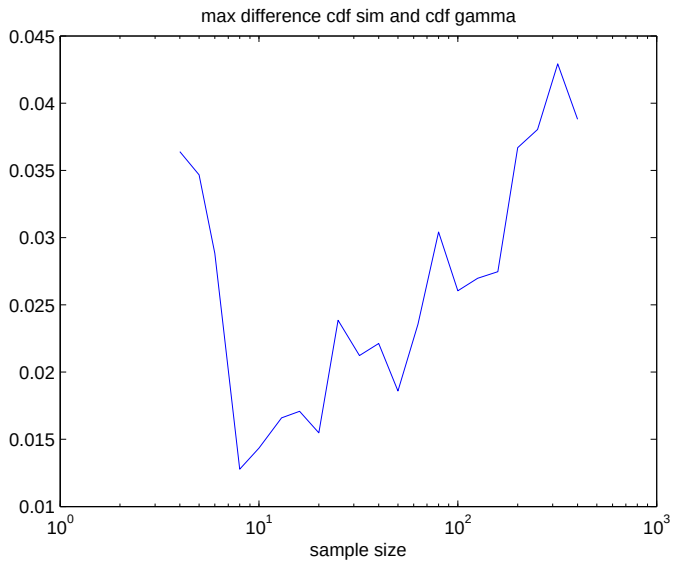

FIGURE A.8: Using a Gamma distribution to approximate the distribution of $MPS|\hat{s}$ appears to produce estimates that are in error by around 0.025 and not more than 0.045. This is consistent with what we observed in Figure A.6 when $\hat{s}$ varied from 0.0 to 10.0 and the sample size was fixed at $n = 30$. These errors are sufficiently small that we may comfortably use the Gamma approximation to the distribution.

the cumulative probability estimated by the simulation and that estimated by the Gamma distribution. We see that for all of these sample sizes, that difference is smaller than 0.045, and is more typically around 0.025. This is consistent with what we saw when we simulated many $\hat{s}$'s for a sample size of $n = 30$, and it is sufficiently small that we may feel reasonably confident in using the Gamma distribution as an estimator.

Now consider Figure A.9, which shows the relationship between the α and β estimates and the sample size. In one way, these plots are encouraging. It is clear that α and β vary in a fairly predictable way with respect to the sample size. However, with the exception of the log-log plot of β with respect to sample size (red diamonds in Figure A.9), none of the other plots look linear or suggest an obvious regression model. It may be that we will need to resort to a "library" of α's and β's after all,

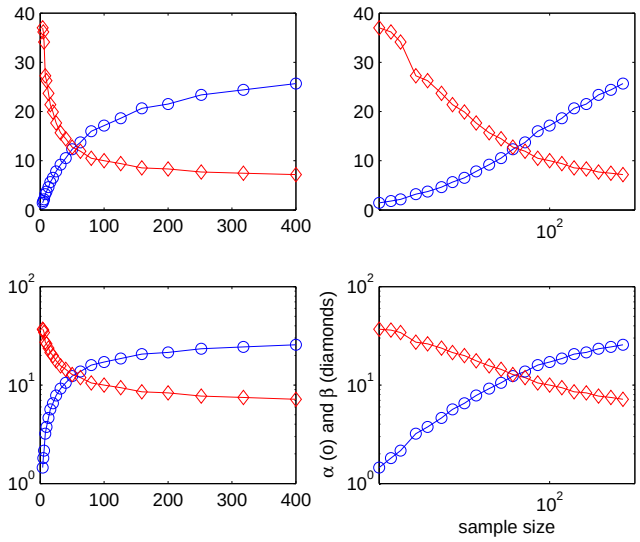

FIGURE A.9: These figures show the relationship between the α (blue circles) and β (red diamonds) estimates in the Gamma distributions and the sample size. All plots show the same data, but the different combinations of linear and log scales are shown in the four plots. $\log \beta$ appears to have an approximately linear relationship with $\log n$, but none of the other plots show linearity.

though we won't need whole shelves of MPS simulations.

If it seems that we have been making too many assumptions, recall that whatever statistic we end up producing and calling $P_{MPS\hat{s}}$, we will examine according to its own definition, and will simulate its own distribution under different models. So again: if it isn't exactly what we set out to construct originally, that isn't necessarily fatal to our evaluation of the performance of MPS or the pair (MPS,$\hat{s}$) as a tool for identifying planets.

Additionally, after we have performed a regression analysis to predict α and β, we will use our equation to predict a few other distributions and "spot-check" them against simulations of their own. Hopefully that will confirm the reasonableness of the definition.

Lest we forget why we are doing this regression and Gamma approximation:

Before this present chapter is over, we would like to have evaluated the statistic $P_{MPS\hat{s}}$ under model M0, under M1 with a realistic distribution of planets, and under M1 with various particular types of planets. And we'll want to do all of this with many different sample sizes so that can see how the power of the statistic is affected by the number of radial velocity observations. (It is possible that we'll be able to find something better, like the maximum likelihood ratio, which will be studied in the next chapter.) But computing values of $P_{MPS\hat{s}}$ for such a large number of combinations of $\hat{s}$ and n would require a very large "library" of reference distributions, each of which requires a time-consuming simulation to fill.[7] Using the Gamma approximations will be much faster. The danger is that we are "blurring" our precious statistic—i.e., we are introducing additional variability due not to planetary presence, but merely to estimation error. However, that blurring should be kept below 0.03 or so, and the statistic itself has the full range of the interval [0,1]. So that blurring is a relatively small price to pay for the considerably increased rapidity of computation.

A.3 α and β as a function of $\hat{s}$ and n

We would like to estimate α and β as a function of $\hat{s}$ and n. The plots we have seen suggest that such a function may exist, and they suggest that a log scale on n is appropriate. But they do not offer up an obvious regression model. In Figure A.7 we saw what appeared to be an approximate linear relationship between α and $\hat{s}$, but β did not look exactly linear, nor exactly quadratic. Then in Figure A.9 we saw an

[7] For example, the graphs in Figures A.6 and A.7 resulted rom 500,000 simulations that required about 10 hours of computing time on my laptop, and that was with a sample size of only $n = 30$. Naturally the larger sample sizes would take longer still. With a faster computer, and possibly more efficient coding, perhaps one could construct a large and useful library in a reasonable length of time, and these Gamma approximations would not be necessary. I am asserting, however, that the Gamma approximations can produce a univariate statistic that can be computed much more quickly at each iteration, and without the requirement of a full library for every sample size, and which will have roughly the same planet-detecting ability as would a version of $P_{MPS\hat{s}}$ based solely on millions of simulations.

apparent linear relationship between $\log \beta$ and $\log n$, but nothing simple presented itself for α.

But one thing seems likely: plots of α and β over the two-dimensional space $\hat{s} \times \log n$ are probably fairly smooth and monotonic or nearly so. So perhaps instead of attempting regression, we should simply use simulations to estimate α and β for certain lattice points on the $\hat{s} \times \log n$ plane, and then treat those as a reference library. For values of $\hat{s}$ and $\log n$ that don't happen to lie on the lattice points, we can interpolate using some smoothing function to get estimates of α and β.

Let's begin by constructing a lattice. In this paper we'll only consider sample sizes between 4 and 400. Let's divide that span up into 10 parts, evenly separated on the log scale. So we'll consider the following sample sizes: 4, 6, 10, 16, 25, 40, 63, 100, 159, 252, 400. We've already remarked that actual stellar jitters of observed stars are very seldom greater than 10, so let's consider stellar jitters estimates 0.0, 0.1, 0.2, ..., 12.0. (If $s = 10$, then $\hat{s} = 12$ may occur, though $\hat{s} = 13$ would be rare.) That ought to create a lattice structure that has sufficiently many points that we'll be able to interpolate α and β for any reasonable combination of $\hat{s}$ and n.

Figures A.10 and A.11 show respectively the α and β surfaces as functions of $\hat{s}$ and $\log n$. They are smooth and monotonic in both $\hat{s}$ and $\log n$, and the α surface looks as though it may even be independent of $\hat{s}$ and nearly linear in $\log n$, though it's hard to say. The β surface does not immediately suggest a parametric model. Figure A.12 shows the same two surfaces, but as profiles in $\hat{s}$ and $\log n$.

These surfaces are not perfectly smooth. Neither are they polynomials. However, it might be reasonable to perform use polynomial regressions on $\log n$ and $\hat{s}$ to estimate values of α and β for future data sets. After all, when we eventually get a real data set (or, in the upcoming sections, more simulated ones) we will need to transform the pair MPS,$\hat{s}$ (and $\log n$) into the statistic we're calling $P_{MPS\hat{s}}$, and that will require determining which (approximate) *Gamma* distribution MPS came

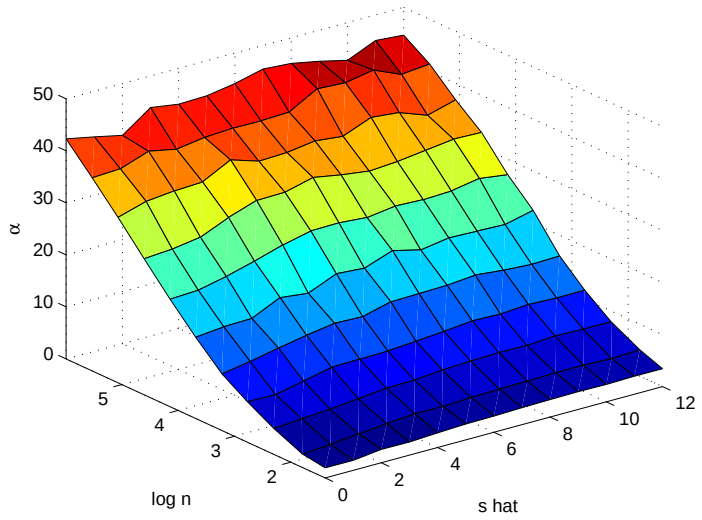

FIGURE A.10: The relationship between α and the $\hat{s} \times \log n$ plane is a rather pretty one, being smooth and monotonic in both dimensions. It suggests that $\hat{s}$ has relatively little to with α, and that the relationship between α and $\log n$ may be approximately linear.

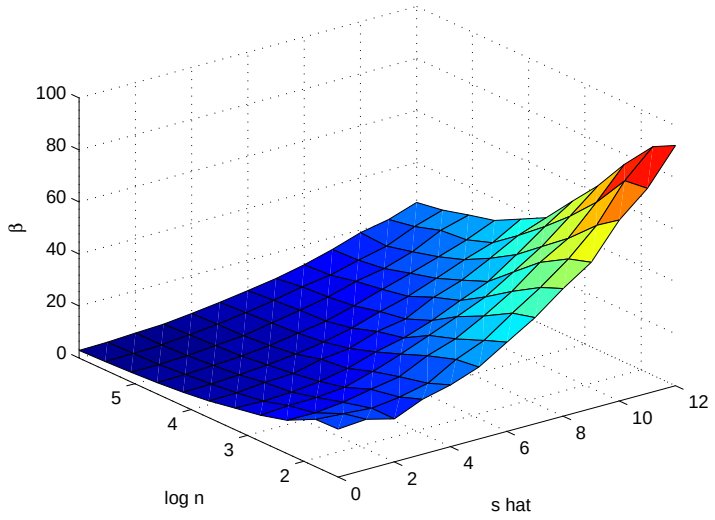

FIGURE A.11: We see from this surface that β rises with $\hat{s}$ and falls with $\log n$, though neither is in a linear way, and a regression model is not immediately suggested by the graph.

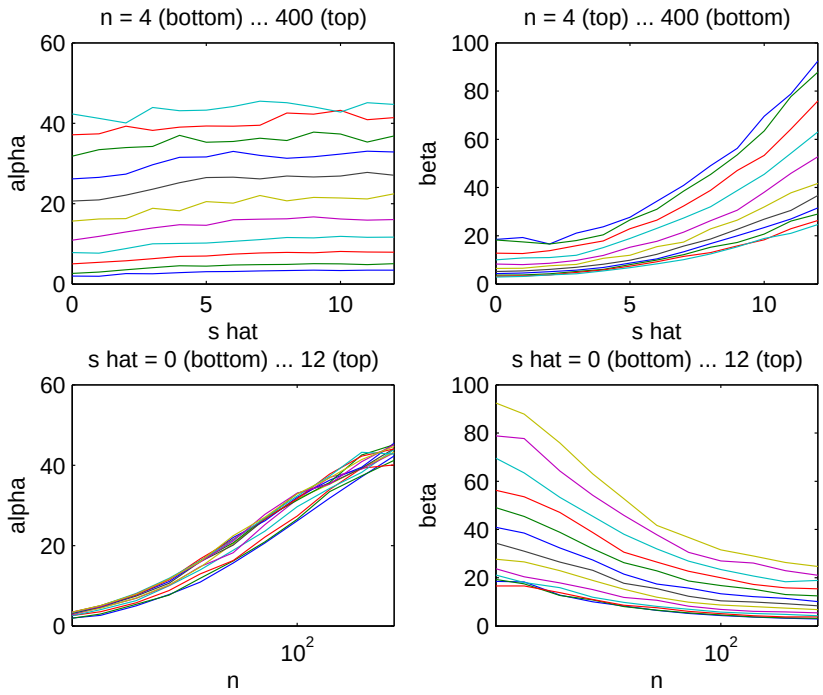

FIGURE A.12: The upper two plots show α and β against $\hat{s}$ for different values of n, and the lower two plots show α and β against n for different values of $\hat{s}$.

from.

There are two dangers in performing polynomial regression. If we use powers of $\log n$ and $\hat{s}$ that are too low, then the curvatures of the surfaces will not have enough wiggle room to conform themselves to the data. But if we use powers that are too large, then we risk surfaces that exhibit extreme behavior near the perimeters of our spaces. After examining several polynomial regression surfaces for both α and β, it seems that reasonable estimates of both can be made using a fifth-degree polynomial for α and a fourth-degree polynomial for β—that is, polynomials whose terms are of the form $MPS^{n1} \cdot \hat{s}^{n2}$, where $n1 + n2 \leq 5$ for the α surface, and where $n1 + n2 \leq 4$ for the β surface. Those regression surfaces are shown in Figure A.13 and Figure A.14. Residual stem plots for the two surfaces are shown in Figure A.15 and Figure A.16.

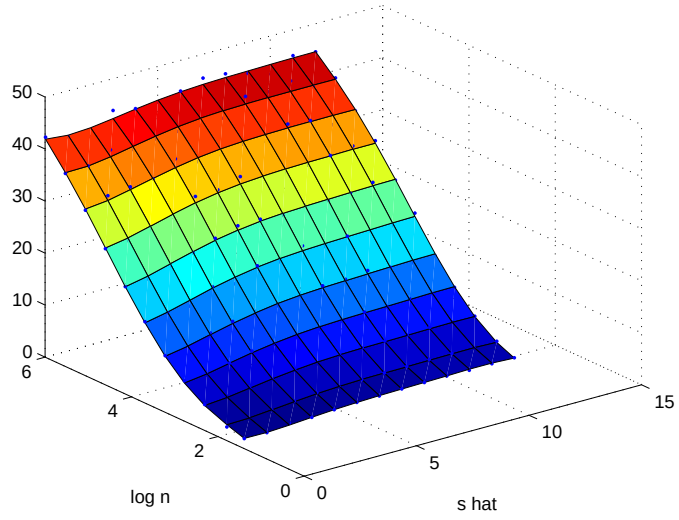

FIGURE A.13: The estimated α surface over $\log n \times \hat{s}$, using fifth-degree polynomial regression.

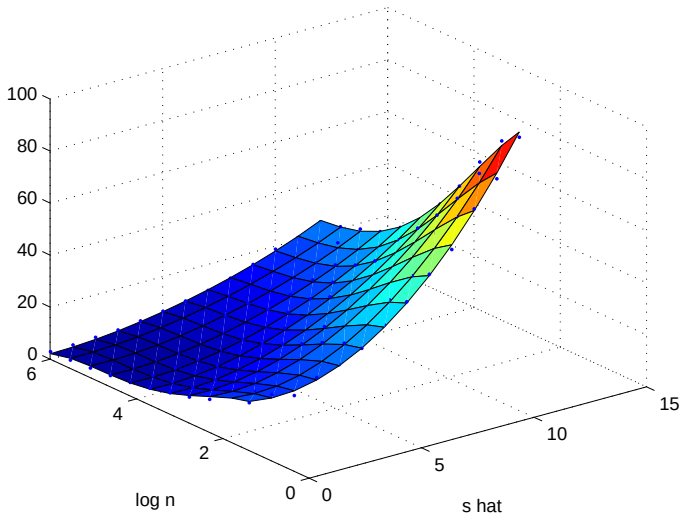

FIGURE A.14: The estimated β surface over $\log n \times \hat{s}$, using fourth-degree polynomial regression.

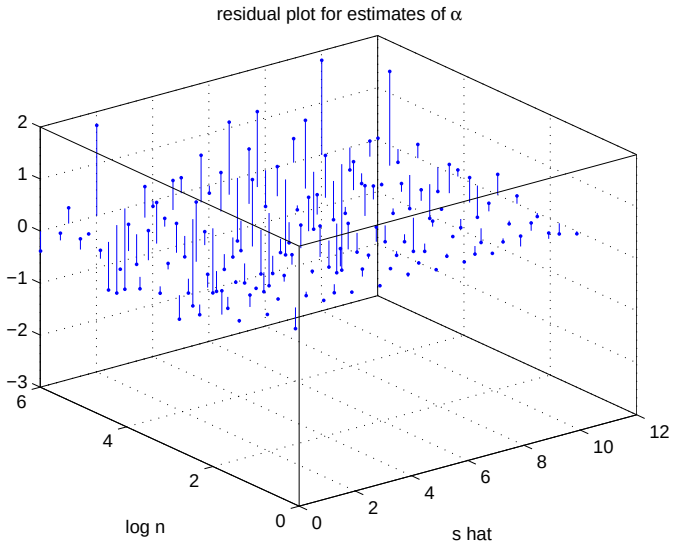

FIGURE A.15: Residual stemplot from the estimated α surface.

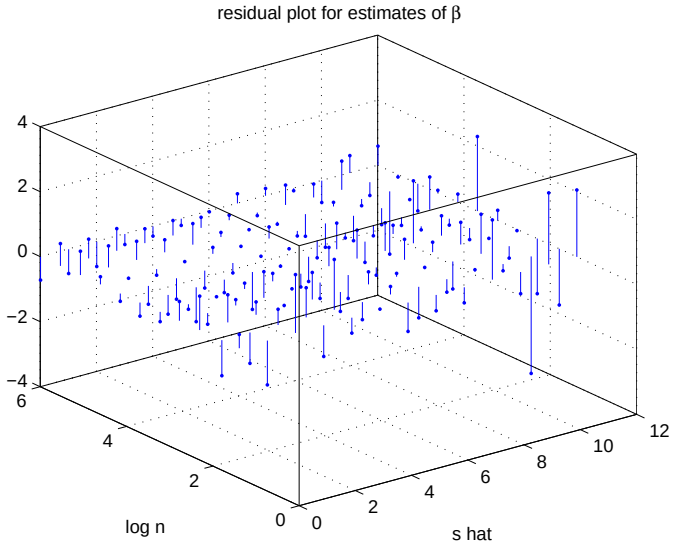

FIGURE A.16: Residual stemplot from the estimated β surface.

The magnitudes of the residuals are not large relative to the estimated values of α and β. However, it is slightly troubling that the residuals from the α surface grow larger when $\log n$ is large. Let's keep that in mind, and "proceed with caution".

Let us pause to recall what we have done. If we have radial velocity data from a star, we may fairly quickly determine the MPS—the height of the tallest spike in the periodogram—and also $\hat{s}$—the maximum likelihood estimate of the stellar jitter parameter s, which is seldom greater than 10 m s^{-1}. Given those two values, we can use the regression surfaces described in this section to estimate the α and β parameters corresponding to the $\mathsf{Gam}(\alpha, \beta)$ distribution that best approximates the distribution of MPS under the no-planet model M0. We may then quickly determine the cdf associated with that MPS in the identified Gamma distribution, and call that statistic $\mathsf{P}_{MPS\hat{s}}$. Due to the way this statistic is constructed, its distribution under M0 should be roughly uniform over its domain of $[0, 1]$ (though we'll want to verify that). If $\mathsf{P}_{MPS\hat{s}}$ is large, that will suggest the presence of a planet. A small value of $P_{MPS\hat{s}}$ will not provide evidence of a planet's presence. How large is large enough to suggest a planet is what we will consider in the next sections.

A.4 The distribution of $P_{MPS\hat{s}}$ under M0

We have said that our statistic $P_{MPS\hat{s}}$ under M0 should be roughly uniform over the unit interval. In this section we'll see how true that is. We'll simply simulate a large number of values of the statistic, drawing the stellar jitter s at random from the $\mathsf{Gam}(3, 1)$ distribution for each star, and the measurement errors from the $\mathsf{Gam}(8, 0.5)$ distribution. (We'll continue to let $C = 0$, the "background velocity" of the star, which has no impact at all on the periodogram.) First let's consider a sample of size $n = 30$. Figure A.17 shows the distribution of $P_{MPS\hat{s}}$ under M0, as estimated by 10,000 simulations.

