ON THE MASS-PERIOD DISTRIBUTIONS AND CORRELATIONS OF EXTRASOLAR PLANETS

ING-GUEY JIANG,1 LI-CHIN YEH,2 YEN-CHANG CHANG,2 AND WEN-LIANG HUNG3

Received 2007 May 20; accepted 2007 August 28

ABSTRACT

In addition to fitting the data of 233 extrasolar planets with power laws, for the first time in this field we construct a correlated mass-period distribution function of extrasolar planets. An algorithm to generate a pair of positively correlated β-distributed random variables is introduced and used for the construction of correlated distribution functions. We investigate the mass-period correlations of extrasolar planets both in linear and logarithmic spaces, determine the confidence intervals of the correlation coefficients, and confirm that there is a positive mass-period correlation for extrasolar planets. In addition to the paucity of massive close-in planets, which are the main contribution to this correlation, there are other fine structures in the data in the mass-period plane.

Key words: methods: data analysis — methods: numerical — methods: statistical — planetary systems

1. INTRODUCTION

Observational efforts have led to the discovery of more than 200 extrasolar planets (exoplanets) in the mass range 0.03–20 M\textsubscript{J}, with orbital periods from a few days to about 4000 days. Many interesting problems regarding the formation and evolution of planetary systems have therefore been studied with the information provided by these detected systems (Jiang & Ip 2001; Laughlin & Chambers 2001; Kinoshita & Nakai 2001; Gozdziewski & Maciejewski 2001; Ji et al. 2002, 2003; Jiang & Yeh 2004).

Moreover, due to the growing number of detected extrasolar planets, several groups have been working on statistical distributions and possible correlations. Assuming that the mass and period distributions are two independent power-law functions, Tabachnik & Tremaine (2002) used the method of maximum likelihood to determine the best power index. Although they found that the uncertainties in the mass and period distributions are coupled, the study of possible mass-period correlations was beyond their scope due to the principal assumption of two independent power-law functions.

In addition, Zucker & Mazeh (2002) calculated the correlation coefficient between mass and period for the detected data in \(\ln P - \ln M\) space. They used Monte Carlo simulations to determine the \(p\)-value for testing whether the correlation is significant. They concluded that the mass-period correlation is significant. However, at the time of their study, the number of detected exoplanets was limited, so only 66 planets were used.

Over the years, many more exoplanets with different properties have been discovered. For example, more hot Jupiters have been found using transit surveys, a few newly detected exoplanets have been found to be moving on extremely eccentric orbits, and exoplanets with mass on the order of Earth mass have been discovered. These results have in fact brought this field into a completely new era. Jiang et al. (2006) did cluster analysis on 143 samples and found that the data grouping could be related to the dynamical processes of planetary systems. This approach was followed by Marchi (2007), in which an extrasolar planet taxonomy was presented.

Therefore, it is about time to construct a new distribution function. As in Tabachnik & Tremaine (2002), assuming the mass and period are two independent power-law distributions, we took 233 examples of exoplanets from The Extrasolar Planets Encyclopaedia\textsuperscript{4} on 2007 July 6 to construct updated distribution functions. Our samples do not include OGLE-235-MOA-53b, 2M1207b, GQ Lup b, HD 187123c, AB Pic b, SCR 1845b, or SWEEPS-04 because either their period or their mass is unknown. The three outliers, PSR 1257+12b, HD 154345b, and PSR B1620–26b, with either extremely small masses or huge periods, are also excluded. (Note that “masses” means the value of projected mass in this paper.)

Moreover, because the mass and period are likely to be correlated according to the results of Zucker & Mazeh (2002), a new distribution function without the assumption that the mass and period are independent would be more satisfactory. That is, we hope to construct a new distribution function in which the mass and period can be coupled. This was not possible until an algorithm for generating two positively correlated \(\beta\)-distributed random variables was provided in Magnussen (2004). We therefore have to employ the \(\beta\)-distribution for this part of the calculations.