The distribution is perhaps not quite so uniform as we might have hoped for, but

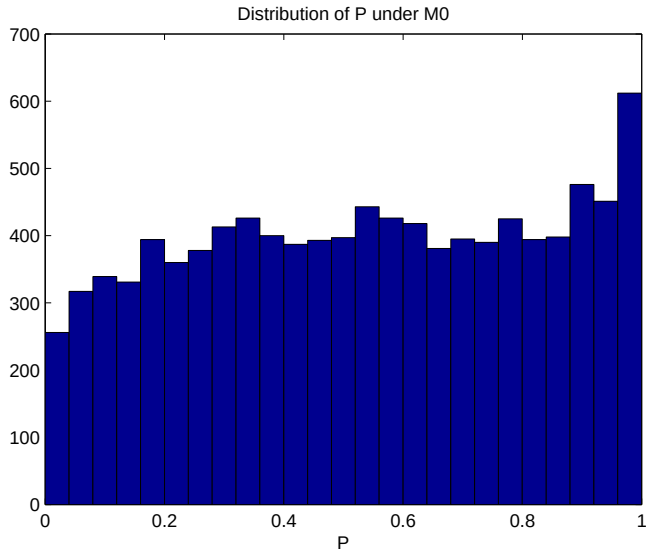

FIGURE A.17: The distribution of $P_{MPS\hat{s}}$ under M0 with $n = 30$ is not quite so uniform as we might like, but it is not so grossly different as to suggest that the construction process is seriously flawed. This histogram is based on 10,000 simulations. *Note: replace with a histogram showing 100K simulations.*

neither is it so grossly different as to suggest that the process of constructing $P_{MPS\hat{s}}$ from a data set is seriously flawed. Remember that by taking into consideration *this* distribution, we have the ability to evaluate the actual performance of $P_{MPS\hat{s}}$, not that of a hypothetical but unachieved statistic. The only meaningful deviations from the uniform distribution occur at the endpoints, where there is less probability mass near 0 and more near 1 than we had hoped for. That is most likely due to the imperfect approximation of the distribution of MPS by a Gamma distribution, which would have the most dramatic effect near the tails. The most important region is that very close to 1, as it is the high values of our statistic that will indicate planetary presence. It appears that the probability density near 1 is perhaps at most twice what it would be under the uniform distribution, and only in the small region where $P_{MPS\hat{s}} > 0.99$. This should not be too problematic.

We will need to see whether the sample size has an effect on the shape of the

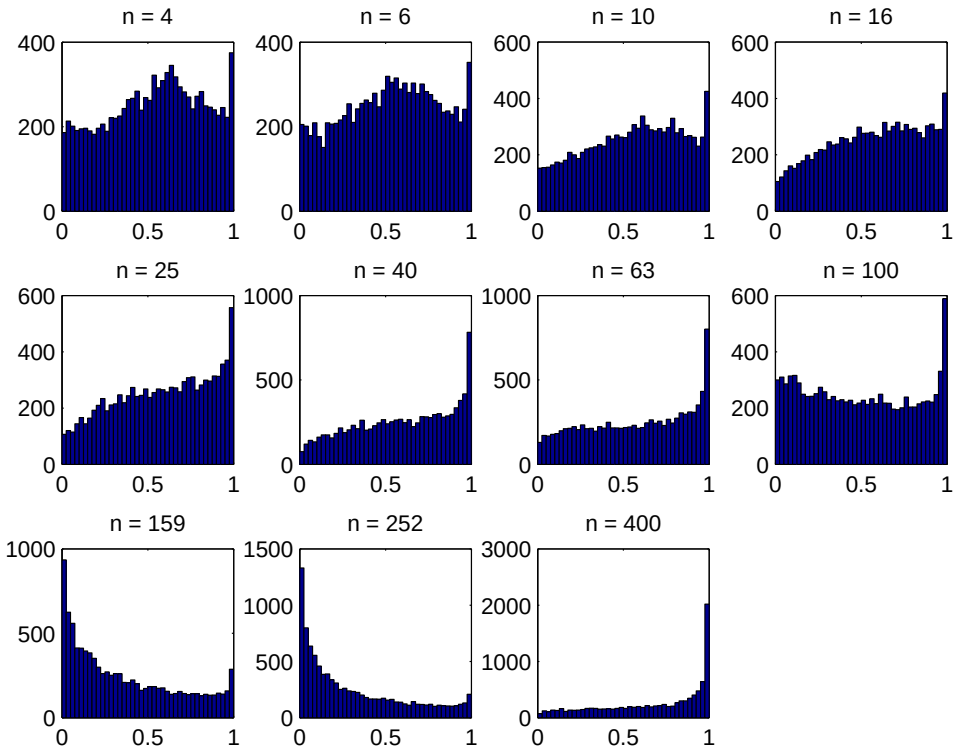

FIGURE A.18: The distribution of $P_{MPS\hat{s}}$ under M0 for 11 different sample sizes. Each one is based on 10,000 simulations. It would have been nice if these had been uniform over [0,1]; or at least if they had been more similar to one another—or even if they evolved in a predictable way. But they do not.

distribution of $P_{MPS\hat{s}}$. Let us simulate the distribution of the statistic for 11 different sample sizes, evenly separated on a logarithmic scale: 4, 6, 10, 16, 25, 40, 63, 100, 159, 252, and 400. Figure A.18 shows these distributions. Each one is based on 10,000 simulations.

It would have been nice if these distributions had been more uniform. Barring that, it would have been nice if they had been more similar to one another. Barring that, it would have been nice if, as the sample size increased, the distribution of $P_{MPS\hat{s}}$ had evolved in a nice predictable way. But it does not. Instead, as n increases we get an initial "sloshing" of the distribution towards values close to 1, until we

reach $n = 159$, at which point the distribution "sloshes" back towards small values. But then at $n = 400$, the largest sample size we'll consider in this paper, there is once again an abundance of large values of the statistic. This unhappy, irregular behavior of the distribution is likely due to two decision that I made earlier when simplifying the process of computing $P_{MPS\hat{s}}$ from the values of MPS and $\hat{s}$: I decided to approximate every distribution of MPS with a Gamma distribution, and I decided to estimate the two parameters of that Gamma distribution with functions of $\hat{s}$ and n that were based on polynomial regression. That regression has likely produced imperfect estimates of the α and β parameters of an imperfect Gamma model. And since the polynomial surfaces we used to estimate α and β were sometimes overestimates and sometimes underestimates of the αs and βs that resulted more directly from the simulations, that is why our distributions now "slosh" around with varying values of n. Recall in the previous section that the residuals from the α surface were cause for concern, but that we would "proceed with caution". Well, now we have proceeded and we must pay the price. We will see late in this chapter that the price is not trivial. [8]

However, rather than beat ourselves with a stick, it is worth noting that at each point along the way, we have faced problems with no easy solutions. The distribution of MPS is not exactly Gamma, but it may not conform to any particular functional form. α and β are not related to MPS and $\hat{s}$ in a polynomial way exactly, but once again: there may be no functional form relating them. We'll keep the simulated density estimates we have, and we'll only consider data sets having sample sizes among those we've simulated: 4, 6, 10, 16, 25, 40, 63, 100, 159, 252, and 400. That

[8] **Spoiler Warning!** We will in fact see that because the distribution of $P_{MPS\hat{s}}$ under the model M0 is suddenly very right-skewed when $n = 400$, the statistic loses much of its power to distinguish the one-planet model from the zero-planet model for that sample size. When that time arrives, it will be clear that it is a problem. But we will then recall the source of that problem (this, right here, the polynomial regression) and by understanding its source, we will be able to tell what part of our analysis is useful and what part is not.

should be all we need to see how well our statistic is performing for different sample sizes.

But there is a good reason to perform one final transformation. We'll look at that in the next section.

A.5 One last transformation

We now see that instead of $P_{MPS\hat{s}}$ having a distribution that is uniform over the unit interval, it tends to produce an abundances of values close to either 0 or 1, depending on the sample size. This is not so problematic near 0, since a small value of this statistic would never lead us to suspect a planet was present. But it will be a problem near 1, since that is precisely the region of interest to us. We will be asking questions such as: How many times more likely is model M0 to produce our observed $P_{MPS\hat{s}}$ than is model M1 with thus-and-such a planet? So the density of $PMPS\hat{s}$ for values near 1 will matter a great deal. But right now it's hard to estimate that density. It will be easier if we transform the statistic one last time so that it spreads out a great deal over the positive real numbers. A natural choice is the logit transformation. While we're at it, let's also admit that the name of our statistic has grown increasingly cumbersome with each transformation. So let us now define: $Q = logit(P_{MPS\hat{s}}) = \log(P_{MPS\hat{s}}) - \log(1 - P_{MPS\hat{s}})$. If Q is positive and big, it suggests a planet. If Q is negative and big, it does not.

The shape of the distribution of Q will not necessarily be symmetric, since we've seen in Figure A.18 that $P_{MPS\hat{s}}$ did not have a very symmetric distribution. However, it seems likely that the distribution will be unimodal. There is a four-parameter family of distributions called *skewed t* distributions that will come in handy right now. [13] The skewed t distribution family has four parameters: η, which describes the center of the distribution; σ, which describes the spread; ν—the counterpart to the degrees of freedom in the standard t distribution—which describes the kurtosis; and

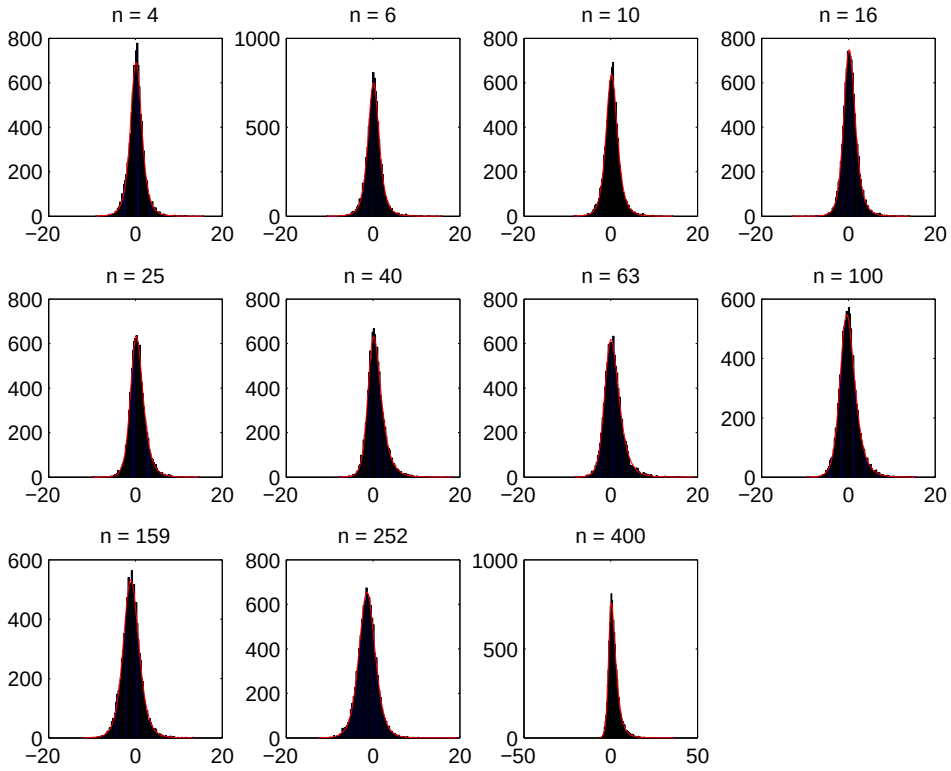

FIGURE A.19: For each of the 11 sample sizes under consideration, we look at the distribution of Q (in blue) and the "best fit" (by the MLE criterion) skewed t distribution (in red).

α, which describes the skew. We will simply denote this function as $t(x; \eta, \sigma, \nu, \alpha)$, or sometimes $t(x)$, if the parameters are obvious in context.

Figure A.19 shows a histogram of the simulated values of Q for the 11 sample sizes we're considering. Superimposed on each histogram is a rescaled skewed t distribution, having parameters determined by the MLE for the simulates. Figure A.20 shows the same thing, only as cdfs rather than pdfs.

The plots show satisfying empirical fits. It should be fairly accurate to use the skewed t distribution to estimate the actual density of Q. Figure A.21 shows plots of the four parameters of the skewed t distribution against n. The plots are not well-

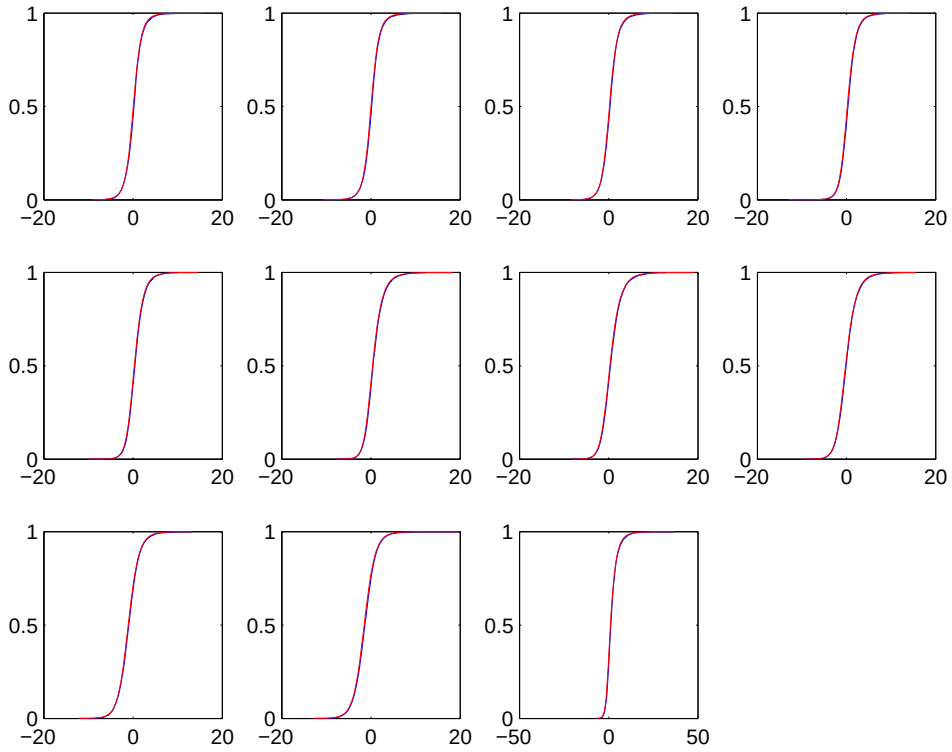

FIGURE A.20: The cdfs of the Q simulates (in blue) and the best-fit skewed t distribution (in red). They are so similar that we can be reasonably comfortable using the skewed t to estimate the density of our statistic.

behaved. The unpredictable behavior of these parameters reflects the "sloshing" of the distribution of $P_{MPS\hat{s}}$ with n that we saw in Figure A.18. For example, the very large skew parameter α corresponding to $n = 400$ reflects the fact that the distribution of $P_{MPS\hat{s}}$ when $n = 400$ had a lot of values near 1.

This is not very pretty—don't we always like to see predictable patterns?—but it doesn't undermine everything we've done. So long as we stick to the sample sizes we've already simulated, we have a reasonably good estimate of the density of the distribution of Q, and we can estimate it quickly at any Q.

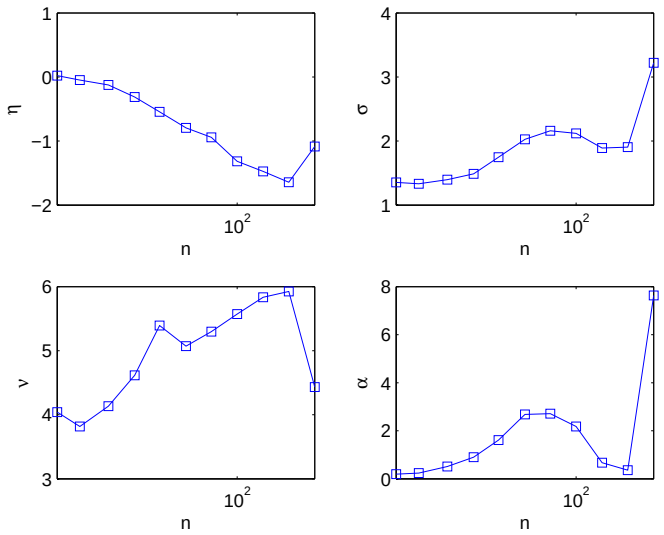

FIGURE A.21: MLE estimates of the four parameters η, σ, ν, and α of the skewed t distribution, plotted against the sample size n. The erratic behavior of the parameters reflects the "sloshing" of the simulations with changing sample sizes, and can trace its roots back to the use of polynomial regression to estimate parameters of the Gamma distribution.

A.6 Five versions of model M1

At last we are ready to see how well this statistic Q, derived from the height of the tallest periodogram spike, performs when it comes to deciding whether a planet is present or not. We will simulate seven different scenarios.

The main guiding principle for choosing the scenarios is variety in planet sizes and eccentricities. However, the detectability of a planet depends not upon its size per se but upon its host star's radial velocity function's semiamplitude K. When K is very large, it is fairly easy to detect a planet. Indeed, with stellar jitters being rarely larger than 12 m s^{-1}—and often considerably smaller—then if K is, say, 20 m s^{-1}, then it may be possible to detect with great confidence the presence of a planet with as few as 2 observations! It would of course take at least several more observations than that to nail down the orbital characteristics of the planet, but 2

may be sufficient to detect the planet's presence—and that is all we are looking at in this paper.

All other things being equal, a large K indicates a large planet. But the size of the planet also is related to the period of the orbit, for a planet that has a large period is far from its host star, and thus has little gravitational pull on it, which lessens K.[9] Witness, for example, the star HD117207. It hosts an orbiting planet and its r.v. semiamplitude is only $K = 26.6$ days, rather small compared to most other detected planets. But its period is 2597 days, over 7 years. That puts the planet at a great distance from its star, so its gravitational influence is weakened. In fact, this is a fairly massive planet, at least 1.88 times the mass of Jupiter. In fact, this is one of several sources of selection bias among exoplanets. Planets that are far from their host stars are hard to detect not only because you have to patiently make observations over many years, but also because they must be massive enough to induce a K that is distinguishable from stellar jitter. So small planets far from their host stars are selected against, and none have been detected so far. That could be because there are none, or because we just can't see them.

So among our scenarios we will not focus on the size of the planet, but upon the magnitude of K. And we will continue in our simulations to do as we've done before, simulating data as random times uniformly distributed over several periods. But the choice of a period is fairly arbitrary in these simulations. It has no effect on the periodogram other than scale (of period, not MPS). So we will choose an arbitrary period of $P = 7$ days for all of our planets. Only K determines the detectability of a circular-orbit planet. So in our interpretation we must keep in mind that for a given detectable K, a small period would correspond to a smaller planet, and a large

[9] If the mass of the star m_s is measured in sun's masses, the velocity of the star v_s is measured in ms^{-1}, and the period of the star P is measured in days, then the mass of the planet, measured in masses of Jupiter, is approximately equal to $m_p \approx \frac{1}{200} v_s m_s^{\frac{2}{3}} P^{\frac{1}{3}}$.

period to a larger planet.

In addition to a variety of K magnitudes, we will also consider two additional eccentricities corresponding to a K that yields moderate detectability when $e = 0$. The idea is that eccentric orbits may be harder to detect than circular ones, since the star may go longer spells without much variation in its radial velocity. Also, the periodogram is designed to detect sinusoidal patterns, so its spikes may not be as tall when the patterns are Keplerian but not sinusoidal.

And finally, we will make the supposition (known to be unrealistic, due to various selection biases) that the stars in our galaxy have a distribution of planets very much like those 300-odd ones that we've already detected. We'll mix them all up and consider sampling from them, and seeing how our statistic performs.

Thus, the following seven scenarios will be considered. A justification is given with each one.

A.6.1 $K = 50$, $e = 0$

In this section we will simulate many sets of observations from a star that does in fact host a companion planet. We will use the same distributions of measurement errors and of stellar jitter as before. We will (arbitrarily) let $P = 7$ days. And we will consider here only circular orbits ($e = 0$) and the rather large perturbations that induce $K = 50$ m s^{-1}.

Let us first consider the case of $n = 6$ observations, which will be useful as an illustration. We simulate 10,000 data sets of the sort described above and compute Q for each one. Figure A.22 shows four plots for comparing the simulated values of Q to the best-fit skew-t distribution. The first is a histogram of Q with the skew-t pdf superimposed. The second is the ogive of Q with the skew-t cdf superimposed. So far so good. The next is a qq-plot, comparing Q simulations to the skew-t quantiles. It doesn't look so good. But the fourth plot is a qq-plot with the outer 1% of the

simulations trimmed (i.e., $\frac{1}{2}\%$ on either side). Now the qq-plot looks much better. But is that "fair", to just toss away the outer 1% of the simulations because we don't like them?

The answer is, more or less, Yes, it is fair. The reason is that our purpose for having the skew-t approximation for the distribution of Q is so that we can estimate the *probability* that the statistic Q will lead us to various conclusions. And 99% of the time, the skew-t density is a very good approximation of the actual density of Q, as seen in the trimmed qq-plot. So although the skew-t is not a good density approximation of Q's density in the remote tails, the fact is that values in those tails occur so seldom that probability estimates of Q behaving in various ways will be reasonably accurate when approximated by the skew-t.

Still, we note that there is a deviation in the trimmed qq-plot at around Q=55. Inspecting the simulations reveals that the greatest gap between the predicted Q and the actual Q occurs at a place where the predicted Q is about 50.0 while the corresponding actual Q is 48.7. The ratio of the skew-t pdf at these two values is about 0.51, a discrepancy that we can live with. It will cause an occasional underestimate in the density of Q, resulting in apparently weaker evidence of planetary presence. But even that occurs rarely: for more than 90% of the simulations, the estimated pdf is between 0.8 and 1.2 times the actual pdf. And so, onward.

We may similarly simulate Q under the same model for the other sample sizes that are under consideration, ranging from $n = 4$ to $n = 400$. The best-fit skew-t distributions are not shown here for all of these, as they are similar to those in Figure A.22. They, too, show their greatest deviations between Q and the predicted Q in tails, but they similarly have density ratios that deviate greatly from 1 only rarely. Thus, once we estimate the four defining parameters of each sample size's resulting skew-t pdf, we may use those parameters to compare the estimated density of Q under this model M1 to the estimated density of Q under the model M0 that was

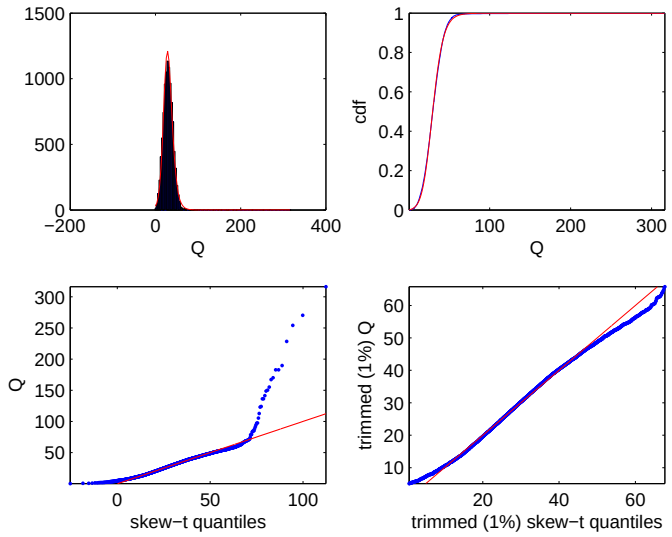

FIGURE A.22: Under this version of model M1, the distribution of Q is reasonably well approximated by a skew-t distribution over 99% of the time, as these plots show. Within the middle 99% of the simulations, at the point of largest discrepancy between the Q simulation and the corresponding skew-t quantile, the ratio of the pdfs is pdf(T quantile)/pdf(Q) ≈ 0.51. That discrepancy is one we can live with. In this case it will tend to underestimate the density, which will lessen the apparent evidence in favor of planetary presence, but only by a factor of about 2, and even then only a fraction of the time. For more than 90% of these simulations, the ratio of the skew-t pdf estimate to the actual density is between 0.8 and 1.2.

discussed earlier.

But then, what to do with the information? We will consider two ways to evaluate the effectiveness of Q at planetary detection. The first will be using a series of "ROC curves".[10] These curves show true positive detection rates as a function of false positive detection rates, when a frequentist hypothesis test is performed with an unspecified threshold for the rejection region. For example, Figure A.23 shows such a curve for the $n = 6$ case. The blue dotted line shows what the curve would look like if the information in the statistic were useless for distinguishing model M0 from model M1: the true positive rate would be equal to the false positive rate, and you

[10] ROC stands for "receiver operating characteristic" and comes from signal detection theory.

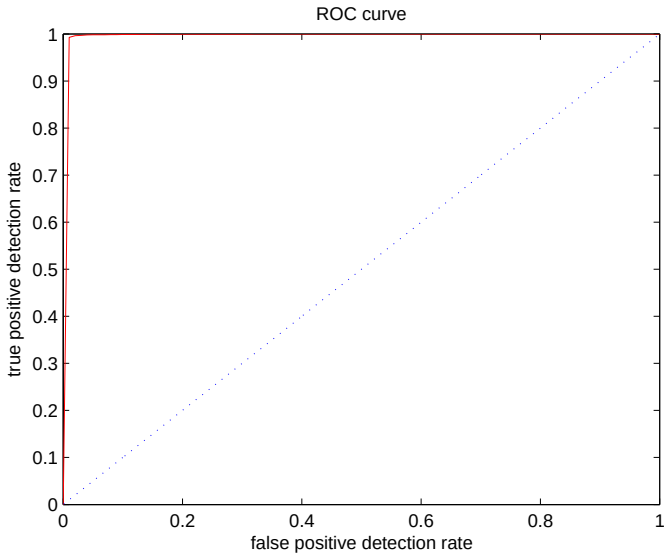

FIGURE A.23: ROC curve when $n = 6$. The blue dotted line of equality between true and false detection rates corresponds to random guessing, completely uninformed by the data—or, put another way, it corresponds to using the information in the data so poorly that you might as well be guessing. The red plot is the ROC curve for the statistic Q with $n = 6$, comparing this particular model M1 to the no-planet model M0. The higher the curve, the better: and this ROC curve is very high indeed.

might as well just be guessing randomly. The red curve is the ROC curve, and is quite appealing: for nearly all false detection rates—even quite small ones—the true detection rate remains very high. That indicates that Q is extremely useful for detecting a planet, even with a sample size as small as $n = 6$. This is quite reassuring. Our intuition tells us that when $K = 50ms^{-1}$—much larger than the largest plausible combination of both stellar jitter and measurement error—we shouldn't require very many observations to detect a planet. But with so few data, the periodogram will have spikes all over the place, and it is not *obvious* that the height of the tallest one would have a distribution much different from two data taken from a non-planet-hosting star. Yet happily, the two distributions are evidently quite different, for this ROC curve shows that Q is quite useful.

Clearly, Q would be even more useful for larger data sets. The only case we may

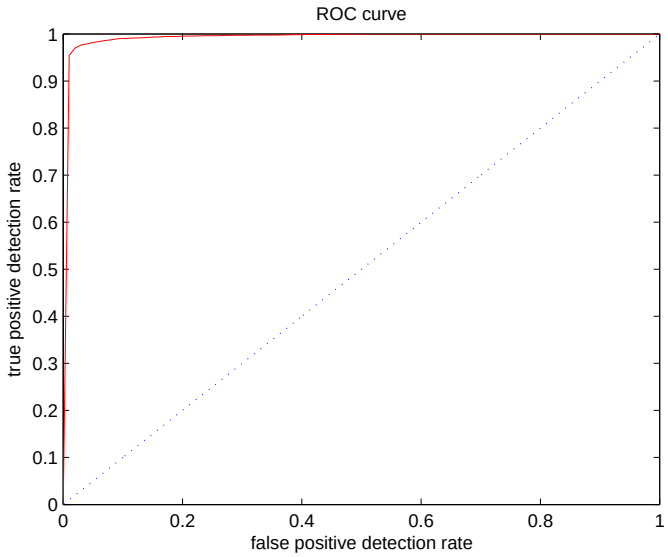

FIGURE A.24: ROC curve when $n = 4$. With fewer data than in the previous case, there is a greater chance that the observed radial velocities will happen to be sufficiently similar so as to have plausibly come from a non-planet-hosting star. But even so, the ROC curve is still quite satisfying. Clearly, Q is a good statistic for detecting this type of planet.

want to consider still is $n = 4$. Does the statistic Q still help us distinguish M0 from M1 so easily? Figure A.24 shows the ROC curve for this case.

We see that even with $n = 4$ observations, the ROC curve is still quite satisfying, indicating that even for very small false detection rates, the true detection rate is still quite high. But we also note that the ROC curve is slightly lower than it was in the case of $n = 6$, the reason being that with fewer observations, there is a greater chance that they will all happen to be sufficiently similar so as to have plausibly come from a non-planet-hosting star.

A useful way to summarize the usefulness of a statistic is to compute the area between the line of uselessness ("$y = x$") and the ROC curve, and then double it. This must produce a number between 0 and 1, and the greater the number, the more useful the statistic. For this reason, we will name this number the "Index of Model

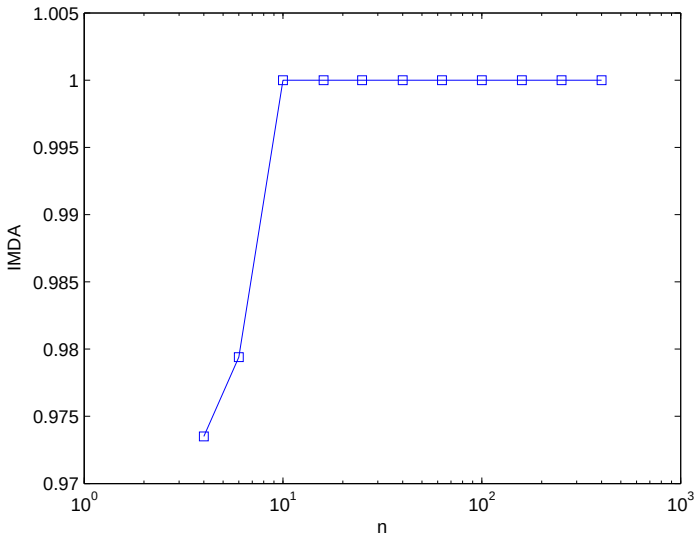

FIGURE A.25: Estimated IMDA's for values of n from 4 to 400. For all $n \geq 10$, the entire collection of simulated Q's under model M1 were greater than all those under model $M0$, yielding an estimated IMDA of 1.

Distinguishing Ability", or the IMDA for short. In figure A.25 we see a plot of the IMDA's for all the sample sizes we're considering.

What is a good IMDA? Many researchers would like the power of a hypothesis test to be at least 0.8 when their significance level is $\alpha = 0.05$. That corresponds to the point (0.05, 0.80) on the ROC curve—a true detection rate (i.e., rejection of the null hypothesis) of 80% when the false detection rate (i.e., Type I error rate) is 5%. The triangle connecting the points (0,0), (0.05, 0.80), and (1,1) on the ROC curve would have an area of 0.375, giving an IMDA of 0.75. So The IMDA's in Figure A.25 are all indeed quite good by that standard.

We note in this plot that once we hit $n = 10$ observations, the entire collection of Q's from model M1 were greater than all those from model M0, leading to an IMDA estimate of 1—a flawless ability to distinguish model M0 from model M1. In fact, the IMDA must be slightly less than 1, but by such a negligible amount that we may consider Q to be conclusive.

Clearly there is no need to look at the actual ROC curves for $n \geq 10$. However, we will look at the ROC curves for all sample sizes in the remaining six sections, when the planets will be harder to detect.

But before continuing on to those other planets, let's look at a different way we might evaluate the usefulness of Q. The ROC curves are based on a frequentist decision-making rule: reject M0 in favor of M1 if Q exceeds a given threshold. That is classical hypothesis testing. A different way to make a decision is based on likelihood ratios. Let L be the ratio of the pdf of our observed Q under model M1 to its pdf under model M0.

So what do our simulations of Q from model M1 say? Using our estimate of the skew-t density for the $n = 6$ case, they say the following:

- Prob(Very strong correct evidence favoring M1) ≈ 0.977;
- Prob(Pretty strong correct evidence favoring M1) ≈ 0.011;
- Prob(Ambiguous evidence) ≈ 0.011;
- Prob(Pretty strong misleading evidence favoring M0) ≈ 0.001;
- Prob(Very strong misleading evidence favoring M0) ≈ 0.

(We could do a similar thing using simulations from model M0, but that would not be as useful. Here we are addressing the question: if we happen to be looking at a star hosting a planet such that $K = 50ms^{-1}$ and $e = 0$, what is the probability that Q will indicate that a planet is present (and how strong will the evidence be)? But using simulations from M0 and making this comparison would be addressing a peculiar question: if we happen to be looking a star hosting no planet, what is the probability that Q will provide evidence favoring *this particular type of planet*? Why would we single out that kind of planet? In the final section of this chapter, when we

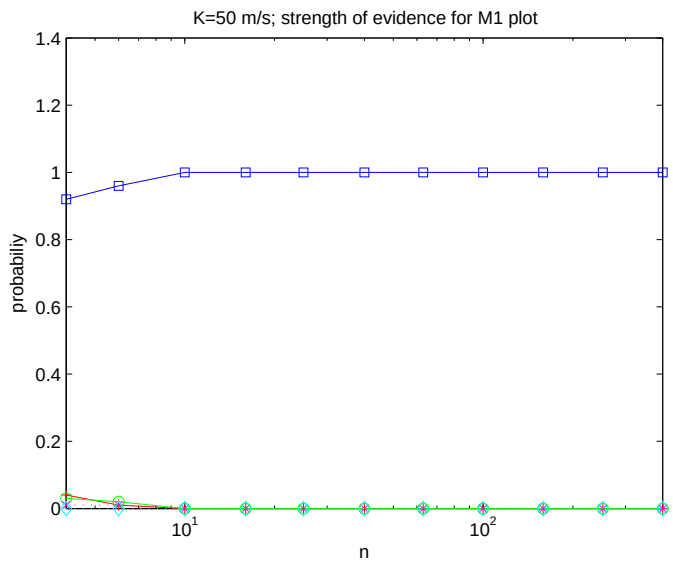

FIGURE A.26: Although it's hard to distinguish four of them, there are five plots here: (1) blue squares show Pr(very strong correct evidence favoring M1); (2) red plusses show Pr(Pretty strong correct evidence favoring M1); (3) green circles show Pr(ambiguous evidence); (4) magenta x's show Pr(pretty strong misleading evidence favoring M0; and (5) cyan diamonds show Pr(very strong misleading evidence favoring M1). The evidence for M1 over M0 is strong in all cases, but does get even stronger for larger n, as we would expect.

are simulating data from a distribution of stars, some of which host planets and some do not, we can then look at the probability of these different types of evidence *in general*; i.e., the probability of strong correct evidence, strong misleading evidence, and so on.)

We can do the same thing for all of our different sample sizes and make a plot of the probabilities of the types of evidence against the sample size. Such a plot is shown in figure A.26. The evidence is so strong favoring M1 for all the sample sizes that four of the plots can hardly be distinguished. However, we will look at similar plots for cases where the planet is harder to detect, and what we would like to see in those cases is how many observations are required before the evidence favoring M1 becomes compelling.

And now at last it is time to move on to those other types of planets. Now that we have discussed our methods of evaluating Q, we won't have to go into those methods again at great length. We may, for each of the remaining cases, simply look at the following:

- plots of ROC curves for different sample sizes;
- a plot of IMDA versus the sample sizes; and
- a plot similar to Figure A.26, showing the evolving probabilities of the different types of evidence in Q as n increases.

In future chapters we will try to produce similar plots for two alternatives to Q: the maximum likelihood ratio, and the Bayes factor. Comparisons of those plots to these will enable us to decide whether these statistics differ in their usefulness as planet-detection summary statistics.

A.6.2 $K = 20$, $e = 0$

We now consider planets with circular orbits having $K = 20ms^{-1}$. Recall that if $P = 7$ days, as in our simulation, then this corresponds to a planet whose minimal mass is about 0.2 times the mass of Jupiter, or about 60 Earth masses. But it would be about as easy to detect a planet with $K = 20ms^{-1}$ and $P = 3$ days, which would be about 25% smaller–and among the smallest 8% of planets so far detected, including those small planets detected as parts of planetary systems.

Figures A.27, A.28, and A.29 show, respectively, ROC curves for the 11 different sample sizes under consideration, the IMDA as a function of sample size, and the probability of different types of evidence found in the likelihood ratio.

And now, Dear Reader, we must now at last pay the price for having too-poorly approximated the distribution of $P_{MPS\hat{s}}$ under M0 by a Gamma distribution—or,

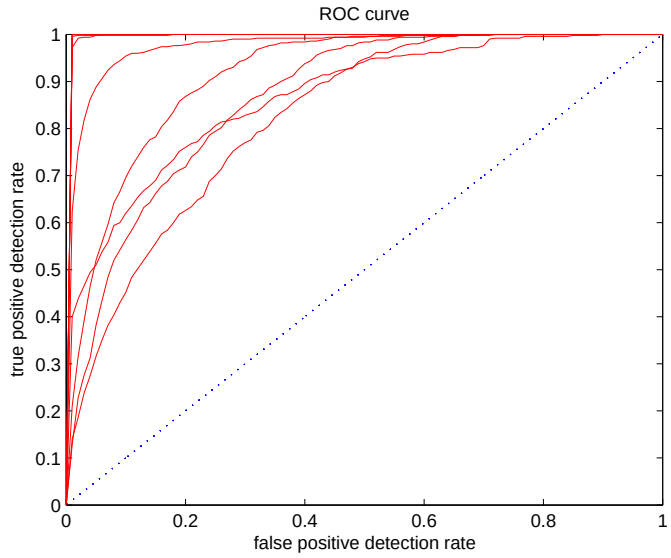

FIGURE A.27: The ROC curves at first appear as you would expect, receding from the line of non-information as the sample size increases. But in fact, the ROC curve corresponding to $n = 400$ is the among the three *lowest* curves, rather than the highest one.

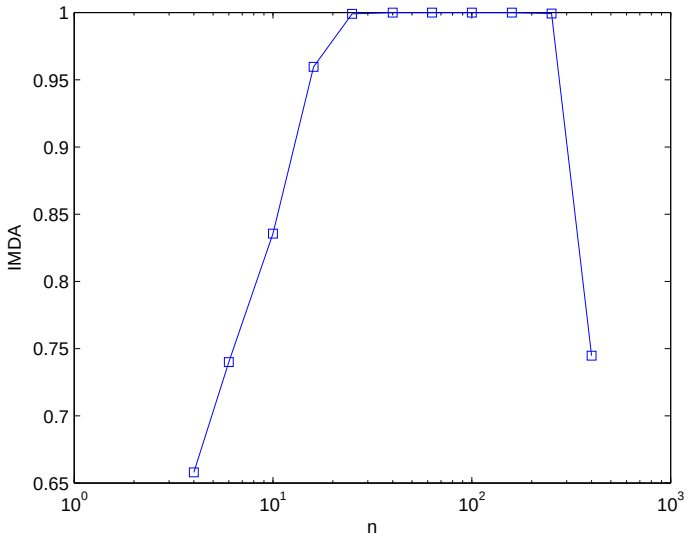

FIGURE A.28: The IMDA increases with increasing sample size... until we reach $n = 400$. Then the model-distinguishing ability plummets. This is the "price we would pay" that was referred to in an earlier section.

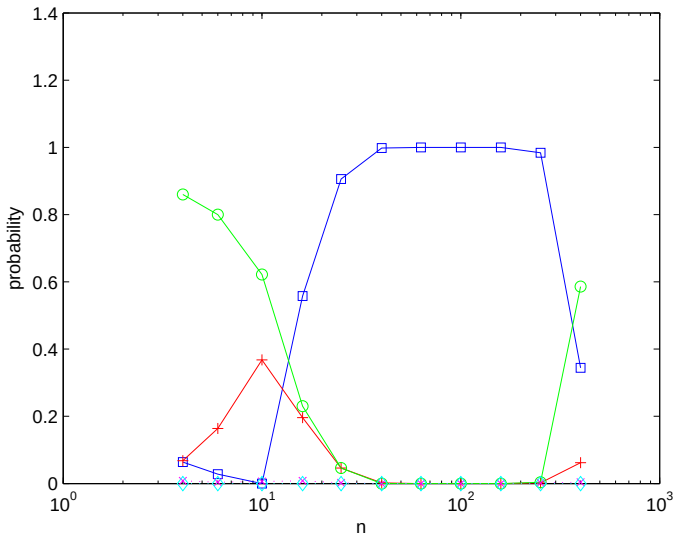

FIGURE A.29: There are five plots here: (1) blue squares show Pr(very strong correct evidence favoring M1); (2) red plusses show Pr(Pretty strong correct evidence favoring M1); (3) green circles show Pr(ambiguous evidence); (4) magenta x's show Pr(pretty strong misleading evidence favoring M0; and (5) cyan diamonds show Pr(very strong misleading evidence favoring M1). As with the ROC curves, the utility of Q when $n = 400$ diminishes relative to smaller sample sizes. We also see here a small diminution of utility going from $n = 4$ through $n = 6$ to $n = 10$.

more likely, for having too-poorly estimated the parameters α and β of that distribution using polynomial regression. Way back then we noticed something peculiar about the distribution of $P_{MPS\hat{s}}$ under model M0 as n increased. Instead of the seeing the distribution evolving in a predictable way, we remarked that it appeared to "slosh around", first towards 1, then back towards 0, and then (this is where our present unhappy situation originated) heavily back towards 1 again at $n = 400$. The residuals in our polynomial regression (see Figures A.15 and A.16 were small relative to the values of α and β that they were estimating, but they were not so small that they were negligible. In particular (hindsight is always 20/20) the residuals of α's regression were particular positive for $n = 400$. An overestimate of α means a too-large mean and variance, creating right-skew. Right-skew in the distribution of

$P_{MPS\hat{s}}$ under M0 means its values will be easily mistaken for ones having come from model M1, which tends to produce larger values of that statistic.

Also we note that the residuals of the β regression were mostly negative for the smallest two sample sizes. This would tend to increase the variance of the Gamma distribution, again blurring values of $P_{MPS\hat{s}}$ that come from M0 with those that come from M1.

This is an unhappy situation to be in, but it need not be fatal. First, in defense of what now reveals itself to have been a poor decision, let us look back once more at Figures **??** and A.14. The surfaces look like awfully good fits. Perhaps the decision to estimate the distribution of $P_{MPS\hat{s}}$ for all sample sizes was itself a problem. At any rate, let us think about the present consequences.

We began with a statistic that I will for brevity call simply P right now: it was the bivariate pair (MPS,$\hat{s}$). We wanted to know how useful P was at distinguishing M0 data from M1 data. A series of transformations turned P into Q. A transformation that does not harken back to the original data cannot impart greater informative power to a statistic than it had already. Thus, P must be at least as useful as Q. So we are not now able, nor could we ever have been able, to say "P is exactly this useful". Rather, we were always required to say only "P is at least this useful—i.e., at least as useful as Q". If it turns out in the next two chapters that Q is about as useful (or more so) than our other two statistics, that's good news for P. If not, it doesn't necessarily mean that P was bad all along; it could be that the transformation from P to Q leached out much of its information.

But there is more. It should be clear that P's usefulness cannot diminish with increasing n. But Q has that unhappy property, at least for this model M1, and for a few sample sizes. The plot of IMDA against n does as we expect it to until $n = 400$. And that plot was not based on any skew-t approximations of the distribution of Q. The strength-of-evidence plot based on likelihood the likelihood ratio L shows a

similar drop in information at $n = 400$, but it also shows a small drop in information going from $n = 4$ through $n = 6$ to $n = 10$. Since that didn't show up in the IMDA plot, which was based entirely on the simulated values of Q, it could possibly be due to the skew-t approximation being too poor an approximation.

We know the likely culprits in creating these problems: Using a Gamma distribution to estimate the actual distribution of $P_{MPS\hat{s}}$; using polynomial regression to estimate α and β parameters for that Gamma distribution using $\hat{s}$ and $\log n$; and using a best-fit skew-t distribution to approximate the density of Q. We do not have an easy solution to the problems apart from returning to the drawing board.

But knowing that P utility cannot decrease with increasing n permits us to get around this problem in one of three possible ways:

- We may simply remove from the IMDA plot and the L strength-of-evidence plot any point corresponding to a Q that appears to have less informative power than one with a smaller sample size.

- We may artificially raise the usefulness estimate of Q for any such point to that of its most useful predecessor, forcing monotonicity.

- We may interpolate between points of increasing usefulness to estimate the usefulness of a point whose usefulness drops, or in the case of $n = 400$ where interpolation is not possible, we may extrapolate from its predecessors.

The last of these solutions has a certain appeal, since it would produce the "prettiest" plots. But we should not forget that this is how the present problem arose in the first place: polynomial regression is, after all, essentially interpolation. So shall we drink from that poisoned fountain again?