After that, in order to carefully investigate the possible mass-period correlations, we work on the correlation coefficients for the data in both linear and logarithmic spaces. A standard method in statistics called the “bootstrap method” is used to find the confidence intervals of the correlation coefficients.

We present the construction of the new mass-period power-law distribution function in § 2 and the correlated mass-period distribution function in § 3. We present the bootstrap method in § 4 and describe the results of the correlation coefficients and confidence intervals in § 5. Finally, we provide concluding remarks in § 6.

2. THE POWER-LAW DISTRIBUTION FUNCTION

In this section we construct a new distribution function, assuming that the mass and period are two independent power laws. We consider the probability density function (pdf) of the power law to have the form

\[
f_{\text{power}}(x) = C(k)x^{-k}, \quad 0 < a < x < b < \infty, \tag{1}
\]

\textsuperscript{4} Available at http://exoplanet.eu/catalog-all.php.

---

1 Department of Physics and Institute of Astronomy, National Tsing-Hua University, Hsin-Chu, Taiwan; jiang@phys.nthu.edu.tw.
2 Department of Applied Mathematics, National Hsinchu University of Education, Hsin-Chu, Taiwan.
3 Graduate Institute of Computer Science, National Hsinchu University of Education, Hsin-Chu, Taiwan.
where the exponent \( k \) is an unknown parameter and the constant \( C(k) \) is given by the normalization requirement that
\[
1 = \int_a^b f_{\text{power}}(x|k) \, dx = \frac{C(k)}{1-k} (b^{1-k} - a^{1-k}).
\]
That is,
\[
C(k) = \frac{1-k}{b^{1-k} - a^{1-k}}.
\]

When sampling from a population described by equation (1), the parameter \( k \) yields knowledge of the entire population. Hence, it is natural to seek a method of finding a good estimator of \( k \). The method of maximum likelihood is one of the most popular techniques for deriving estimators. Letting \( X_1, \ldots, X_n \) be independent and identically distributed (IID) samples from the pdf \( f_{\text{power}}(x|k) \), the likelihood function is given by
\[
L(k|x_1, \ldots, x_n) = \prod_{i=1}^n f_{\text{power}}(x_i|k).
\]
Taking the logarithm of both sides, differentiating partially with respect to the parameter \( k \), and setting the result to zero we can determine the maximum likelihood estimate (MLE) of \( k \) by solving
\[
\frac{a^{1-k} - b^{1-k} + (1-k)(b^{1-k} \ln b - a^{1-k} \ln a)}{(1-k)(b^{1-k} - a^{1-k})} = \frac{\sum_{i=1}^n \ln x_i}{n}.
\]

Now we use the power-law distributions
\[
\begin{align*}
    f_{\text{power}}^M(m|k_m) &= C(k_m)m^{-k_m}, \quad a_m < m < b_m, \\
    f_{\text{power}}^P(p|k_p) &= C(k_p)p^{-k_p}, \quad a_p < p < b_p,
\end{align*}
\]
to fit the 233 observed data in the \( M \) and \( P \) spaces, respectively. First, we choose the range of \( M \) as follows:
\[
a_m = M_{\text{min}}, \quad b_m = M_{\text{max}},
\]
where \( M_{\text{min}} \) and \( M_{\text{max}} \) are the smallest and largest mass of the data set. That is, \( a_m = 0.012 \) and \( b_m = 18.4 \). Using equation (2) we obtain the MLE of \( k_m \), which is \( k_m = 0.7805 \). Similarly, the range of \( P \) is \( a_p = 1.211909, \quad b_p = 4517.4 \), and the MLE of \( k_p \) is \( k_p = 0.9277 \).