Yes, we shall, but we won't dunk our heads all the way in as we did before. Since this is the end of the road for the analysis, any errors we make at this point can

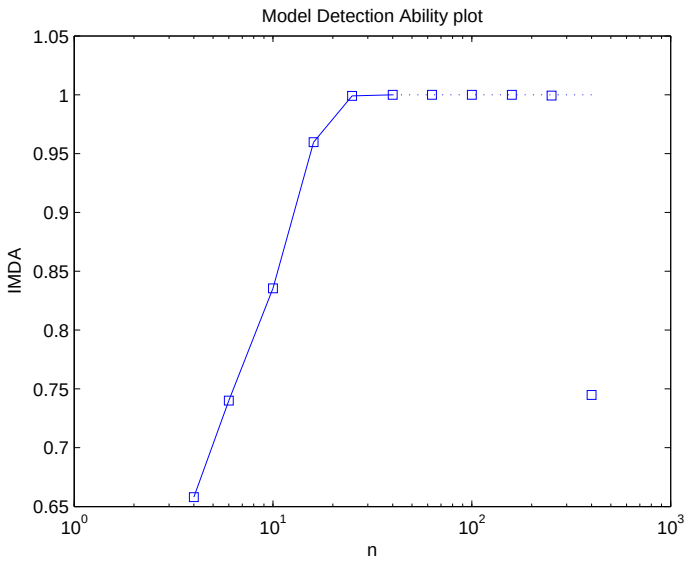

FIGURE A.30: In this new "corrected" IMDA plot, we take advantage of the fact that although Q's usefulness diminishes at $n = 400$, the information in P cannot.

no longer propagate. And what we'll do in our plots is draw all the original points generated by the analysis method described so far, but we'll connect them with dotted lines wherever we are interpolating or extrapolating. For example, Figures A.30 and A.31 show the new, "corrected" IMDA plot and the L-strength-of-evidence plot.

A.6.3 $K = 7$, $e = 0$

If $K = 7$, $e = 0$, and $P = 3$, then we're talking about a very small planet: about one-twentieth the mass of Jupiter, or about sixteen times the mass of the Earth. Let's see how well Q does at distinguishing such star's data from model M0.

Figures A.32 and A.33 show, respectively, the "corrected" IMDA plot and the L-strength-of-evidence plot.

The first of these two plots (which was constructed without the approximating skew-t distributions) suggests that if one commits to making somewhere between 25 and 40 observations, one can have a pretty high true detection rate of planets

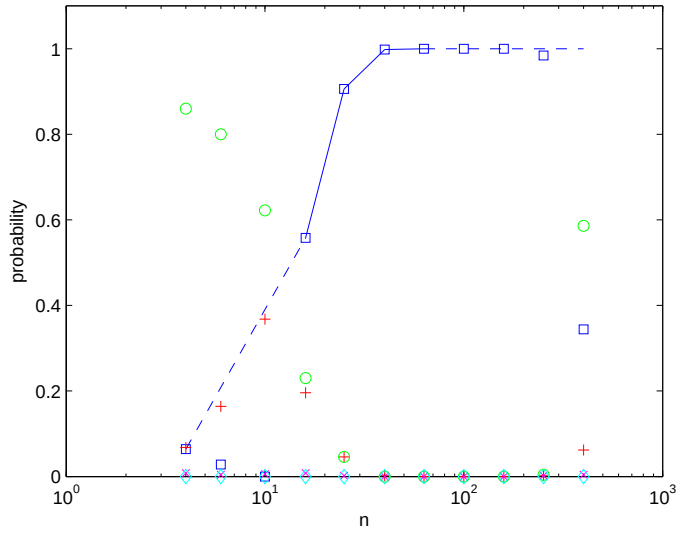

FIGURE A.31: In this "corrected" L-strength-of-evidence plot, we force the probability of "very strong" correct evidence favoring model M1 to be increasing. However, we don't know how the probabilities of "pretty strong" correct evidence and ambiguous information are changing, since they may increase or decrease with increasing n, gaining some probability from the next-weaker evidence class while simultaneously losing it to the next-higher evidence class. So where we have interpolated "very strong" evidence, we don't connect the points for the other plots at all.

like this even while maintaining a small (less than 5%) false detection rate. The second plot (which did use the approximating skew-t distributions, and therefore is somewhat less trustworthy) suggests that at around 63 observations, the chance of the data providing very strong evidence favoring model M1 is at least 80%. These two pieces of evidence are not greatly at odds with one another. In order to detect a circular-orbit planet with a K this small, they suggest a ballpark figure of 40 to 70 observations, committed in advance.

A.6.4 $K = 3$, $e = 0$

Now we come to the first case in which K itself is comparable to the magnitude of stellar jitter, and is considerably smaller than the combination of stellar jitter and measurement errors. Can the signal of a whispering planet be heard through the din

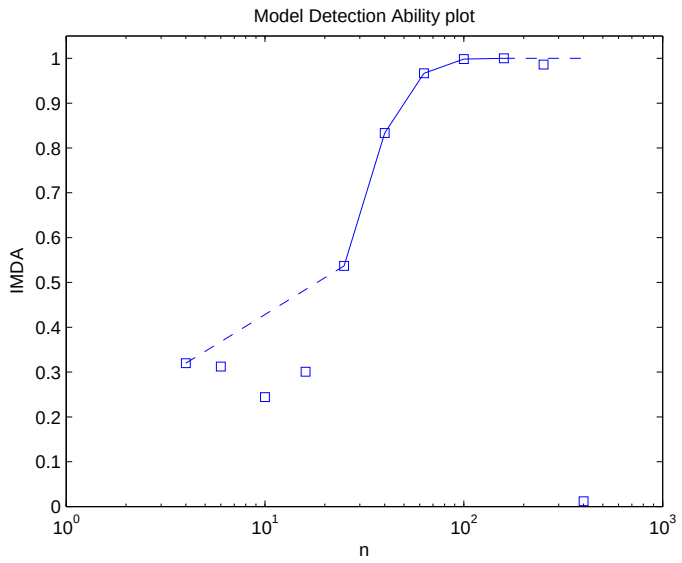

FIGURE A.32: IMDA plot when $K = 7$. Recalling that an IMDA greater than 0.75 is desirable, this plot suggests that occurs somewhere between $n = 25$ and $n = 40$.

of these additional noise sources? The evidence that it can is suggested by Figures A.34 and A.35.

Between these two plots, we get the idea that if we commit in advance to making somewhere between 150 and 250 observations on a star, then we should have a very good chance of picking out this whisper-quiet planet, were it to exist. As of this writing, only one planet that small and "quiet" has been detected which was not part of a multiple-planet system. Probably no stand-alone planet (that is, a planet not known to be part of a larger planetary system) has been found with a K as small as $3ms^{-1}$. The smallest is probably that of HD 4308 b, with $K = 4.07ms^{-1}$. Its minimal mass is about 0.05 times the mass of Jupiter. It was first described in 2005, after 41 observations were made on it.

A.6.5 $K = 20$, $e = 0.3$

Now we leave circular orbits and small semiamplitudes to consider another way in which planets may be hard to detect, particularly with a periodogram, which is

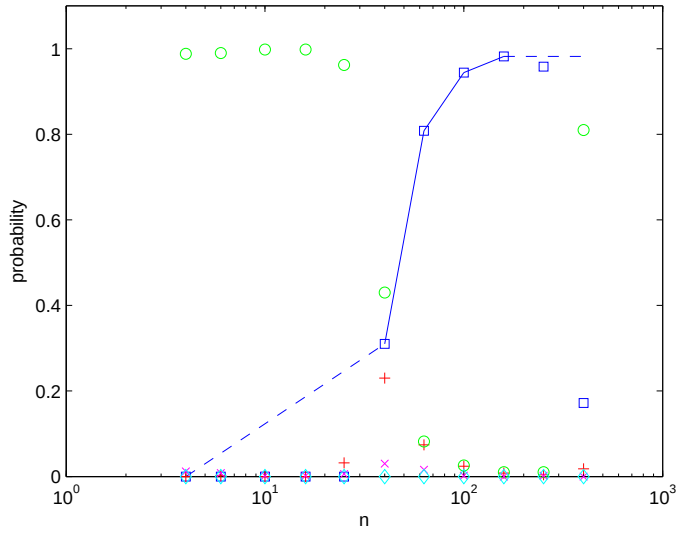

FIGURE A.33: The IMDA plot suggested that $n = 40$ would be a sufficient number of observations to detect this sort of planet. In this plot of "strength of evidence", we find a suggestion that as many as $n = 63$ observations are required before there is a high probability of "very strong" evidence correctly pointing to model M1. But since these calculations (unlike the IMDA) rely on the skew-t approximations to the Q distribution, they may imply that the evidence is weaker than it really is.

designed to find sinusoidal patterns. We will here consider planets and stars whose orbits are somewhat eccentric: we'll simulate $e = 0.4$ stars in this section and $e = 0.6$ stars in the next. In both cases we'll let $K = 30$ m s^{-1}, since we found that for circular orbits, a moderate number of observations were required to detect the planet. We now want to see how many more observations are needed when the orbits are eccentric.

We will assume that orientations are arbitrarily distributed, so in our simulation we'll draw both ω and M_0 from the distribution that is uniformly distributed over $[0, 2 * pi)$.

Figures A.36 and A.37 show the now-familiar IMDA plot and the L-strength-of-evidence plot, respectively.

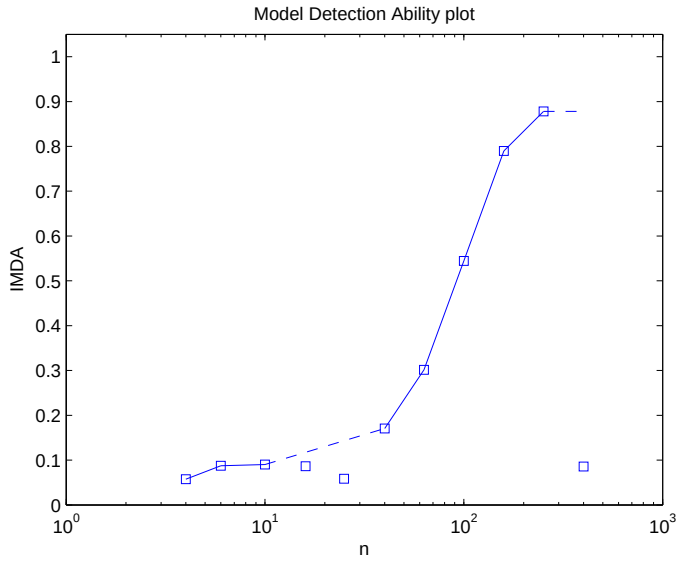

FIGURE A.34: IMDA plot when $K = 3$. Recalling that an IMDA greater than 0.75 is desirable, this plot suggests that occurs somewhere between $n = 100$ and $n = 160$.

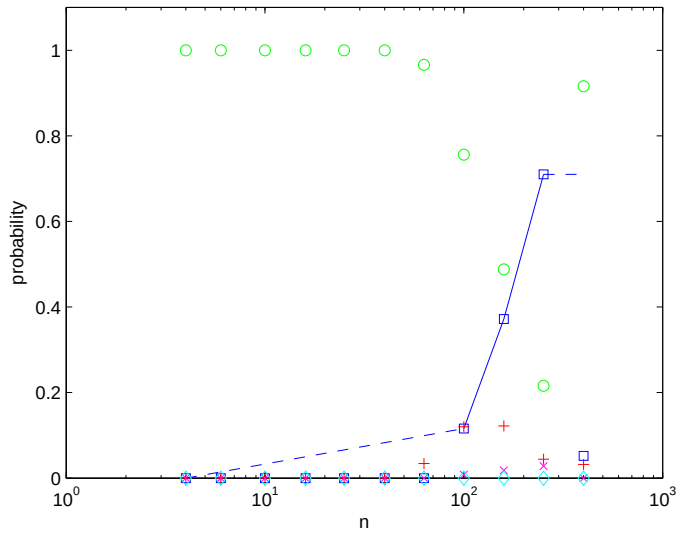

FIGURE A.35: In this plot of "strength of evidence", we run smack up against the problem we created when we badly estimated the distribution of $P_{MPS\hat{s}}$ when $n = 400$. If the apparent trend were to continue, then we would expect a high probability of "very strong" evidence correctly pointing to model M1 when n is just a little above 250 observations.

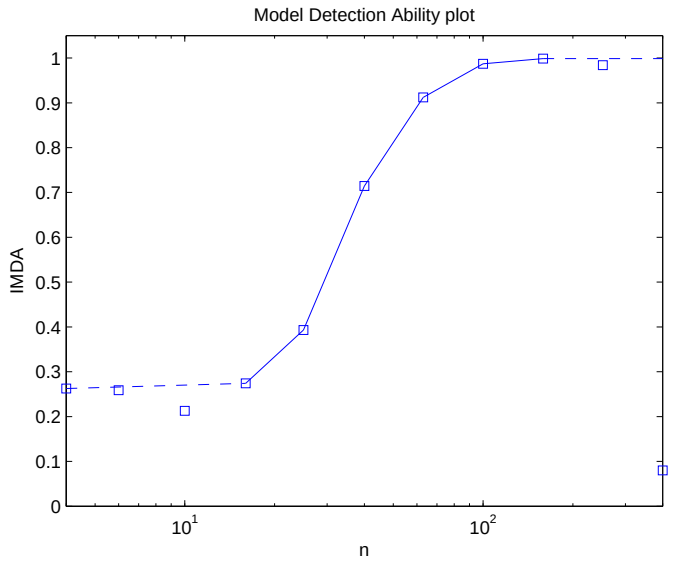

FIGURE A.36: $K = 7$ m s^{-1} and $e = 0.3$ in this IMDA plot.

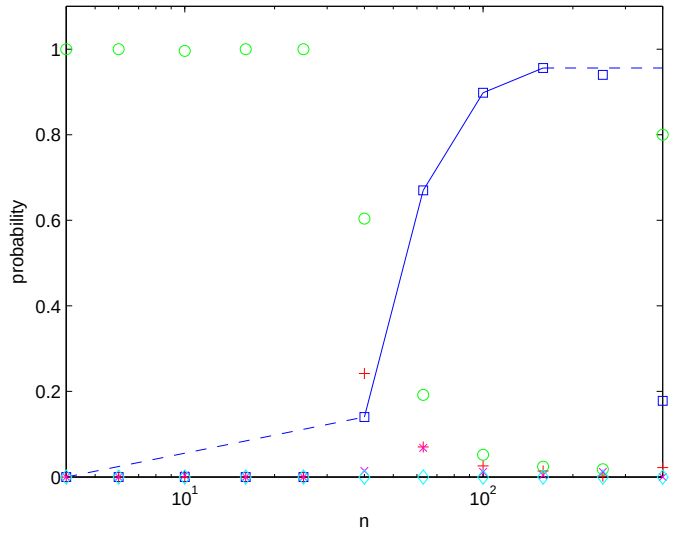

FIGURE A.37: $K = 7$ m s^{-1} and $e = 0.3$ in this L strength-of-evidence plot.

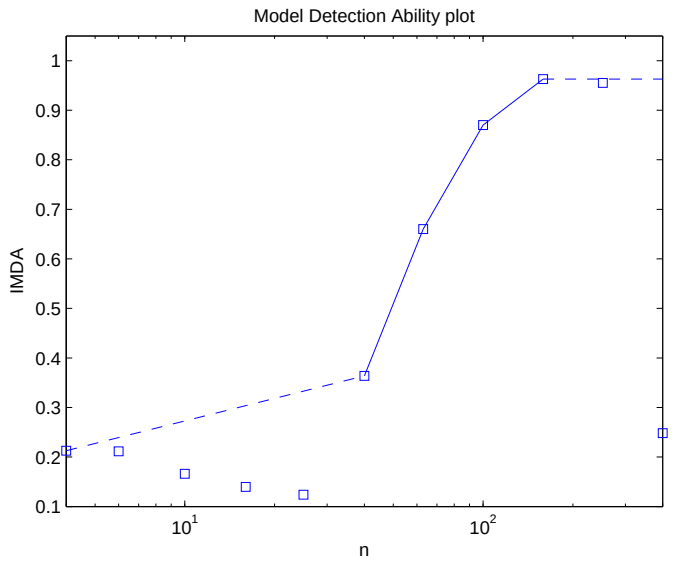

FIGURE A.38: $K = 7$ m s^{-1} and $e = 0.6$ in this IMDA plot.

A.6.6 $K = 20$, $e = 0.6$

More eccentric still: $e = 0.6$.

Figures A.38 and A.39 show the now-familiar IMDA plot and the L-strength-of-evidence plot, respectively.

A.6.7 Summarizing the information so far

Figure A.40 shows the four corrected IMDA plots for the circular orbits all shown together for comparison. Figure **??** shows, for each of the circular orbit scenarios, a plot of the probability of "very strong evidence favoring M1". Both of these plots tell a similar story: with enough patience, even orbits with $K = 3$ m s^{-1} should be detectable. But it requires commitment in advance to making a fairly large number of observations—between perhaps 150 and 300, depending upon the stringency you demand. The IMDA plot suggests fewer numbers of observations are required than does the $L > 32$ plot. One reason may be that L was "blurred" somewhat by using the skew-t distribution to approximate the actual distribution of Q for each sample

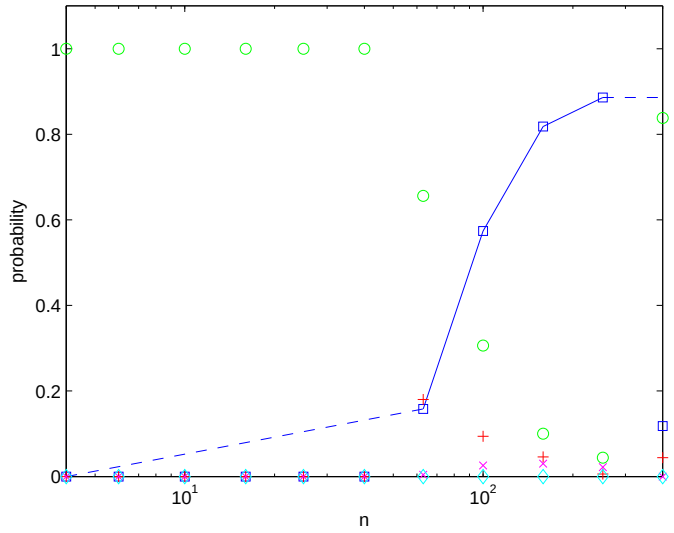

FIGURE A.39: $K = 7$ m s^{-1} and $e = 0.6$ in this L strength-of-evidence plot.

size, and we saw that the difference in density was sometimes (though not often) in error by a factor of 2. Another reason may be simply that requiring an IMDA of 0.75 is a less strict demand than requiring $L > 32$.

But we stress here once again that if you want to see a star with $K < 3$ m s^{-1}, then a commitment must be made in advance to make a fairly large number of observations, even if along the way nothing seems to be happening. And supposing one commits to $n = 300$ observations on a promising star, and then after those observations are all made, still nothing shows up. Then what? Well, one may continue to make observations if one wants, perhaps hoping to find an even smaller K, or one may then give up, knowing that had there been a $K > 3$ m s^{-1}, it would have shown up by now. Since only about one in eight observed stars has revealed any planet so far, it's possible that a commitment to 300 observations on a star may really amount to a commitment to 300 observations on each of 8 different promising stars, still with no guarantees of success. 2400 observations are neither easy nor cheap. And of course, fewer may be required than are suggested in this paper, since

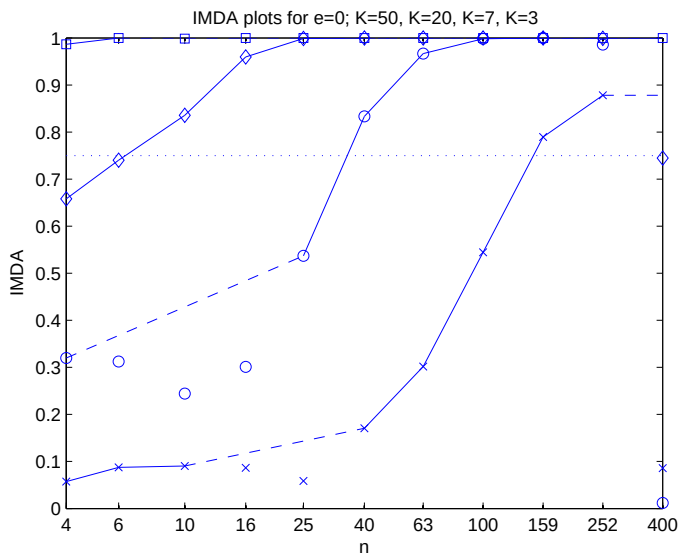

FIGURE A.40: For comparison, the four corrected IMDA plots for the circular orbits are shown together. A horizontal dotted line is shown at 0.75, the IMDA corresponding to a power of 0.8 when the false detection rate is 0.05 The plot indicates that even smallish planets are detectable with enough patience.

we have been simulating data based on random observation times, when in fact smart scheduling has been shown to reduce, at least somewhat, the number of observations required. [15] But still: if small planets are out there, the key to their detection may not be better equipment, but simply greater patience.

Better measuring tools continue to reduce the magnitudes of measurement error, but they can never do anything about stellar jitter. It is possible that the detection of small planets may *always* require patience, since stellar jitter may remain a persistent source of noise in radial velocity data. And unfortunately, once K decreases below the level of noise in stellar jitter, it is not at all certain that the patterns observed in this paper will continue, as regards the relationship between K and number of observations required. Roughly speaking, this paper has suggested the following pattern: when K goes *down* by a factor of ϕ, the number of observations required to discern the patterned wobble goes *up* by a factor of about *phi*. But that pattern

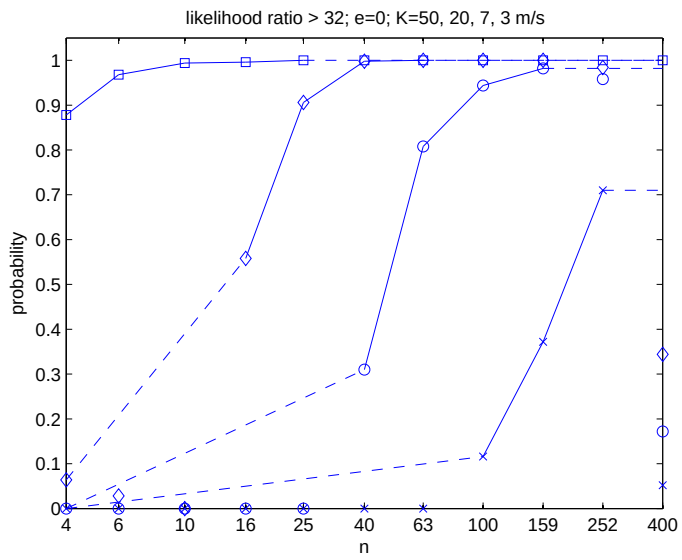

FIGURE A.41: For comparison, the four corrected L-strength-of-evidence plots for the circular orbits are shown together, only showing the probability of "very strong evidence favoring M1 ($L > 32$). As did the IMDA plot, this one indicates that even smallish planets are detectable with enough patience. If it seems to suggest greater patience than did the IMDA plot, that may be because the skew-t approximations of the distributions of Q are not perfect. We saw earlier that they were sometimes (though not often) in error by a factor of 2. Another reason may be simply that demanding $L > 32$ may be a more stringent requirement than merely demanding an IMDA ¿ 0.75.

may not continue once K drops below the magnitude of stellar jitter. Indeed, some simulations not included in this paper suggest that the number of observations requires skyrockets. It is therefore possible that stand-alone planets as small as the Earth, and as cool as the Earth (coolness requiring a distance from the host star that diminishes induced wobble) may forever require an utterly impractically large number of observations to be detected using radial velocities.

Of course there are more interesting things about planets than mere smallness and coolness. But for this writer—for me, Floyd Bullard—what has been the most disappointing thing about writing this dissertation has been the discovery of just how much patience it may require to detect an Earth-like planet using radial velocities.

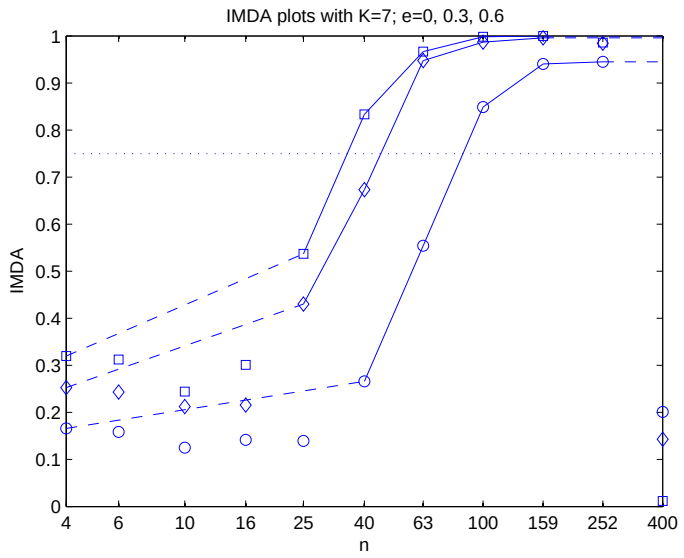

FIGURE A.42: IMDA plots. $K = 7$ m s^{-1}; $e = 0$ (squares), $e = 0.3$ (diamonds), and $e = 0.6$ (circles) are shown.

My interest in exoplanets has always been largely spurred by the hope that someday we would find inhabitable planets out there, and perhaps even life. That may yet happen, of course. But due to the persistence of stellar jitter and the inherent difficulty in picking out tiny patterns from huge noise, the detection of Earth-like planets would appear to be down some path other than radial velocity measurements.

Before continuing to the last section in this chapter, let's also look at plots comparing the IMDA and the probability of $L > 32$ for the simulations in which $K = 7$ m s^{-1} and e is 0, 0.3, or 0.6. These are shown in Figures A.42 and A.43.

Earlier in this chapter we commented that for a given K, it was not obvious whether greater eccentricity would make an orbital pattern easier or harder to detect. What may make it easier is that for a given K, the range of possible radial velocities is greater with greater e, which would make the unusual variation stand out if you were lucky with the times you happened to be looking. But what may make it harder is that when e is larger, the orientation of the orbit relative to the earth may

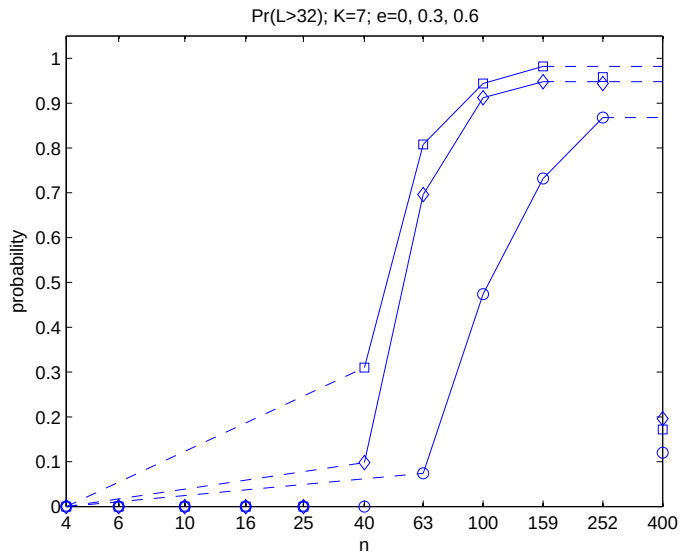

FIGURE A.43: $\Pr(L > 32)$ plots. $K = 7$ m s^{-1}; $e = 0$ (squares), $e = 0.3$ (diamonds), and $e = 0.6$ (circles) are shown.

cause there to be long stretches with relatively little change in radial velocity. If you don't happen to be watching during the "bursts", then you will think nothing is there at all. A third problem related to the MPS uniquely is that the periodogram is designed to detect sinusoidal signals, not the patterns created by more variable Keplerian motion.

So in the end, was it easier or harder to detect planets with greater eccentricity? It was harder, but not hugely so. Orbits with e as large as 0.6 have been found only rarely. These plots suggest that although selection bias may be at work, it is probably not the principle reason that such orbits are so rare. At least as likely an explanation is that they really are relatively rare in the galaxy.

A.6.8 A distribution of stars mimicking those we've observed.

In each of the last six sections of this paper we've looked at particular types of orbits and asked the question: "supposing this were exactly what we were looking for, how good would our statistic be at telling us that we were looking at a planet and not

mere noise?". What we've learned is this: if you are interested in finding planets of a certain smallness (or, more accurately, finding stars with a certain smallness of K), and you are planning to use the maximum periodogram spike to tell you whether you've found a planet or not, we now know a maximal number of observations you need to make, by which time you may quit looking if you still haven't found anything, because chances are slim that what you're looking for is there at all.

But it is not always the case that astronomers are looking for something in particular. They may be looking just for "something"—anything. If that is so, then we may still want to decide when, if we've seen nothing, it's time to quit looking and give up. So in this section we will repeat the simulations and analysis of the last six sections, only this time we will simulate orbits in M1 not as having a particular fixed K and e, but as having roughly the same distribution as those found so far. We will take $\log K$ to be normally distributed with mean $\mu = 4$ and standard deviation $\sigma = 1.2$. And we will take 10% of planet-hosting stars to have exactly circular orbits with $e = 0$. The rest will have eccentricities drawn from a triangular distribution, decreasing linearly as e goes from 0 to 1.

Figures A.44 and A.45 show the IMDA plot and the $Pr(L > 32)$ plot for this simulation.

Both of these plots indicate that it's fairly easy to detect something that's similar to what has already been detected. But that shouldn't surprise us. If we're looking for things like what we've already found, then we're essentially looking for the things in the galaxy that are most easily found. Within just four observations we should be able to tell with great confidence whether the star we're looking at is a planetless star, or if it's one "like" those hosting planets that have already been discovered. But that isn't a great reflection on the statistic. It just reflects the fact that the majority of planets we've found so far induce large wobbles on their host stars. And such a star would be easy to detect the next time, too.

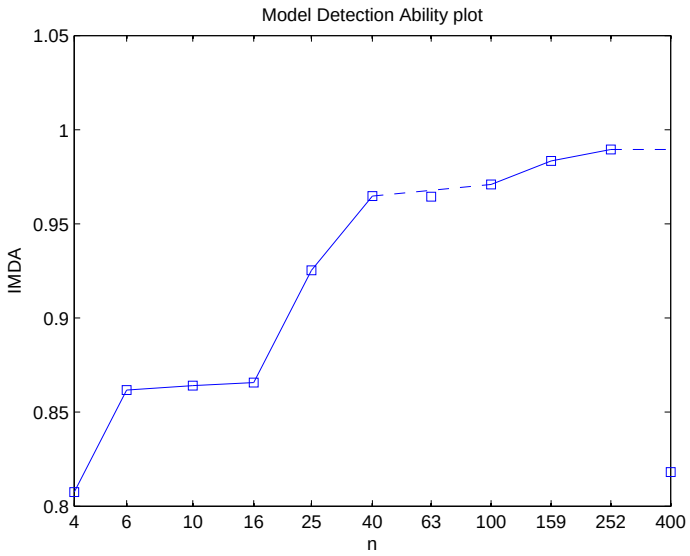

FIGURE A.44: IMDA plot for stars mimicking those in our galaxy for which we've already found something.

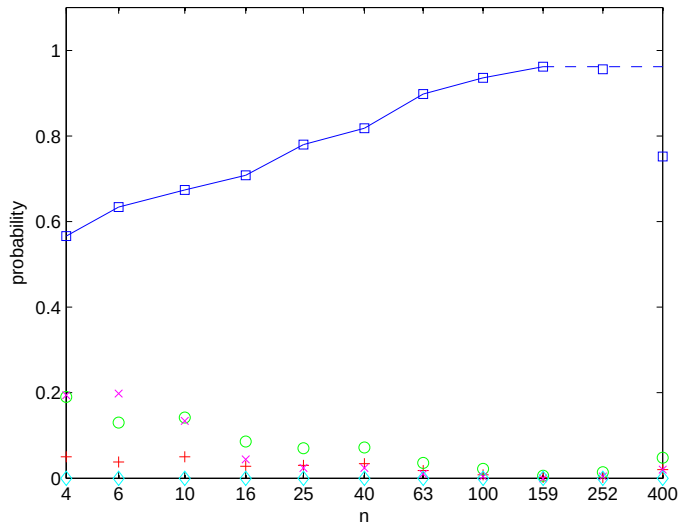

FIGURE A.45: $\Pr(L > 32)$ plot for stars mimicking those in our galaxy for which we've already found something.

If all we were after was finding more of the same, as quickly as possible, then the best use of observation times would be to make four observations on a star, and quit if nothing shows up. (And quit if something does show up, too: more observations will help determine the orbital characteristics, but if our goal is mere detection, then additional observations are of little added benefit.) But that is of course not our only goal. Hence the preceding sections: what if we wanted to find *this?* How hard would that be? What if we wanted to find *this?* Or *this...* ?

A.7 A missed opportunity

I remark here on the fact that an opportunity was missed to study the statistic $MPS|\hat{s}$ in a way that involved fewer approximations, and in particular, fewer forced parmaetric density estimations. At the time when we had "shelves" full of values of MPS, each shelf corresponding to an approximate estimate $\hat{s}$, we did not need necessarily to estimate, or even to preserve, the density of everything on the shelf. All that was ultimately of consequence was a threshold value for the MPS: how large is "too large" to be typical from that shelf. Thus, we might have used the simulations only to determine a quantile—say, the 95th—and then, rather than modeling parameters α and β for a supposed Gamma distribution model, we could perhaps have modeled this quantile directly as a function of $\hat{s}$ and n. That would have given us a tool—a function of $\hat{s}$ and n—with which we could then, for any future real or simulated data set, determine whether the evidence strongly suggested the presence of a planet.

Having already concluded this chapter, I offer this up as a missed opportunity. I will not now return and re-do the analysis, but simply comment that the methods applied in this chapter may not have been ideal, and the possibility remains that the two-dimensional statistic $(MPS, \hat{s})$ may be a very good planet-detecting tool when used most effectively; perhaps it would be more useful than either the MPS alone,

or the likelihood ratio Λ discussed earlier in this paper.

Appendix B

Nested Sampling: An Approach Perhaps Not Best Suited to the Problem

B.1 Nested Sampling: An Approach Perhaps Not Best Suited to the Problem

Recall that the model selection problem under consideration is this: given some radial velocity measurements on a particular star, does it host an orbiting planet or not?[1] Also recall that the zero-planet model involves two unknown parameters—the background motion of the star C, and the stellar jitter s^2–and that the one-planet model involves five additional parameters—the velocity semi-amplitude K, the period P, the eccentricity e, the argument of periastron ω, and the mean anomaly M_0. When performing model selection, we must integrate out all of the additional parameters. This is not a problem in the no-planet model with only two parameters and a unimodal joint distribution to boot. However, in the one-planet model, it is a problem.

Integrating out the seven parameters in the one-planet model is very hard, due to

[1] Another problem for astrophysicists is whether a star may host *multiple* planets. However, in this dissertation we are only considering the zero- and one-planet models.

the number of dimensions and the multimodality of their joint posterior distribution. This was discussed in Chapter 2. In this section we consider an unusual approach to integration, introduced by John Skilling [6]. This is a Monte Carlo approach and is called *nested sampling* due to the way in which additional samples from the posterior distribution are drawn. Below are the steps of the approach. The words *live* and *bank* are in italics because they are ones I have chosen and do not, to my knowledge, appear in other papers about nested sampling. I will use them throughout this chapter, hopeful that they add clarity.

- Randomly sample N points from the prior distribution of the parameters, and compute the likelihood at each one. These points are all initially called *live*.

- Determine which point among the N live points has the lowest likelihood, and *bank* that point along with its likelihood—*i.e.*, record it and remove it from the set of live points.

- Draw a new point to replace the banked one, this time drawing points according to their relative prior densities, but restricted only to that part of the parameter space where the likelihood is greater than that of the most recently banked point. The previous $N - 1$ points along with this new point are now the current live set.

- Repeat the previous two steps many times, always banking that point with the lowest likelihood among the live points and replacing it with a new one drawn from the prior, conditional upon the likelihood being greater than that of the most recently banked point.

- Monitor the set of banked points—especially the most recent ones—and at some point stop banking points one by one, and bank all the remaining live points. There are different decision rules you can choose for when it's time

to stop, but the general idea is that the remaining live points will eventually be "corralled" into a tiny pocket surrounding (ideally) the point of highest likelihood. In a sufficiently tiny neighborhood of this point the likelihoods will all be about the same, so sampling more points will lose its usefulness.

- Estimate the integral of the likelihood with respect to the prior by taking into account the fact that the earliest banked points were sparsely distributed in regions of low likelihood, and the latest points were densely distributed in a region of high likelihood. (The details of how to estimate the integral will be given later in this section.)

In the following sections we will:

- look at the mathematics behind the nested sampling algorithm;
- apply the algorithm to a very simple integration problem whose solution is known so as to demonstrate how it works in the ideal situation;
- apply the algorithm to the problem of integrating out the seven parameters in the 1-planet model for the star HD88133; and finally,
- consider numerous modifications to the basic nested sampling approach, all of them attempts to overcome the essential difficulty of applying nested sampling to a multimodal, multi-dimensional integrand.

Radial velocity data from the star HD88133 will be useful to consider because the dominant mode in the posterior distribution of its parameters is so much greater in magnitude than its secondary mode that it is obvious which mode is correct, and one can therefore estimate the marginal likelihood of the one-planet model in this case by using importance sampling. Figure B.1 shows the radial velocity data for this star shown in three separate plots with the models corresponding to the

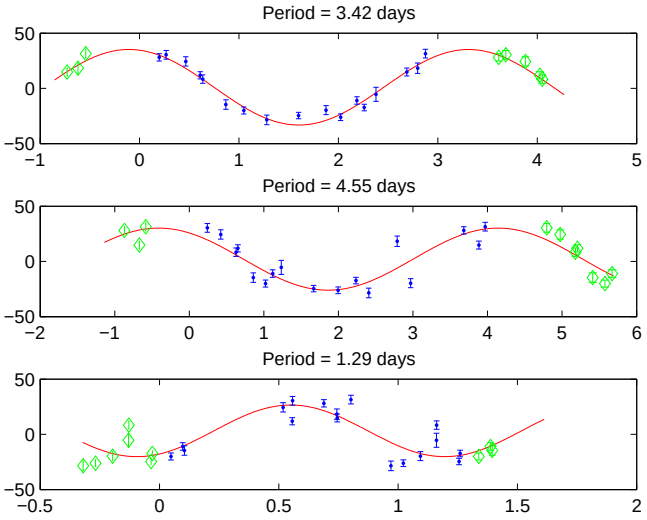

FIGURE B.1: HD88133. Shown are models corresponding to the primary, secondary, and tertiary modes in the likelihood function. Their local maximum log likelihoods are, respectively, -60, -97, and -121.

primary, secondary, and tertiary modes of the likelihood function. The three modes are sufficiently far removed from one another that each one's contribution to the total integral $\int_\Theta \pi(\theta)L(\theta)d\theta$ may be estimated separately using importance sampling; the importance functions do not "bleed" into one another. The log likelihoods of these three modes are, respectively, -60, -97, and -121, and their contributions to the desired integral are comparably far apart in magnitude.

The radial velocity data collected for the star HD88133 are therefore sufficient to make it obvious that there is one very dominant mode in the posterior distribution of the parameters—there is no doubt at all that a planet is present and its orbital parameters are estimable with great precision. As such, the actual normalizing constant that we seek—that is, the integral of the prior times the likelihood over the parameter space—may be estimated with great precision using importance sampling. Thus, we will have a "gold standard" by which to compare our efforts using nested sampling. We will see that, for reasons which will become clear, nested sampling

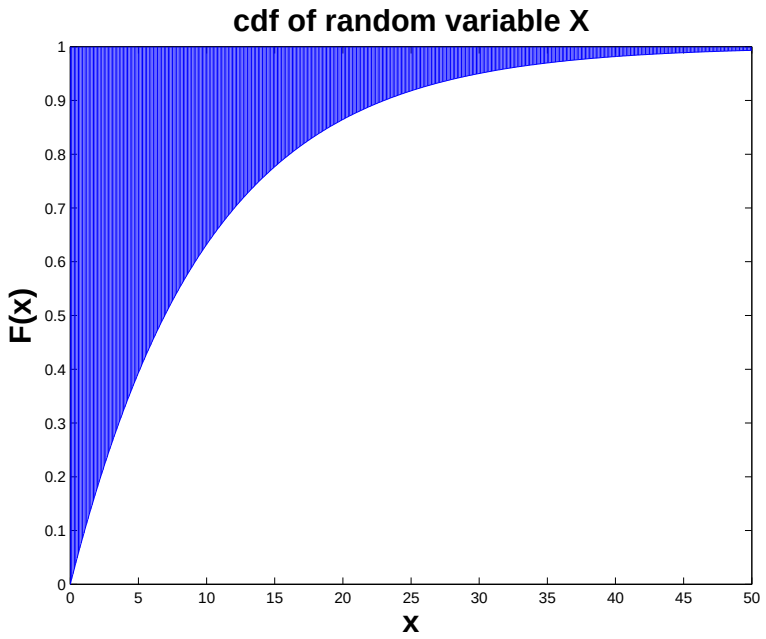

FIGURE B.2: The area of the shaded region equals $\mathsf{E}(X)$.

does not do at all a good job of estimating this integral.

Unfortunately, even after considering numerous modifications to the nested sampling algorithm, we will end the chapter ever more convinced that nested sampling is simply not a good fit for this problem.

B.2 The nested sampling algorithm: a summary

We wish to evaluate an integral which may be represented generally as $\int_\Theta \pi(\theta)L(\theta)d\theta$.

Lemma: For any positive-valued random variable, the area above the cdf and below 1 is equal to the expected value of the random variable. *i.e.*, if X has pdf f and cdf F, then

$$\int_0^\infty (1 - F(x))dx = \int_0^\infty xf(x)dx \equiv \mathsf{E}(X).$$

Figure B.2 illustrates this.

(For a proof of this lemma, see Casella & Berger, *Statistical Inference*, Second

Edition, p. 79.)

$$\begin{aligned}
\int_0^\infty (1 - F(x))dx &= \int_0^\infty (1 - \mathsf{P}(X < x)) \cdot dx \\
&= \int_0^\infty \mathsf{P}(X > x) \cdot dx \\
&= \int_0^\infty \int_x^\infty f(y) \cdot dy \cdot dx \\
&= \int_0^\infty \int_0^y f(y) \cdot dx \cdot dy \\
&= \int_0^\infty f(y) \int_0^y dx \cdot dy \\
&= \int_0^\infty f(y) \cdot y \cdot dy \\
&= \mathsf{E}(X)
\end{aligned}$$

Let Θ be a space and $\pi(\theta)$ be a probability density function defined on Θ. Let L be a non-negative real function defined on Θ. Then $L(\theta)$ has its own distribution. It should be easy to see that the cumulative distribution function of $\lambda = L(\theta)$ is defined by

$$F(\lambda) \equiv \int_{L(\theta)<\lambda} \pi(\theta)d\theta.$$

Recall that we want to estimate the integral $\int \pi(\theta)L(\theta)d\theta$. That is just the expected value of $L(\theta)$ under π. But we have seen from the earlier lemma that the expected value of a non-negative random variable with cdf F is equal to $\int_0^\infty (1 - F(x))dx$. So the integral we want to estimate must be equal to $\int_0^\infty (1 - F(\lambda))d\lambda$.

Note that if we define

$$X(\lambda) \equiv 1 - F(\lambda) = \int_{L(\theta)>\lambda} \pi(\theta)d\theta,$$

then the desired integral is equal to $\int_0^\infty X(\lambda)d\lambda$.

Intuitively, $X(\lambda)$ is equal to the probability that a θ randomly drawn from the parameter space Θ will have a likelihood greater than λ. It may be helpful to think of Θ as a plane and $L(\theta)$ as a hilly surface over that plane. The set of all θ's whose likelihoods exceed a threshold value λ is therefore rather like a collection of islands that rise above the "water" level λ. The probability (under π) associated with those islands is equal to $X(\lambda)$. As λ increases, the "water level" rises, some islands are submersed completely, and $X(\lambda)$ decreases.