Histograms of the observed data in the \( M \) and \( P \) spaces are shown in Figure 1. Figure 1a shows \( M \), and the area covered by this histogram is 115.5. We define the curve,
\[
    f_{\text{hist}}^{M} = 115.5 f_{\text{power}}^M(m|k_m),
\]
and plot it as a dotted curve in Figure 1a for comparison. Similarly, the histogram in Figure 1b shows \( P \), and its area is 11,650. We define another curve,
\[
    f_{\text{hist}}^{P} = 11650 f_{\text{power}}^P(p|k_p),
\]
and plot it as a dotted curve in Figure 1b.

3. THE CORRELATED DISTRIBUTION FUNCTION

In §2 the distributions of mass and period of the extrasolar planets are described by two independent power laws. However, using a data set of 66 exoplanets, Zucker & Mazeh (2002) suggested a possible mass-period correlation. Further, our data of 233 exoplanets show that the correlation coefficient in \( M-P \) space is 0.1762, and this indicates that there exists a positive correlation between \( M \) and \( P \). It is thus not suitable to use two independent power laws to describe the joint mass-period distribution. Therefore, we need to know how to simultaneously describe and use probability models to elicit information from the mass and period measurements. It is necessary to construct a new distribution function in which the mass and period can be positively correlated and coupled. This was not possible until an algorithm for generating two positively correlated \( \beta \)-distributed random variables was provided in Magnusson (2004). This is why we decided to proceed with the \( \beta \)-distribution here.

The \( \beta \)-distributions are very general, have many possible different functional shapes, and have the advantage that variable boundaries and normalizations are automatically considered. The \( \beta \)-distributions are continuous on the finite interval \((c, d)\), \(-\infty < c < d < \infty\), indexed by two positive parameters \( \alpha \) and \( \beta \). The pdf is given by
\[
f_{\beta}(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \frac{(x-c)^{\alpha-1}(d-x)^{\beta-1}}{(d-c)^{\alpha+\beta-1}},
\]
\(c < x < d, \quad \alpha > 0, \quad \beta > 0\),
\[
B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} \, dt.
\]
In equation (3) the pdf \( f_{\beta}(x|\alpha, \beta) \) satisfies \( \int_c^d f_{\beta}(x|\alpha, \beta) \, dx = 1 \). The \( \beta \)-function \( B(\alpha, \beta) \) can be expressed as
\[
B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)},
\]
where the $\Gamma$-function is
\[
\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} \, dt.
\] (4)

Considering the transformation
\[
y = \frac{x - c}{d - c},
\]
we then have the pdf
\[
f(y|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad 0 < y < 1,
\] (5)

which is called the “standard $\beta$-distribution.”

The $\beta$-distribution is often used to model a phenomenon which could be described by the values of random variables defined in a finite interval. As parameters $\alpha$ and $\beta$ vary, the $\beta$-distribution takes on many shapes, as shown in Figure 2. The pdf can be strictly increasing, strictly decreasing, or U-shaped. The case $\alpha = \beta$ yields a pdf symmetric about $(c + d)/2$. If $\beta = 1$, the distribution is called a “power-function distribution.” That is, this distribution is one kind of power-law function.

From the above discussion, it is clear that the $\beta$-distribution is very versatile and can be used for many different purposes. This flexibility encourages its empirical use in a wide range of applications. For example, Wall et al. (2000) successfully used the $\beta$-distribution to model both the subgrid-scale pdf and the subgrid-scale Favor pdf of the mixture fraction. Ettoumi et al. (2002) used $\beta$-distributions to analyze solar measurements in Algeria. Flynn (2004) suggested the $\beta$-distribution as a suitable model for human exposure to airborne contaminants. Ji et al. (2005) proposed a $\beta$-mixture model to analyze a large number of correlation coefficients in bioinformatics.