Note that the area of the region in Figure B.2 may be integrated in the other direction and the area will remain the same. This leads to the following theorem:

Theorem: Assuming the function X is invertible, then

$$\int_\theta \pi(\theta)\pi(X|\theta)d\theta = \int_0^1 X^{-1}(p)dp.$$

. (See illustration in Figure B.3.)

$$\int_\theta \pi(\theta)\pi(X|\theta)d\theta = \int_0^1 X^{-1}(p)dp.$$

We now have a new integral that equals our original one. We want to estimate the integral $\int_0^1 X^{-1}(p)dp$, where $X^{-1}(p) = \lambda$ is that likelihood λ such that $\mathsf{P}(L(\theta) > \lambda) = p$. For example, if $X^{-1}(0.9) = 0.0042$, then 90% of draws θ from the prior will have likelihoods greater than 0.0042.

(*Aside*: note that even if Θ is a multi-dimensional parameter space, this new function X^{-1} is always one-dimensional and monotonic.)

Typically—especially in higher dimensional problems—the likelihood function will be much higher in one small region of the parameter space than everywhere else. If the prior is fairly diffuse, then the result is that X^{-1} is a very skewed function, with nearly all of its "area" clumped up near $p = 0$. (See Figure B.4.) We will be

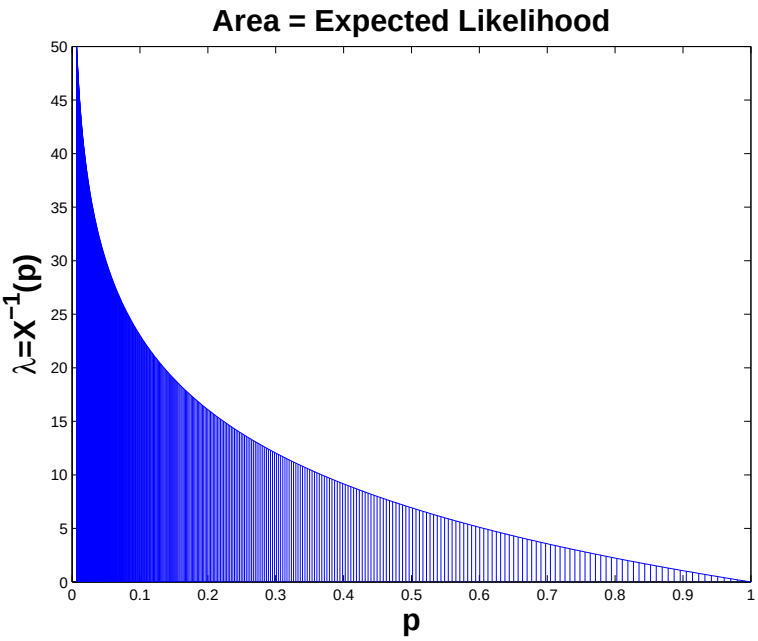

FIGURE B.3: The area of the shaded region equals the expected likelihood.

looking at graphs meant to approximate $X^{-1}(p)$ versus p, and because of this strong skew it will be useful to look at the graphs with p on a logarithmic scale. Also, the nature of nested sampling is such that sampled point will be far denser towards the left of this plot and sparser to the right—another reason to draw the plot with p on a logarithmic scale.

B.3 Another summary of nested sampling, with computational formulas

We have just seen that if X is defined as $X(\lambda) \equiv 1 - F(\lambda) = \int_{L(\theta)>\lambda} \pi(\theta) d\theta,$, then (assuming X is invertible—not a very strong assumption) then $\int_\theta \pi(\theta)\pi(X|\theta)d\theta = \int_0^1 X^{-1}(p)dp.$. We now fill in more details about how draws made from Θ using the nested sampling algorithm lead to a computational estimate of $\int_0^1 X^{-1}(p)dp.$ and therefore of $\int_\theta \pi(\theta)\pi(X|\theta)d\theta$, which is the integral we require for our model selection problem.

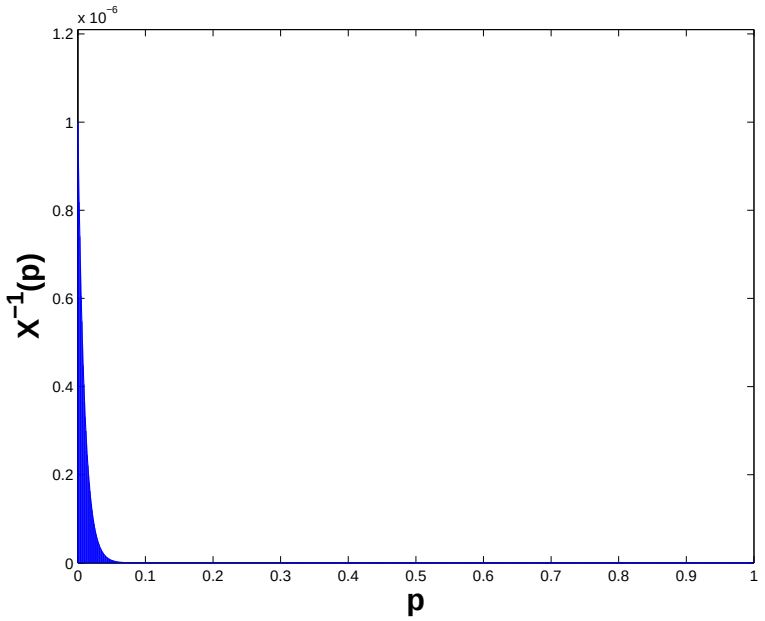

FIGURE B.4: All the "action" in this integral occurs near $p = 0$.

We recall that the nested sampling algorithm is as follows:

- Sample a set of N values of θ from the prior distribution π and order them according to their likelihoods. Call this set the *live* set.

- Find the value of θ among the live points that has the smallest likelihood and call it $\theta^{(1)}$. Put it in a sequenced set called the *bank*. It is paired with the expected N^{th} order statistic from among N points uniformly drawn from the interval $[0, 1]$; this expected order statistic is $\frac{N-1}{N}$. By assuming that this *banked* $\theta^{(1)}$ is representative of the lowest $\frac{N-1}{N}$ likelihoods, we take its contribution to the integral estimate is $[1 - \frac{N-1}{N}] \times L(\theta^{(1)})$.

- Replace the $\theta^{(1)}$ that was just banked with a new θ, this time drawn from among the prior distribution π, but restricted only to that part of the parameter space Θ where the likelihood exceeds that of $\theta^{(1)}$. This new θ along with the $N-1$ θ's that have not yet been banked constitute the current live set. Sort the current

live set according to their likelihoods.

- Again find the value of θ from among the live points that has the smallest likelihood, and this time call it $\theta^{(2)}$. Put it in the bank, along with the expected N^{th} order statistic from among N points uniformly drawn from, this time, the interval $[0, \frac{N-1}{N}]$; this expected order statistic is $\left(\frac{N-1}{N}\right)^2$. This time the contribution to the integral estimate is $\left[\frac{N-1}{N} - \left(\frac{N-1}{N}\right)^2\right] L(\theta^{(2)})$.

- Repeat this many times: banking the θ from the live set that has the lowest likelihood and replacing it with a new θ drawn from the prior π, but always restricted to that part of Θ where $\pi(\theta) > L(\theta^{(i)})$—i.e., that part where the likelihood exceeds that of the most recently banked θ. The contribution of $\theta^{(i)}$ to the integral estimate is equal to $\left[\left(\frac{N-1}{N}\right)^{(i-1)} - \left(\frac{N-1}{N}\right)^i\right] L(\theta^{(i)})$.

- Ideally, the contributions to the integral estimate will first increase, as the likelihoods they represent increase, and then eventually decrease, as the set of live θ's all congregate near the global mode (and hence have nearly equal likelihoods) and the difference $\left[\left(\frac{N-1}{N}\right)^{(i-1)} - \left(\frac{N-1}{N}\right)^i\right]$ diminishes. When these contributions have decreased sufficiently for it to appear that the remaining region of Θ has roughly constant likelihood, then the algorithm may stop, and the remaining live points may each contribute to the integral estimate a value equal to its likelihood times $\left(\frac{N-1}{N}\right)^i$. (One stopping rule, and the one we will use throughout this section, is to stop the sampling process when the greatest and the least likelihoods among the live points differ by no more than a certain relative amount–say, 0.1%.)

- The sum of all the contributions to the integral is the desired integral estimate.

One way to show the accumulating integral is in a pair of graphs, one of which

shows the most recently banked likelihood plotted against the expected order statistics (these will move *from right to left*) and the other of which shows the individual banked points' contributions to the integral, also plotted against the expected order statistics (these will also move from right to left). One would like to see the likelihood plot rise (moving right to left) and eventually plateau; and the contributions plot rise (moving right to left) and eventually fall. We will see such plots in the next section following a "toy problem". (They are shown in figure B.6.)

B.4 A toy problem

Consider the "toy problem" where there are two unknown parameters α and β with a joint prior uniform over the unit square, and with likelihood $L(\alpha, \beta) = \alpha^{30}(1-\alpha)^{30}\beta^{30}(1-\beta)^{30}$. This may arise if a treatment and a control group each see 30 successes and 30 failures among 60 trials for each group, and α and β are the respective probabilities of success per trial in the groups. We know that the integral of this likelihood with respect to the prior is equal to $\left(\frac{\Gamma(31)^2}{\Gamma 62}\right)^2 \approx 1.9215 \times 10^{-38}$.

Figure B.5 shows one implementation of the nested algorithm with only $N = 5$ points, connected according to the order in which they were banked. Figure B.6 shows the two graphs described earlier, with the likelihood plot on top and the integral contributions below. This implementation stopped after just 190 steps, and its integral estimate was 3.34×10^{-38}, which is about 1.74 times the correct value.

Such a large error may be avoided by using a much larger N. That is because with a larger population of live points, the movement from regions of low likelihood to regions of high likelihood is more gradual, with tinier steps. However, the tinier steps are, the more must be taken; and unfortunately, as we shall see later, computation time for the HD88133 problem permits N on the order of 10,000, but not on the order of 1,000,000.

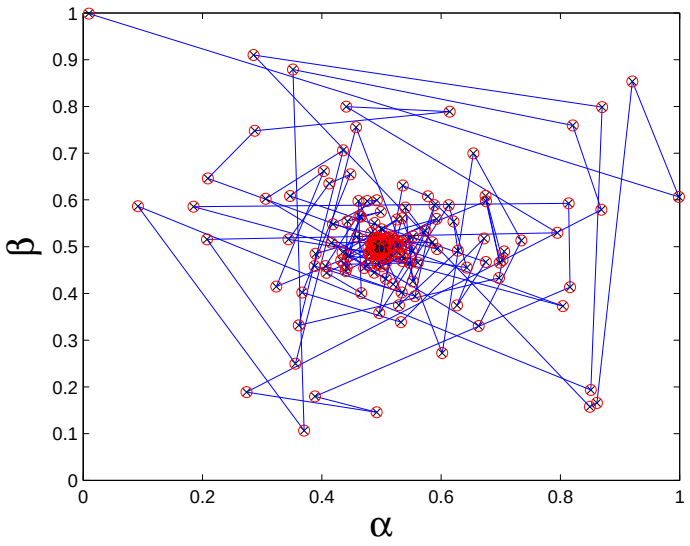

FIGURE B.5: With only $N = 5$ points, one can see the migration of points towards the mode at (0.5,0.5)—but the integral estimate will necessarily be a very crude one.

This section of this paper, and this toy problem, are meant to convey a sense of how nested sampling works in an ideal situation. For the most part, we see from the resulting plots that the live points start out scattered all over the parameter space, and then move quickly to concentrate in a small neighborhood of the MLE of (α, β). Such rapid movement away from regions of low likelihood results in more efficient sampling from the parameter space than could be achieved by, say, direct sampling from the prior. Likewise, the plot of the contributions to the area rises (moving right to left) as the live points begin to vacate the vast region of low likelihood, then it falls approximately exponentially, once the likelihoods remain approximately constant and their contributions diminish according to $\left[\left(\frac{N-1}{N}\right)^{(i-1)} - \left(\frac{N-1}{N}\right)^{i}\right]$.

If we repeat this algorithm many times for the toy problem, we may see whether the method tends to produce an unbiased estimate, and what the spread of the estimates is. Figure B.7 shows histograms of 1000 estimates for this toy problem under four different live point population sizes N.

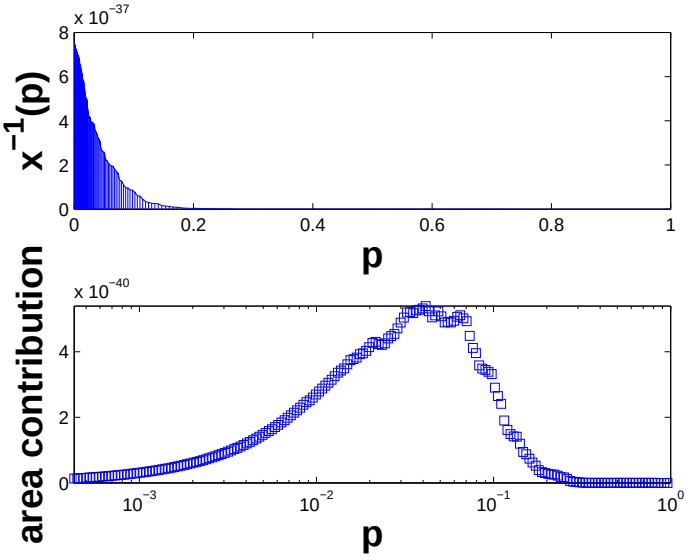

FIGURE B.6: In the upper plot, the scale on the horizontal axis is kept linear so that the shaded area corresponds to the desired integral. In the lower plot the scale on the horizontal axis is logarithmic so that the rise and fall of the contributions can be seen clearly. The "bumpiness" of this lower plot where the integral contributions peak is a result of such a tiny N (5 points). An N on the order of a thousand or even a million points would be more desirable. While $N = 1,000,000$ might be possible with this toy problem, we found it impossible with the real problem of the HD88133 data (see the section on computing time).

We see from the histograms that the estimates appear to be unbiased and also that the variability decreases with larger N. The decrease unfortunately does not appear to be very rapid, and a population of $N = 100$ points still regularly produces estimates that are off by 25% to 35% of the actual value. This would not necessarily be a problem in model selection if the marginal likelihood estimates of two models were different by a very large factor. And that's a good thing, because the computation times associated with actual star data are orders of magnitude greater than those required for this toy problem (more on that shortly). But in any case, an error by a factor of 35% will be the least of our problems. A far greater problem, and one that nested sampling does not seem able to overcome, is that of sampling from the

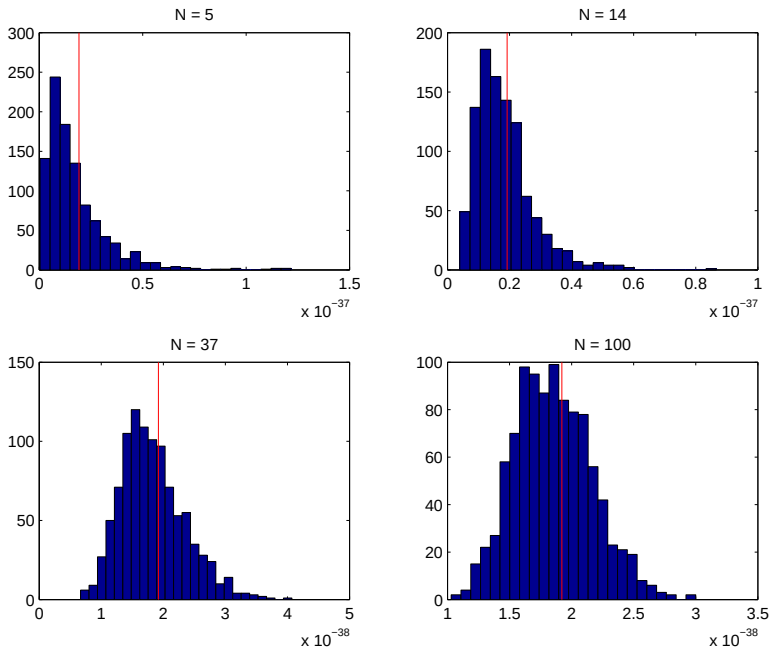

FIGURE B.7: The nested algorithm appears to be unbiased, and it produces estimates with less variability with larger values of N. In each plot a vertical line is drawn at the actual integral value of 1.9215×10^{-38}.

restricted prior in a multimodal space.

Before continuing to sections that look at a real star problem, here is an important remark about the sampling step.

The most difficult step in nested sampling is the sampling of a point according to the prior restricted to a region of likelihood higher than that of the most recently banked point. In the toy problem above, this was done according to a method suggested by Mukerjee, Parkinson & Little. [7] At each iterative step, an imaginary rectangular "box" was drawn around all the live points, then it was expanded around its center by a constant factor of, in this case, 25%. A point was drawn uniformly from the box, and then to make the sampling follow the prior distribution within the box, a rejection step was performed in which the prior density of the sampled point was compared to the maximal value of the prior within the box.

The "box" naturally shrinks as the points congregate around MLE, making the sampling more efficient than continual sampling from the prior or from the full parameter space, and the rejection step forces the sampling to follow the relative values of the prior. In fact, as the box shrinks, the ratio of the highest to the lowest values of the prior within the box approaches 1, making this step quite efficient as well.

This sampling algorithm has the advantage of not requiring previous knowledge of where the MLE lies: as the points naturally abandon regions of low likelihood, the box will shrink like shrink-wrap around the MLE. However, the algorithm only works under one critical assumption: that the likelihood function is unimodal; or at least that if it is multimodal, then there is a dominant mode occupying so much of the parameter space that points sampled from the shrinking "box" will with relatively high probability land near that mode.

But if the likelihood function is multimodal, then the box will have a very hard time shrinking. The live points will all be scattered around in pockets of different modes, and between those modes will be vast regions of likelihood that are smaller than the threshold. If we return to the analogy of the hilly surface $L(\theta)$ and the "water level" of λ, then the nested sampling algorithm requires us at each step to sample from only those islands that rise above the current water level;and the oceans between them may be quite vast—especially if Θ is a multi-dimensional space.

B.5 HD88133: importance sampling estimate as a "gold standard"

17 radial velocity measurements have been made on the star HD88133. [8] They are shown below in two plots; the upper plot is the raw data and the lower plot is of the data with time shown modulo 3.4153 days, which is the MLE of the period given the data.

Aside: that MLE was found in the following way. First, the Lomb-Scargle periodogram was used to identify candidate periods for the HD88133 data. The largest,

second-largest, and third-largest peaks were identified. Next, the local mode corresponding to that period was found by using the Nelder-Mead search algorithm, with the period fixed at the one indicated by the L-S periodogram; so the search was done only in the other six parameters. Then the Nelder-Mead algorithm was performed again, beginning where the previous Nelder-Mead search ended, only this time the period parameter was freed as well, so that a local mode in seven dimensions was truly found. Finally, the likelihoods at the primary mode, secondary mode, and tertiary mode were compared to see whether the ordering suggested by the periodogram was in fact the true order of the likelihoods. (It was.)

This is in fact the method that was used throughout this paper to identify local modes. When one mode is exceptionally dominant, as in this case, then the problem may be well-understood without worrying about the behavior of the likelihood function in other regions of the parameter space. However, when the periodogram leads to multiple modes that are comparable in magnitude (which is necessarily the case with sufficiently small data sets, or when a planet is not present at all) then we must keep on looking at shorter and shorter spikes in the periodogram, which lead to smaller and smaller modes in the likelihood function. This will be discussed further in Chapter 5 of this paper.

For the purpose of model selection, we will need two integrals:

$$\pi(X|M_0) = \int_{\Theta_0} \pi_0(\theta) \cdot \pi(\theta, X) d\theta, \tag{B.1}$$

and

$$\pi(X|M_1) = \int_{\Theta_1} \pi_1(\theta) \cdot \pi(\theta, X) d\theta, \tag{B.2}$$

The first of these integrals is not hard to find, since the parameter space Θ_0 has

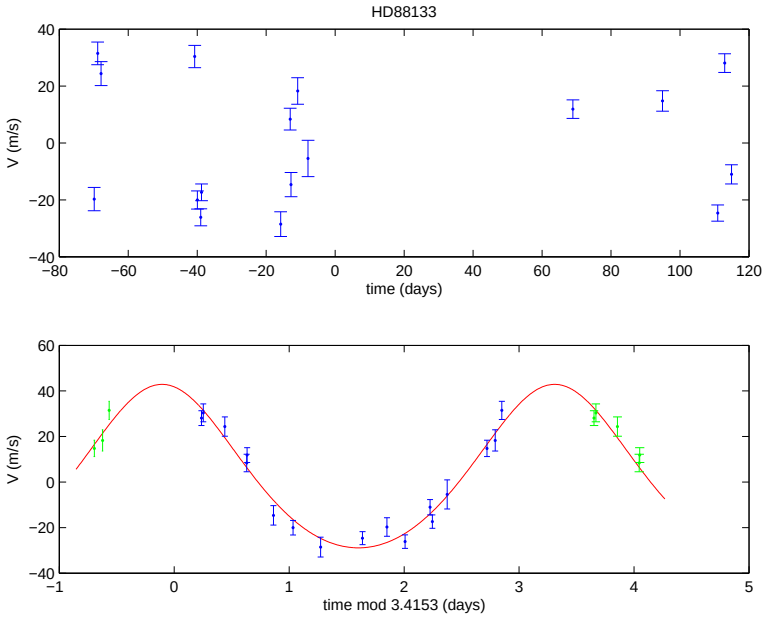

FIGURE B.8: Green points in the lower graph (those at $t < 0$ or $t > 3$) are duplicates of points in blue.

only two dimensions: C and s. Using importance sampling with 100,000 draws and an effective sample size of 62,000, the first integral is estimated to be $(1.683 \pm .004) \times 10^{-37}$. This agrees with an integral estimated by gridding finely on C and s axes (200 divisions along both axes, spanning six standard deviations on either side of their means) and estimating using a Riemann sum. That Riemann sum estimate is 1.681×10^{-37}.

The latter integral is more challenging. It cannot be easily estimated with a Riemann sum, as there are five additional dimensions. The integrand is multimodal, but if we can trust that nearly all the non-trivial mass lies in a region surrounding the one dominant mode (whose MLE is shown in the lower plot in figure ??)—and we can so trust in the case of HD88133, as we have seen—then we may again use importance sampling, drawing from a $t_{(4)}$ distribution centered on the MLE and having a covariance matrix estimated from the sample covariance of many draws

from a previously-constructed Metropolis chain. Our result of such an importance sampling estimate with 100,000 draws and an effective sample size of 14,200 is $(2.96 \pm 0.03) \times 10^{-35}$.