The standard $\beta$-function is for one variable only. If there is more than one variable and they are independent of each other, the extension to multivariable cases is straightforward. However, it is not easy to create a generalized $\beta$-function with a pair of correlated variables. A few algorithms have been proposed to numerically generate pairs of correlated $\beta$-distributed random variables (see Johnson 1987; Loukas 1984; Michael & Schucany 2002). Due to the limitations of these algorithms, they can only be used for particular types of data sets. The algorithm in Magnusson (2004) can generate a pair of positively correlated $\beta$-distributed random variables without any limitations. We thus use it to construct the numerical mass-period distribution function $f(M, P|\alpha_m, \beta_m, \alpha_p, \beta_p)$.

The probability of a planet with mass and orbital period in the range $[M, M + dM]$, $[P, P + dP]$ is given by
\[
f(M, P|\alpha_m, \beta_m, \alpha_p, \beta_p) \, dM \, dP,
\]

where the marginal distributions of $M$ and $P$ follow the $\beta$-distribution as in equation (3), with parameters $(\alpha_m, \beta_m)$ and $(\alpha_p, \beta_p)$, respectively. From the data, the boundaries can be set such that $m_1 < M < m_2$ and $p_1 < P < p_2$. Then the marginal distributions of the random variables
\[
M_1 = M - m_1 \quad \text{and} \quad P_1 = P - p_1
\]
should follow the standard $\beta$-distributions as in equation (5), with parameters $(\alpha_m, \beta_m)$ and $(\alpha_p, \beta_p)$, respectively.

Let us define
\[
\delta_1 = \rho(M_1, P_1)(1 + \alpha_m + \alpha_p)C, \quad \delta_2 = \rho(M_1, P_1)(1 + \beta_m + \beta_p)C,
\] (6) (7)

where
\[
C = \sqrt{\alpha_m \alpha_p \beta_m \beta_p (1 + \alpha_m + \beta_m)(1 + \alpha_p + \beta_p)}
\]

\[
\times \left\{ (1 + \alpha_p)(1 + \beta_m)(1 + \beta_p)
\right. \\
\left. + \alpha_m [1 + \beta_m + \beta_p + \beta_m \beta_p + \alpha_p (1 + \beta_m + \beta_p)] \right\}^{-1}
\] (8)

and $\rho(M_1, P_1)$ is the correlation coefficient between $M_1$ and $P_1$. Then $M_1$ and $P_1$ generated by the following equations would be a pair of correlated $\beta$-distributed variables:
\[
M_1 = \frac{G(\alpha_m^*) + G(\delta_1)}{G(\alpha_m^*) + G(\delta_1) + G(\beta_m^*) + G(\delta_2)},
\] (9)
\[
P_1 = \frac{G(\alpha_p^*) + G(\delta_1)}{G(\alpha_p^*) + G(\delta_1) + G(\beta_p^*) + G(\delta_2)},
\] (10)

where $G(\alpha)$ is a random variable distributed as a $\Gamma$-distribution with parameters $\alpha$ and 1 and $\alpha_m^*$, $\beta_m^*$, $\alpha_p^*$, and $\beta_p^*$ are defined by
\[
\alpha_m^* = \alpha_m - \delta_1,
\] (11)
\[
\beta_m^* = \beta_m - \delta_2,
\] (12)
\[
\alpha_p^* = \alpha_p - \delta_1,
\] (13)
\[
\beta_p^* = \beta_p - \delta_2.
\] (14)
Note that the pdf of a $\Gamma$-distribution with parameters $\alpha$ and $\beta$ is (Hogg & Craig 1989)

$$f(x) = \frac{1}{\Gamma(\alpha)/\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta},$$

where $0 < x < \infty$.