The Bayes factor is therefore 176 in favor of the one-planet model over the zero-planet model; if we place a prior of 0.12 on the one-planet model and a prior of 0.88 on the zero-planet model (which approximately matches present planet detection rates), this gives a posterior probability of 0.96 that HD88133 does host an orbiting planet.

B.6 Computing time

The average CPU time it takes on a 1.66 GHz (512MB) processor laptop for our code (in Matlab) to draw a random θ from its prior distribution and compute the values of $\pi(\theta)$ and $\pi(X|\theta)$ is about 0.95 milliseconds for HD88133, which has 17 observations. For 51 Peg, which has 111 observations, the average time is 4.15 milliseconds. After computing this time for eight stars, it is clear that the relationship between average computation time and the number of points is very nearly linear, with the slope being about 0.008 milliseconds per additional observation, as shown in Figure B.9.

Contrast this with generating standard normals at random and computing their pdfs; on the same laptop this takes on the order of 0.18 *microseconds*—over 5000 times faster than the HD88133 problem. Doing the same with a 7-dimensional multivariate standard normal takes about 1.8 microseconds, and doing the same for a 7-dimensional standard $t_{(4)}$ distribution takes about 64 microseconds.

I include this section on computing time because nested sampling requires a lot of points—Skilling used millions of points in his examples, which had simpler functions than the present star problem. But increasing the number of points used slows down the pace at which one approaches the region(s) of high probability density, and so increases the number of iterations required for a desired precision. What's more,

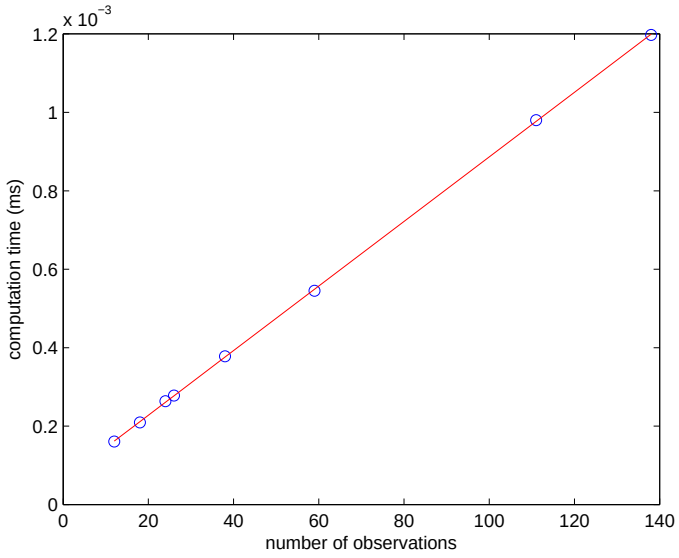

FIGURE B.9: Each additional observation adds about 0.008 milliseconds to the time required to compute the predicted radial velocities.

in his examples Skilling used a fairly direct method for drawing points according a restricted prior. In the earlier "toy problem" section, we used a fairly direct method for so drawing points. With a 7-dimensional multi-modal likelihood function, that is not possible. The "ocean" between the "islands" eventually occupies so much space that the rejection step in the algorithm rejects draws with a frequency that essentially grinds the algorithm to a halt. (This will be discussed further in the following section, in which we show that for the HD8133 data, the rejection algorithm must fall to an acceptance rate of at most 1 in 2 million before it settles into a pocket surrounding the global mode. Since the period parameter is sensitive to the number of data points, then with larger data sets, the likelihood function would be spikier and the rejection rate would be even larger.)

We therefore attempted to use variations on the Metropolis algorithm to perform the step of drawing from the restricted prior. These will be discussed in more detail as this section progresses. It is mentioned here in the context of computing time because

using a Metropolis algorithm to draw a *single point* at each of *many* iterations makes the computing time even greater still. The computing time is great because the likelihood computation is heavy; and also because each new point θ drawn from Θ according to the restricted prior required hundreds or thousands of meandering Metropolis draws along the way.

Using the Metropolis algorithm to find each new point, when there are millions of points and tens of millions of necessary iterations is prohibitively slow with our code. For this reason, throughout much of this report, the number of live points (N) is on the order of one thousand instead of one million. This naturally would decrease the accuracy and precision in any resulting estimate. But one would surely expect the answer to be accurate within one or two orders of magnitude, and this did not occur. A bigger problem seemed to be the fact that the support of the global mode was just too small to be stumbled upon with any reasonable probability (see next section).

B.7 The size of the support of the global mode

If any sampling technique is going to estimate an integral with reasonable accuracy, then it is reasonable to think that at least some points must be sampled from somewhere "near" the global mode. Unless it is told where that mode lies, then it must find the mode itself. That is not hard if the function being integrated is unimodal. (For example, in the earlier toy problem, we used a "box" that encompassed all the live points; as these points coalesced around the mode of the likelihood function, the box shrank, so that newly drawn points were guaranteed to be fairly near the mode.) But if the function is multimodal (and the likelihood function is highly multimodal in the case of HD88133) then it is more difficult.

Figure B.10 shows the Lomb-Scargle periodogram for the HD88133 data. (We will become very intimate with the L-S periodogram in Chapter 3 of this dissertation!)

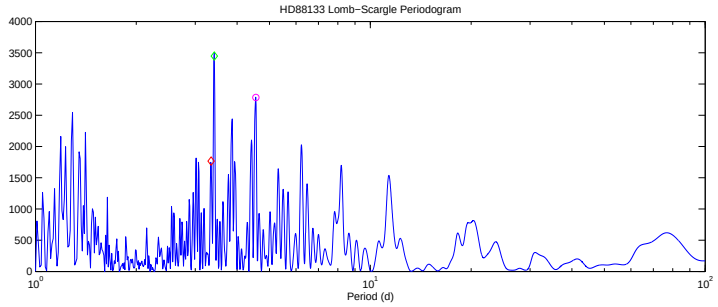

FIGURE B.10: Lomb-Scargle periodogram for the HD88133 data. The two diamonds are at the periods $P = 3.347$ and $P = 3.415$, and the circle is at period $P = 4.552$.

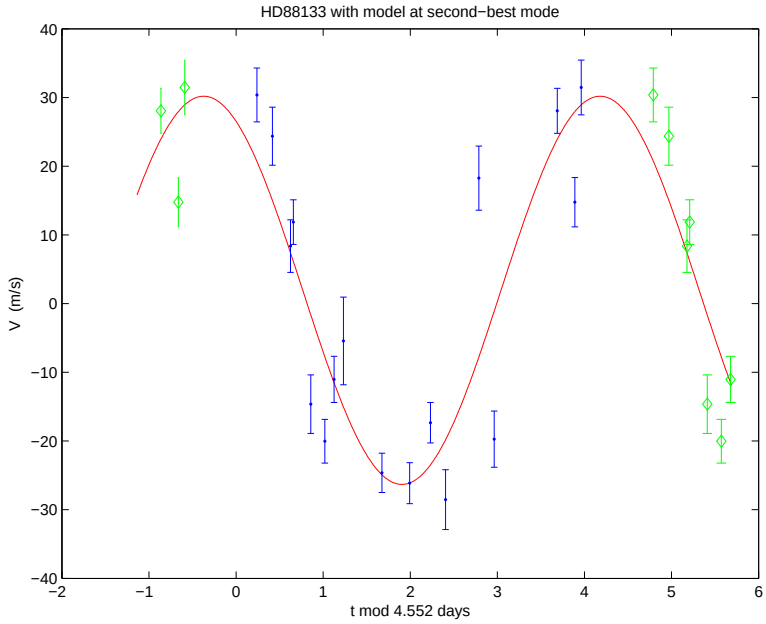

FIGURE B.11: The second best model.

The highest peak in this periodogram occurs at approximately $P = 3.415$ days, the period which was shown earlier for these data. The second-highest peak occurs at approximately $P = 4.552$ days. The data modulo 4.552 days are shown in figure B.11 along with the model whose parameters are fit to their MLE values near that period.

Although the $P = 4.552$ solution may appear to be a fairly good one, in fact the

log likelihood at this local mode is equal to -96, compared with -49 for the dominant mode, a difference of 47 favoring the shorter period.

So now we consider the question: for a nested sampling algorithm that is guided only by the locations and function values of the points it has already found, how hard is it to eventually find the global mode of the function, where the bulk of the probability mass lies? We do not address it in general here, but consider the case of the HD88133 data. The second-highest mode in the posterior occurs at a point where the log likelihood is about -96, compared with -49 at the global mode.

Because these log likelihoods are so dramatically different (the actual likelihoods differ by a factor of 2×10^{20}), it is hard to imagine, though it is possible, that the probability mass is non-negligible anywhere other than the global mode. It is true that the likelihood function is sensitive to small changes in some of the parameters (especially the period), which results in a smaller support for the mode in the posterior probabiltiy density function. But that sensitivity is nearly equally true everywhere, so the support at the secondary mode, while very plausibly larger than that at the global mode, is not likely to be sufficiently greater to give non-trivial mass to that mode.

Let us give our nested sampling algorithm the benefit of having already found, by any means, the second-highest mode. There are two possible ways the algorithm might now continue. It might "stumble upon" a point near the global mode whose posterior probability density is higher than that at the second-highest mode—higher than any posterior density it has yet found. In that case, all of the points near the secondary mode would quickly migrate to the vicinity of the global mode, and when the algorithm concluded, it would have a good chance of being successful—of producing an integral estimate that is approximately correct. But the second way the algorithm might continue is to settle on points surrounding the secondary mode, concentrating them ever tighter and tighter until it terminates, never having gotten

near the global mode. Were this to occur, the resulting integral estimate would be a gross underestimate of the actual integral.

If the nested sampling algorithm were able to draw exactly from the restricted prior, then as the points concentrated more densely around the secondary mode, the volume of their support in the prior would diminish until eventually the restricted prior would shift to the global mode with very high probability. But that is supposing that the algorithm could draw exactly from the restricted prior, as if it "knew" that apart from the mode in which its points were concentrated, there existed another mode elsewhere.[2] Absent this knowledge, what are the chances that the algorithm would "stumble" upon a point near the global mode whose posterior density is higher than at the secondary mode?

By trial and error, we found a rectangular region (in seven dimensions) surrounding the global mode that appears to fully contain all the points whose likelihood exceeds that at the secondary mode. Figure B.12 shows pairwise scatterplots of such points drawn from the prior, restricted to this region. Some of the scatterplots suggest there are actually three modes in this vicinity, but one such extra mode arises because the parameter $z \equiv (\omega + M_0) \mod 2\pi$ "wraps around" at 2π. But it is clear that there are still in fact two modes here. These two modes correspond to the two periods indicated by diamonds in the earlier periodogram (figure ??). So in fact, the secondary mode in the posterior density function does not correspond to the secondary mode in the periodogram after all—$P = 4.552$ is at best a tertiary mode.

We are trying here to determine the chance of finding the global mode by chance, and the presence of this additional mode results in an overestimate, not an underestimate of this chance. We give the nested sampling algorithm the benefit of having already found the mode at $P = 4.552$ and we ask: what is the chance that it will later

[2] The global and secondary modes of the posterior density are not at all near one another. There is a vast chasm between them that is not bridged by other helpful modes.

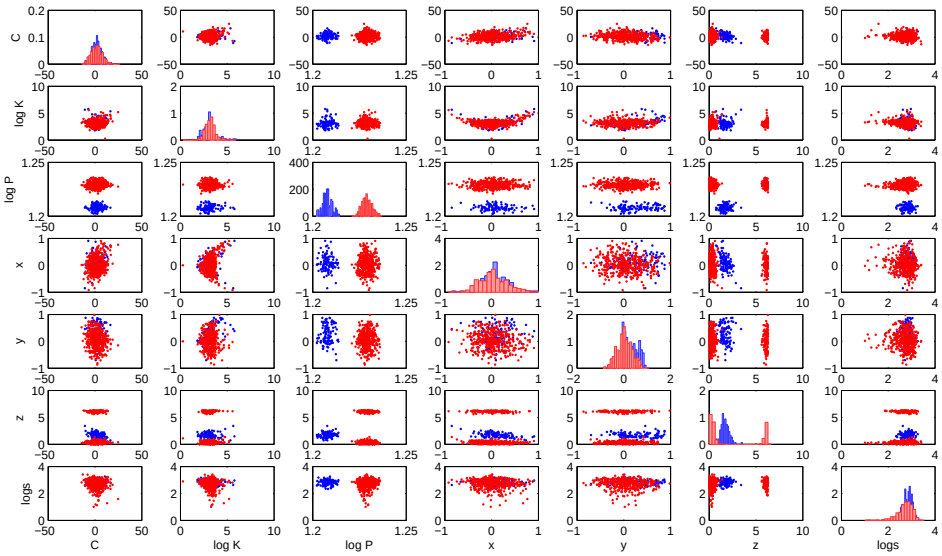

FIGURE B.12: Parameter values where the posterior density exceeds that of the mode at $P = 4.552$. The two colors indicate two different modes in the same vicinity.

find *either* of the modes at $P = 3.347$ or $P = 3.415$? (We may consider the former of these two modes as a possible stepping-stone towards the other.) We used importance sampling to estimate the total prior probability associated with these points whose posterior density exceeds that of the mode at $P = 4.552$. We drew uniformly from the rectangular hyper-box indicated by the boundaries of the scatterplots in figure B.12. After one million such draws, a 99% confidence interval estimate of the prior probability over the region (based on 510 such points found in a million) is [4.6e-7, 5.5e-7]—i.e., at best,about 1 in 2 million.

Now 1 in 2 million is far from impossible. On average, it would take the Matlab code used here only about a half an hour to compute this many priors and likelihoods. So why not simply use a rejection algorithm to draw from the restricted prior at each step of the nested sampling algorithm? There are several reasons why this is problematic:

- First, we have here the benefit of the periodogram and some analysis that lets us know about how many draws we should expect to make from the prior before we land near the global mode. We would not know that in advance for another set of radial velocity data.

- Second, radial velocity data for a star that does host a companion have the property that as the number of data points increases, the mass under the true mode increases, while its support under the prior decreases, due to increased sensitivity to the period parameter. Thus, with more data a periodogram or some other analytic technique makes the true period stand out more, while it gets harder to find using a random search technique, surrounded as it is by a "cocoon" of dramatically lower posterior density.

- But the main reason that it is impractical to draw millions of times, hoping to find the global mode, is that such a draw must be made at every step of the nested sampling algorithm. Spending half an hour once to draw points hoping to find the global mode may be practically possible. But at every step, one must draw from the prior restricted to points of higher posterior density than the lowest one yet found. Supposing that a million draws from the prior were made initially and one of them happened to be near the global mode, it would take at least a million draws from a restricted prior (to replace the other 999,999 points)—and in all probability, at least many thousands of millions—before the entire collection of points was concentrated around the global mode. This would take an impractically long time.

The upshot of all this is that in a multimodal problem, nested sampling may only be practical if (1) all modes carrying substantial posterior probability mass also carry sufficient support in the prior that they may be found fairly easily by drawing directly from the prior—essentially, by rejection sampling, or (2) all modes carrying

substantial posterior probability mass are sufficiently close to other massive modes that as the critical threshold on the posterior times the prior increases, they may be stumbled upon by "mode-hopping".

If we think of the modes as islands in a sea whose water level is rising, then the islands with the most mass must either be large enough to be seen "from the air" as it were (rejection sampling) or else must be close enough to one another that as the water level rises and inundates some islands, the points will always be able to "island-hop" to another, continuing until they have found the highest island of all.

Neither of these seems to be the case with HD88133. The global mode has a very small support in the prior, mainly because of the extreme sensitivity of the likelihood function to small changes in the period parameter. But that mode is also isolated from other modes (excepting, as we have seen, its very nearest neighbor—a close cousin) by vast chasms of low posterior density, so that it is extremely unlikely that a point would happen to find its way by chance from another mode to the global mode.

This is why nested sampling does not seem to be a good solution to the problem of integrating the posterior density function for HD88133, and probably other similar radial velocity data sets as well. Nevertheless, we made several attempts to overcome the inherent difficulties of mode-isolation, listed in the overview at the start of this paper, and the rest of this chapter describes the results of these attempts—unfortunately, all of them unsuccessful.

There is one possible way to overcome the isolated-mode problem that we feel might yet be successful, but which we did not attempt. That is to vary dramatically the number of points in the "live" collection. If one suspects a priori that the posterior density function does in fact have one or a very few isolated modes with tiny support but large posterior mass, then one could begin by making several millions of draws, and thus hope to have at least one land near each mode of substantial mass.

Then one could dramatically decrease the number of "live" points, culling away many points of inferior posterior density without replacing them, until the number is more manageable, perhaps in the thousands. Then one could follow a one-to-one replacement scheme for a while until (one hopes) the points were all concentrated in one or more mode whose mass was non-negligible. And then perhaps at the end one could increase the number of "live" points again as the algorithm played out, so as to increase the accuracy of the integral estimate.

But now we continue to sections describing the attempts we made to integrate the prior times the likelihood function for HD88133 using nested sampling with a constant number of points—at most 50,000, considering how long it took some of our methods to make a draw from the restricted prior, and how slowly the collection of points converged on a mode.

B.8 Sampling from the restricted prior using rejection sampling

As we saw earlier, the modes of the prior times the likelihood function are so tall and narrow that rejection sampling is impossibly inefficient. We tried it anyway.

If $N = 100$ points are used and 2000 steps of nested sampling are performed, you would expect the "banked" points to achieve about the same likelihood level as you would achieve with 1000 points and 20,000 steps of nested sampling—that is, for a given likelihood level, the ratio of nested sampling steps to live population size N is roughly constant. We called this the *depth ratio*. As we have discussed already, a fairly large N is expected for a reasonably accurate integral estimate, but this requires not only more points sampled initially, but more steps as well.

Solely for the sake of seeing whether rejection sampling would be feasible, we ran nested sampling with 100 points until the number of rejections passed a certain "quitting rule" threshold. We did this 500 times for each of four different quitting rule thresholds (10, 100, 1000, and 10,000 rejections). The plot in figure B.13 shows

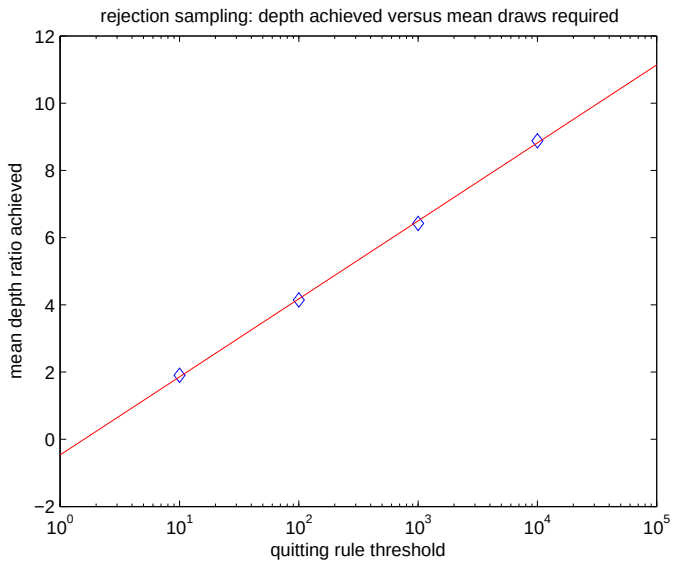

FIGURE B.13: The number of draws required for success in rejection sampling slows down exponentially as the "depth" increases.

the relationship between the quitting rule threshold and the depth ratio achieved. It appears that the relationship is logarithmic, with each factor of 10 increase in the quitting rule threshold resulting in an increase of a little more than 2 in the depth ratio reached. Put another way, as the algorithm proceeds the number of rejections required for a successful draw increases exponentially with "depth". In order for rejection sampling to be feasible, one would need to reach a satisfactory depth ratio before the quitting rule threshold became too large to be practical.

It is not clear what depth ratio is required to achieve a reasonable integral for the HD88133 problem, but eight is not enough. In 500 iterations with a rejection threshold of 10,000 rejections, the greatest log likelihood found was -90.1, the corresponding solution of which is shown in figure B.14. Clearly, this is not a good solution. At the global posterior mode, the log likelihood is about -49.

With the sampling process slowing down exponentially, it does not seem reasonable that rejection sampling will be sufficiently efficient at drawing points from the

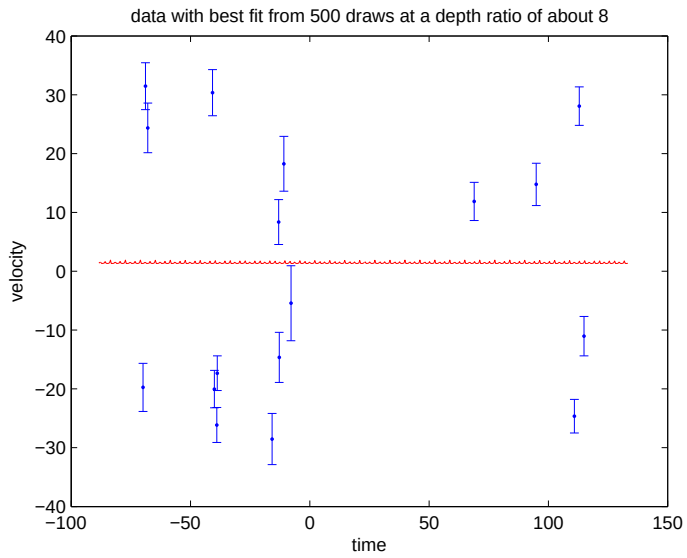

FIGURE B.14: At a "depth" of 8 (about 800 draws starting with 100 points), this is the best solution that was found. Clearly, a greater depth is required.

restricted prior to produce a reasonable integral estimate in a reasonable amount of time. We tried letting one simulation run over several days with only 100 points, and after three days, it had only performed 1305 simulations, achieving a depth ratio of only about 13, and among its 100 resulting points (we aborted simulation) the best solution was still terrible.

All the rest of our attempts to sample from the restricted prior are variations on the Metropolis algorithm.