The above procedure to generate a pair of positively correlated $\beta$-distributed $M_1$ and $P_1$ variables can be summarized as follows (the Magnussen algorithm):

1. Assume that the marginal distribution of $M_1$, i.e., $f_{M_1}(m|\alpha_m, \beta_m)$, is a standard $\beta$-distribution with parameters $(\alpha_m, \beta_m)$. Using the maximum likelihood method, we employ the data to get the best estimation $(\hat{\alpha}_m, \hat{\beta}_m)$ of $(\alpha_m, \beta_m)$. Similarly, we also get the best estimation $(\hat{\alpha}_p, \hat{\beta}_p)$ of $(\alpha_p, \beta_p)$ for $P_1$.
2. Calculate the value of $C$ using equation (8).
3. Calculate the correlation coefficient $\rho(M_1, P_1)$ from the data and use it as the value of $\rho(M_1, P_1)$.
4. Calculate $\delta_1$ and $\delta_2$ using equations (6) and (7).
5. Calculate $\alpha^*_m$, $\beta^*_m$, $\alpha^*_p$, and $\beta^*_p$ using equations (11)–(14).
6. Generate pairs of $M_1$ and $P_1$ using equations (9) and (10).

We now apply the above algorithm to the data set of 233 exoplanets. To avoid a possible singularity, the range of $M$ is chosen to be $(M_{\text{min}}/1.5, 1.5M_{\text{max}})$. That is, $m_1 = 0.008, m_2 = 26.7$. For the same reason, the range of $P$ is taken as $P_1 = P_{\text{min}}/1.5 = 0.8079, P_2 = 1.5P_{\text{max}} = 6776.1$, where $P_{\text{min}}$ and $P_{\text{max}}$ are the smallest and largest periods of the observed data. Thus, the MLEs of $\alpha_m, \beta_m, \alpha_p$, and $\beta_p$ are $\hat{\alpha}_m = 0.6524, \hat{\beta}_m = 5.9076, \hat{\alpha}_p = 0.3697, \text{and } \hat{\beta}_p = 3.8445$, respectively. In addition, the data show that the mass-period correlation coefficient $\rho(M_1, P_1) = 0.1762$. We then get all the parameters’ values as $C = 0.2092, \delta_1 = 0.0745, \delta_2 = 0.3961, \alpha_m = 0.5779, \beta_m = 5.5115, \alpha_p = 0.2952$, and $\beta_p = 3.4484$.

Because the area of the histogram for $M$ in Figure 1a is 115.5, we define the curve,

$$f_{\beta}^{\text{hist}}(m|\hat{\alpha}_m, \hat{\beta}_m),$$

and plot it as a solid curve in Figure 1a for comparison. Similarly, the area of the histogram for $P$ in Figure 1b is 11650, so we define another curve,

$$f_{\beta}^{\text{hist}}(p|\hat{\alpha}_p, \hat{\beta}_p),$$

and plot it as a solid curve in Figure 1b.

These plots indicate that the $\beta$-distribution presents a better fit to the data, compared with the power law. Because there is no closed form for the positively correlated $\beta$-distribution, we numerically plot the joint distribution of $M$ and $P$ as shown in Figure 3. Figure 3 presents the three-dimensional plot of our correlated mass-period distribution function. Figure 4 shows its contour in smaller ranges of mass and period. We thus successfully construct, for the first time in this field, the correlated mass-period distribution function.

Note that for pairs of quantities $(x_i, y_i), i = 1, \ldots, n$, the correlation coefficient $\theta$ is usually given by

$$\hat{\theta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}},$$

where $\bar{x} = \sum_{i=1}^n x_i/n$ and $\bar{y} = \sum_{i=1}^n y_i/n$. The value of $\hat{\theta}$ lies between $-1$ and 1. If the data points all lie on a straight line with positive (negative) slope, the correlation coefficient $\hat{\theta} = 1 (-1)$, which is called “complete positive correlation” (“complete negative correlation”). When the data points are randomly distributed, the variables $x$ and $y$ are uncorrelated and $\hat{\theta}$ is near zero. Thus, the value $\hat{\theta}$ is regarded as one conventional way to quantitatively
describe the strength of the relationship between $x$ and $y$. Thus, our value of $\hat{\rho}(M_1, P_1) = 0.1762$ was obtained by the above equation.

4. THE BOOTSTRAP CONFIDENCE INTERVALS

To assess the statistical significance of the possible correlation, it would be good to have the confidence interval corresponding to a given confidence level. We use the bootstrap method to construct confidence intervals here. Statistically, we determine the characteristics of the population by taking samples. Since the sample should give us information about the population, analogous characteristics of the sample should give us information about the population characteristics. The bootstrap method proposed by Efron (1979) is a simple and straightforward method to calculate the approximate biases, standard deviations, and confidence intervals, for example. It gives the population characteristics by repeatedly taking samples from the original data set.

The bootstrap method is used for IID data. DiCiccio & Efron (1996) found that the bootstrap confidence intervals are more accurate than the classical normal approximation intervals. The standard bootstrap method for confidence intervals can be described as follows. Given an observed IID sample $z = \{z_1, \ldots, z_n\}$ from an unknown distribution function $F$, we want to construct a confidence interval for an interesting parameter $\theta = \theta(F)$ based on $z$. Let $\hat{F}$ be the empirical distribution function, which is defined to be the discrete distribution that assigns the probability $1/n$ to each value $z_i$, $i = 1, \ldots, n$. The key idea of the bootstrap method is a bootstrap sample, which is defined to be a random sample of size $n$ drawn from $\hat{F}$, say, $\hat{z} = \{\hat{z}_1, \ldots, \hat{z}_n\}$. That is, the bootstrap data points $\hat{z}_1, \ldots, \hat{z}_n$ are a random sample of size $n$ drawn from the data set $\{z_1, \ldots, z_n\}$.

Let $\hat{\theta} = s(z)$ be an estimate of $\theta$. Corresponding to a bootstrap data set $\hat{z}^*$ is a bootstrap replication of $\hat{\theta}$,

$$\hat{\theta}^* = s(\hat{z}^*).$$

The quantity $s(\hat{z}^*)$ is the result of applying the same function $s(\ldots)$ to $\hat{z}^*$ as was applied to $z$. For example, if $s(z)$ is the sample mean $\bar{z} = \sum_{i=1}^{n} z_i/n$, then $s(\hat{z}^*)$ is the mean of the bootstrap data set, $\bar{\hat{z}} = \sum_{i=1}^{n} \hat{z}_i/n$. The bootstrap algorithm, described as follows, is a data-based simulation procedure to obtain a good approximation of the confidence interval for $\theta$.

1. Draw a bootstrap sample $\hat{z}^* = (\hat{z}_1^*, \ldots, \hat{z}_n^*)$ according to $\hat{F}$.
2. Evaluate $\hat{\theta}^* = s(\hat{z}^*)$, where $s(\hat{z}^*)$ is the value of $s(z)$ based on $\hat{z}^*$. Repeat the first two steps a large number of times, say, $B$ times, to obtain $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$.
3. Sort $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ into the ordered list $\hat{\theta}_{(1)}^*, \ldots, \hat{\theta}_{(B)}^*$.
4. Let $\hat{\theta}_{(\alpha B)}^*, 0 < \alpha < 0.5$ be the $(100\alpha)$th empirical percentile of $\hat{\theta}_{(b)}, b = 1, \ldots, B$, that is, the $(\alpha B)$th value in the ordered list of $\hat{\theta}_{(1)}^*, \ldots, \hat{\theta}_{(B)}^*$. Likewise, let $\hat{\theta}_{(1 - \alpha B)}^*$ be the $(1 - \alpha B)$th value in the ordered list of $\hat{\theta}_{(1)}^*, \ldots, \hat{\theta}_{(B)}^*$. Then, the approximate $1 - 2\alpha$ confidence interval for $\theta$ is $[\hat{\theta}_{(\alpha B)}^*, \hat{\theta}_{(1 - \alpha B)}^*]$.