B.9 Sampling from the restricted prior using the Metropolis algorithm with a constant proposed jump size

For a highly multimodal function, rejection sampling may be about the only way to guarantee exact draws from the prior restricted to function values exceeding a given threshold. But as it is impractically slow, we next tried using the Metropolis algorithm to draw from the restricted prior. Initially, this will be slower than rejection

sampling, since the latter may require very few draws early in the simulation while the former will require a good deal of meandering to achieve independence between the starting point and the final point. However, as the algorithm progresses, we have seen that rejection sampling will slow down exponentially, while draws by Metropolis will still take about the same amount of time. Thus, the Metropolis algorithm will guarantee a result in a finite, predictable amount of time. However, it is not certain that the resulting draws at each step are good approximations of draws from the restricted prior.

B.9.1 Metropolis starting on a random "interior" point

Nested sampling requires at each iteration dropping out the point with the lowest likelihood and replacing it with a point drawn from the restricted prior which we will henceforth call π^*, or $\pi^{*\lambda}$ if it is necessary to specify the threshold λ. (Recall that the restricted prior has a density equal to zero at every point for which $L(\theta)$ is less than that of the most recently banked point, and is everywhere else proportional to the prior π.) Sampling from π^* is not hard when the threshold likelihood is quite low, for then π^* and π are very similar. The problem arises when the threshold likelihood becomes high enough to create lots of isolated islands in the parameter space Θ. As we saw in the last section, these islands will eventually become so small relative to the oceanic part of Θ over which the likelihood is inferior to the threshold that direct rejection sampling is prohibitively inefficient.

Our first Metropolis approach to sampling from π^* was, at each step of the nested sampling algorithm, to construct a Metropolis chain of values of θ, using π^* as the desired density function, and beginning with a θ selected at random from among the "live" points. We cycled through the seven parameters, proposing for each one a jump in that parameter's dimension only. Jump proposals were symmetric about the present location, normally distributed, and with a variance determined in advance

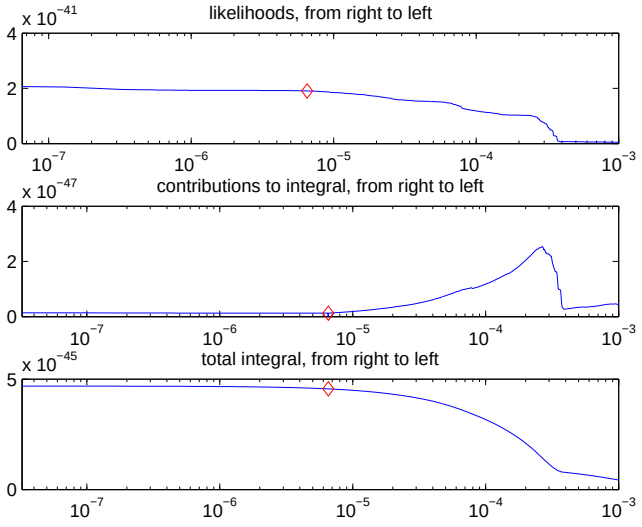

FIGURE B.15: The top graph shows that the highest likelihood discovered was around 2×10^{-41}. The likelihood at the MLE is about 4.6×10^{-22}. The search didn't even come close to finding the MLE. The red diamond in each plot shows where the banked points end (banked points are to the right) and the "leftover" points begin (these are to the left). The middle graph shows the progression of contributions to the integral estimate. Moving right to left, the sudden rise early in the iterations occurs as the algorithm suddenly "discovers" a pocket of higher likelihood than where it has been gathering. The points all quickly migrate to that new pocket. The bottom graph shows the progression of total cumulative integral estimate. Its final value here is 4.7×10^{-45}. The true value is about 3×10^{-35}.

to be such that about 25% to 35% of the jumps would be accepted initially. We used $N = 1000$ points.

The three plots in Figure B.15 are typical of the results we obtained performing the algorithm in this way multiple times. The uppermost of the three plots shows the succession of banked likelihoods. The highest they ever get is about 2×10^{-41}, which is many orders of magnitude smaller than the likelihood at the MLE, which is around 4×10^{-22}. This indicates that the search did not find the MLE. In fact, Figure B.16 shows the model corresponding to the best fit that the algorithm did find.

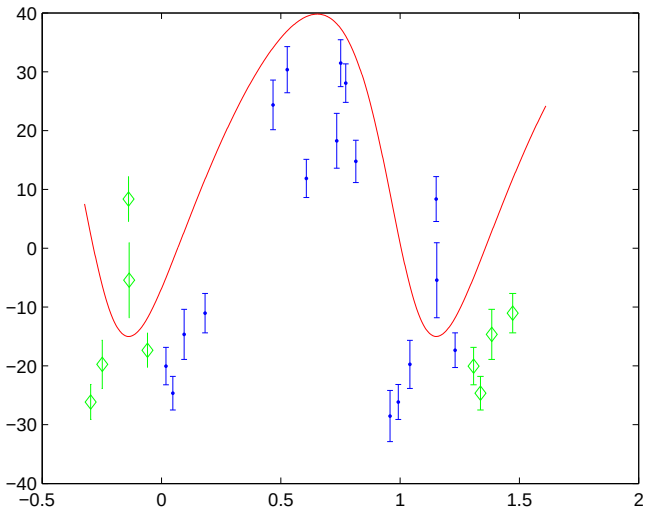

FIGURE B.16: The best solution found by the nested search algorithm is still many orders of magnitude worse than the MLE.

The poor integral estimate of, ultimately, 4.7×10^{-45} is too small by a factor of about 6 billion, due to the fact that the algorithm never found the MLE. Running the algorithm many times produces a very wide range of integral estimates, but all of them gross underestimates of the true value, because none of them ever found the MLE. This approach to integral estimation is horribly biased; the explanation is that the Metropolis step does not successfully draw from the restricted prior π^*. Initially it probably does a reasonably good approximation of it, but as the water level rises and the islands become more isolated, the live points all congregate on some single island, oblivious to the presence of any others. Metropolis moves are accepted with great frequency initially, but eventually they are nearly all rejected, as the live points huddle together like castaways during a terrible storm, unable to escape the island on which they've been marooned, despite countless valiant attempts.

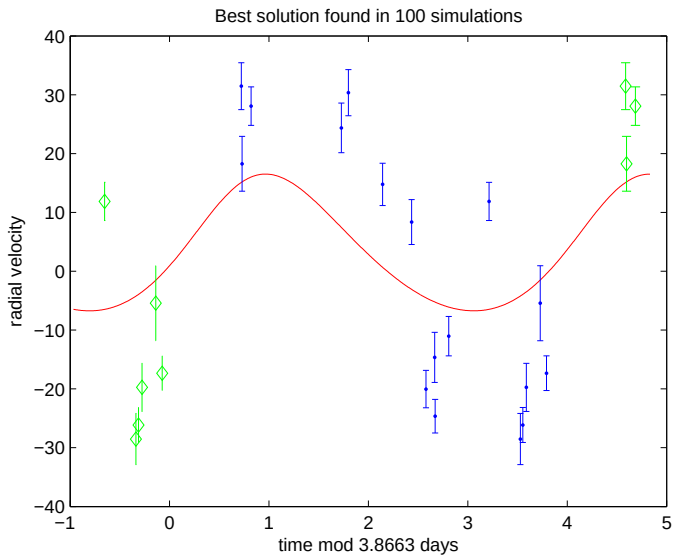

FIGURE B.17: The best solution found by the nested search algorithm after 100 tries is still poor, and the integral estimate is too small by a factor of about two million.

B.9.2 Metropolis starting on the most recently rejected "boundary" point

Since the earlier Metropolis algorithm was repeatedly unsuccessful, we next tried the same thing, only with the starting point for the Metropolis algorithm at each step being the most recently banked point, rather than a randomly selected interior live point. The results were identical. After 100 simulations using $N = 1000$ points, the best integral estimate was 1.7×10^{-41}, and its solution is shown in Figure B.17.

B.10 Metropolis searches with jump proposals shrinking

When using the Metropolis algorithm at each stage of nested sampling to sample the next point from the restricted prior, the space we're sampling from must necessarily be smaller at each step than it was at the step before. In the case of a unimodal likelihood, that shrinking space is very well-behaved, and the Metropolis draw, wherever it begins, needs sufficient steps to "explore" the space before settling on its final

location, hopefully one that is approximately independent from where it began. In the case of a multimodal likelihood, the shrinking space is less well-behaved, being like a collection of distinct shrinking (and vanishing) islands in a rising sea, but one still hopes that the Metropolis algorithm can produce a "draw" from the restricted prior that is approximately independent from where it began.

Because the space being sampled from is shrinking, it is natural to think that a good proposed Metropolis jump size for the *early* in the nested sampling algorithm may not be so good later in the nested sampling algorithm. In particular, one supposes that early jump proposals should be large in magnitude so as to explore the larger space more thoroughly, while later jump proposals should be smaller in magnitude so as to avoid an overabundance of rejections. In this section we briefly look at two attempts to use the Metropolis algorithm to draw from the restricted prior, with jump sizes shrinking according to a predetermined schedule.

B.10.1 Jump proposals shrinking at a constant rate

We first considered jump proposals shrinking in magnitude at a constant rate. We began by considering magnitudes that were good sizes for the full unrestricted space, and then shrank them at a constant rate in each dimension as the algorithm progressed. We were able to monitor for each step in the nested sampling algorithm the rejection rate for the Metropolis draw. It was hoped that if a good rate of "jump-shrinkage" were determined, then the rejection rate could be maintained for each parameter at around 25%-35% throughout the nested sampling algorithm, which we estimated to be a nearly optimal rate [9].

We were not able to achieve this. Typically, no matter what the shrinking rate was for the parameters, some of them would end up achieving acceptance with near certainty while others achieved rejection with near certainty. (A typical outcome is shown in Figure B.18, which records the seven parameters' rejection rates over

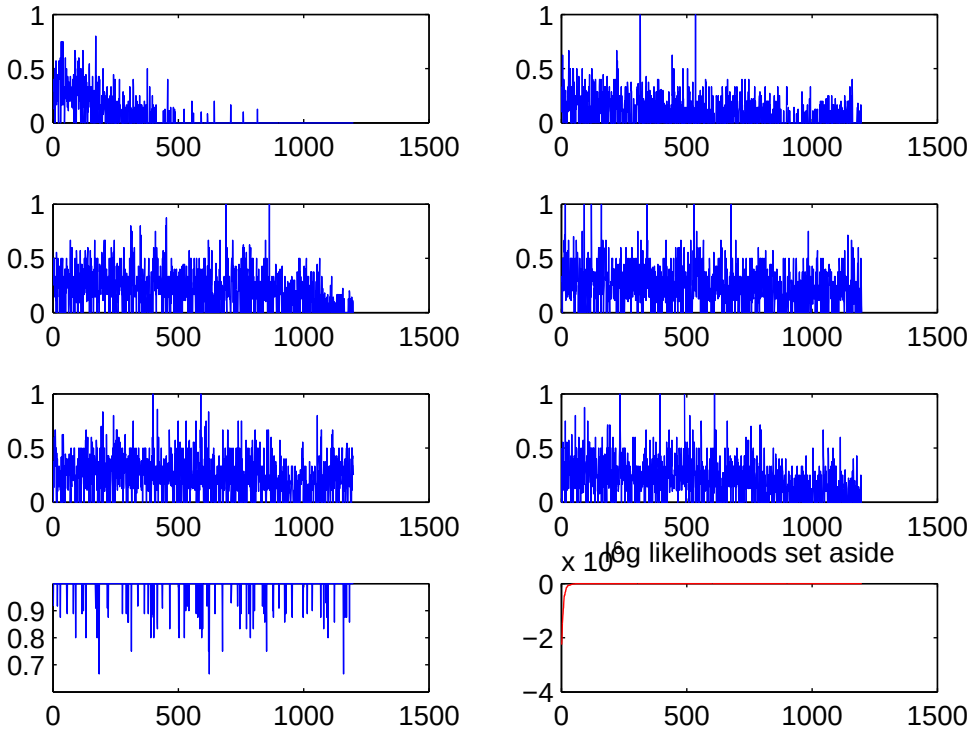

FIGURE B.18: When the proposed jump sizes of the Metropolis algorithm shrink at a constant rate, some parameters achieve near certain acceptance while others achieve near certain rejection.

1200 Metropolis steps.) This suggests that the fractions of the total space Θ that are occupied by the mode and its neighborhood in the seven dimensions are not all the same. In particular, the likelihood function is so very sensitive to the period parameter P that the mode is a very narrow spike in that parameter, while its sensitivity to the stellar jitter parameter s is much smaller. So P ends up with near-certain rejections while s ends up with near-certain acceptances.

After performing many such iterations and obtaining several hundreds of estimates of the desired integral, all fell short by many orders of magnitude. The greatest estimate obtained in 500 estimates was $8/8 \times 10^{-37}$, which is 35 times too great.

The problem, as before, is that the Metropolis algorithm remains unable to find the global mode, so the algorithm ends up settling on some other mode instead.

B.10.2 Jump proposals shrinking at a variable rate

We tried shrinking the proposed jump sizes according to a function that would shrink rapidly at first (since there is so much empty space to vacate so quickly in the initial stages of the algorithm) and then slow down as the live points coalesced around one or more modes.

In our earlier shrinking schedule, the proposed jump sizes shrank linearly towards zero, so if there were to be n jumps in all, then the i^{th} one would be equal to the initial proposed jump size multiplied by a factor of $\frac{n-i}{n}$. Here, we determined the i^{th} jump size by multiplying the initial proposed jump size by a factor of $\left(\frac{n-i}{n}\right)^k$, where k was a parameter that most likely should be greater than 1 (so as to achieve the initially rapid shrinkage).

Even with k chosen separately for each parameter, it was still very unusual for the nested sampling algorithm to discover on its own the location of the global mode. Typical values of the integral estimate were off by factors of 50 to 5000.

B.11 Seeded nested sampling with the known global mode introduced at random times, and then accounted for in the integral estimate

Since the trouble with every attempt to this point has been that the nested sampling cannot by itself find a global mode among many thousands of local modes, we next tried "seeding" the algorithm with one point that was in fact the MLE. The rest of the algorithm would proceed normally. The hope was that the nested sampling algorithm would explore the entire parameter space, accumulating contributions to the integral estimate, but the seeded MLE would guarantee that every so often, the Metropolis algorithm would choose the MLE as its starting point, and so would have

a good chance of selecting a new point near the global mode. In that way, the points would be guaranteed to eventually coalesce around this mode.

However, with the MLE having been seeded initially, we knew that the integral estimates we obtained would be too large—the live points would congregate more quickly in a small pocket than they would naturally have done. So the eventual estimate would have to be adjusted appropriately. We wanted to be sampling at each stage from the restricted prior π^*, but in fact we were sampling from a mixture of π^* and a point mass at the MLE. With N points in the population, the distribution we actually drew from had $\frac{1}{N}$ of its mass on the MLE and $\frac{N-1}{N}$ of its mass on π^*.

Desired integral: $I = \int \pi(\theta)L(\theta)d\theta$.

Actual integral obtained: $I_{obt} = \frac{N-1}{N}\int \pi(\theta)L(\theta)d\theta + \frac{1}{N}L(MLE)$

Solving for the desired integral, we get:

$I = \frac{N}{N-1}I_{obt} - \frac{1}{N-1}L(MLE)$

And here's the problem: The likelihood at the MLE is so many times greater than I (which is, after all, a weighted average of likelihoods from all over the entire parameter space), that although the above equation is correct, estimating I in this way is like trying to estimate the weight of a grain of rice by weighing it simultaneously with an elephant and then subtracting out the weight of the elephant. The answer you'll get will be unbiased (as indeed our estimates were), but with such great variability that the estimate is as likely to be a negative value as a positive one, and in any case off the mark by many orders of magnitude.

We tried systematically choosing the MLE as the starting point exactly once every N times, and also letting it be chosen at random with probability $\frac{1}{N}$. We also tried correcting for its introduction by using the theoretical introduction rate $\frac{1}{N}$ and also by using the actual introduction rate that occurred in the implementation of the nested algorithm. In all cases, the uncertainty associated with the integral we were actually estimating was orders of magnitude greater than the value of the integral

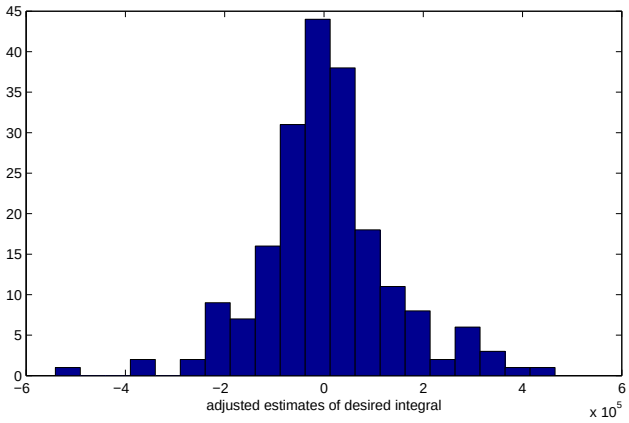

FIGURE B.19: These are unbiased estimates of the desired integral, but since its value is 3×10^{-35}, the variability in the estimates renders them useless.

we really wanted. Here is a typical result: the histogram below (Figure B.19) shows a histogram of the adjusted estimates of the desired integral, resulting from 200 implementations of the nested sampling algorithm. It appears to produce unbiased results, but the variability inherent in the integral we're actually estimating with the nested algorithm is far greater than the integral itself. What we wanted to measure carefully with nested sampling was the weight that the likelihood function carries in lots of different regions of the parameter space. All we've ended up measuring is how often we ended up sampling the MLE by chance, and/or where the Metropolis algorithm happened to take it.

B.12 Conclusion

In conclusion, we do not feel that nested sampling is a good approach to finding the desired integral for this model selection problem. It has the advantage of sampling sparsely from regions of low likelihood and densely from regions of high likelihood—a desirable property. It also does not necessarily require you to know where any modes of the likelihood function are—also desirable, though not essential.

However, it has a fatal problem: it cannot easily find the global mode when the data are sufficiently numerous for that mode to be narrow; and moreover, it can't "know" that it hasn't found the global mode. Instead, it will necessarily eventually settle on some local mode, resulting in an underestimate—probably a gross underestimate—of the actual desired integral. The practical outcome would be a conclusion that the star hosts no planet, when in fact there may be a very high probability that the star hosts a planet, as is the case with HD88133.

No, we have to have a smarter way to find where modes are. Once modes are identified, importance sampling can provide accurate and fairly efficient estimates of the integrals required for model selection, and it can be determined whether a planet exists or not. One method commonly used for identifying modes in the parameter space is the Lomb-Scargle periodogram. In the next two chapters of this paper, we discuss the Lomb-Scargle periodogram and discuss how good it is at finding the modes of the likelihood function.

We hope to show that although it is a very good tool to use when the orbit of a planet is fairly circular, it is less efficient when an orbit is elliptical, and may have already resulted in the undercounting of planets with highly elliptical orbits, due to a premature determination that no planet was present.

Bibliography

[1] Ford, Eric B. & Gregory, Philip C., *Bayesian Model Selection and Extrasolar Planet Detection*, ASP Conference Series, *date?*.

[2] Marcy, Geoffrey W., Butler, R. Paul, & Vogt, Steven S. *Sub-Saturn Planetary Candidates of HD 16141 and HD 46375*, The Astrophysical Journal, Volume 536, Issue 1, pp. L43-L46, June 2000.

[3] Tinney, C. G.; Butler, R. Paul; Marcy, Geoffrey W.; Jones, Hugh R. A.; Penny, Alan J.; McCarthy, Chris; Carter, Brad D.; & Bond, Jade, *Four New Planets Orbiting Metal-enriched Stars*, The Astrophysical Journal, Volume 587, Issue 1, pp. 423-428, April 2003.

[4] Ford, Eric B., *Quantifying the Uncertainty in the Orbits of Extrasolar Planets*, The Astrphysics Journal, *date?*

[5] Ford, Eric B., *Adaptive Scheduling Algorithms for Planet Searches* Accepted to AJ, *date?*

[6] Skilling, John; *Nested Sampling* American Institute of Physics Conference Proceedings, Vol. 735, Issue 1, pp. 395-405, 16 Nov. 2004.

[7] Mukherjee, Pia; Parkinson, David; and Liddle, Andrew R., *A Nested Sampling Algorithm for Cosmological Model Selection*, The Astrophysical Journal, 638:L51–L54, 20 Feb 2006.

[8] Fischer, D. A., et al, *A Hot Saturn Planet Orbiting HD 88133, from the N2K Consortium*, submitted to the Astrophysics Journal, 13 Sept 2004.

[9] Roberts, Gareth & Rosenthal, Jeffrey. *Optimal Scaling for Various Metropolis-Hastings Algorithms* Statistical Science, Vol. 16, No. 4, pp. 351–367, 2001.

[10] Scargle, Jeffrey *Studies in Astronomical Time Series Analysis II. Statistical Aspects of Spectral Analysis of Unevenly Spaced Data* The Astrophysical Journal, 263:835–853, 15 December 1982.

[11] Gregory, P. C., *A Bayesian Analysis of Extrasolar Planet Data for HD 73526* Astrophysical Journal, 22 Dec 2004.

[12] Tinney, C. G., et al, *The 2:1 Resonant Exoplanetary System Orbiting HD 73526* The Astrophysical Journal 647:594–599, 10 August 2006.

[13] Jones, M. C., & Faddy, M. J. *A Skew Extension of the t-distribution, with applications* Journal of the Royal Statistical Society, Vol. 65., Issue 1, pp. 159–174. 28 January 2003.

[14] Royall, Richard *Statistical Evidence* Chapman & Hall, 1997.

[15] Ford, Eric B., *Adaptive Scheduling Algorithms for Planet Searches* submitted to AJ 30 December 2004.

Biography

Name:

- Floyd Andrew Bullard

Date and place of birth:

- 27 January 1969
- Wilson, North Carolina

Degrees earned:

- B.S. in Mathematical Sciences, the Johns Hopkins University, May 1991
- M.S. in Statistics, The University of North Carolina at Chapel Hill, May 1999
- Ph.D. in Statistical Science, Duke University, May 2009

Employer:

- The North Carolina School of Science and Mathematics, Durham, NC (1999–2003 & 2007–present)