Moreover, if $B\alpha$ is not an integer, the following procedure can be used: Assuming $0 < \alpha < 0.5$, let $k = \lfloor (B + 1)\alpha \rfloor$, i.e., the largest integer $\leq (B + 1)\alpha$. Then we define $\hat{\theta}_{(\alpha)}^*$ and $\hat{\theta}_{(1 - \alpha)}^*$ by the $k$th and $(B + 1 - k)$th values in the ordered list of $\hat{\theta}_{(1)}^*, \ldots, \hat{\theta}_{(B)}^*$.

In this paper the bootstrap algorithm is used to construct confidence intervals for correlation coefficients. Assuming that $n$ objects are sampled from a population and two numerical characteristics are measured on each of them, we end up with bivariate random sample points $z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)$. Let $\theta$ be the population correlation coefficient. Based on $n$ data points $z_1, \ldots, z_n$, the population correlation coefficient $\theta$ is estimated by the sample correlation coefficient $\hat{\theta}$ defined in equation (16). Independent repetitions of the bootstrap sampling process give $B$ bootstrap replications $\hat{\theta}_1, \ldots, \hat{\theta}_B$. Then we obtain the approximate $1 - 2\alpha$ confidence interval for $\theta$, which is $[\hat{\theta}_{(\alpha B)}^*, \hat{\theta}_{(1 - \alpha B)}^*]$.

According to Efron & Tibshirani (1993), the number of independent repetitions of bootstrap sampling shall be set as $B = 2000$.

5. MASS-PERIOD CORRELATIONS

We first work on the mass-period correlation in $M$-$P$ space. With the 233 observed data, we calculate correlation coefficients
and use the bootstrap method to determine confidence intervals. We find that the correlation coefficient between $M$ and $P$ is 0.1762 and that the resulting 95% bootstrap confidence interval is (0.0575, 0.3130). This indicates that the mass $M$ and period $P$ have a weak positive correlation. Figure 5 shows the 233 samples in $M$-$P$ space. Indeed, it looks as though the distribution is not completely random. However, the positive correlation is difficult to recognize, so it is consistent with a weak correlation.

On the other hand, for the mass and period distributions in $\ln M$-$\ln P$ space, the correlation coefficient between $\ln M$ and $\ln P$ is 0.3876, and the corresponding 95% bootstrap confidence interval is (0.2668, 0.5001). It clearly indicates that there is a positive mass-period correlation for the data in $\ln M$-$\ln P$ space.

Figure 6 shows the 233 exoplanets in $\ln M$-$\ln P$ space, and it seems that there are many fine structures of data distribution. For example, for those points with $\ln M > 0$, it is far more crowded in the region with $\ln P > 5$. This is, of course, related to the deficit of massive close-in planets, which make the main contribution to the positive mass-period correlation. On the other hand, partially by using transit surveys, the discovery of many hot Jupiters also makes it a bit crowded around ($\ln M$, $\ln P$) = (−0.5, 1). However, the deficit of sub-Jupiter mass planets at separations of about 0.5 AU mentioned in Papaloizou & Terquem (2006) seems to disappear in this plot of 233 data points.

In fact, the mass-period correlation was theoretically studied in Pätzold & Rauer (2002) and Jiang et al. (2003), which focused on the paucity of massive close-in planets. The explanation for this paucity and the correlation could be that tidal interactions with host stars make the massive close-in planets migrate inward and finally merge with the stars.

6. CONCLUDING REMARKS

In this paper we first construct a new mass-period distribution function of exoplanets in which the correlation is considered. This is the first time this has been done in this field and was not possible until the method was proposed in Magnussen (2004). The correlation coefficients of exoplanet data in $M$-$P$ space and $\ln M$-$\ln P$ space are further determined, and the bootstrap method is then used to construct the confidence intervals of correlation coefficients at particular confidence levels. We confirm that there is a mass-period correlation for exoplanets. In addition to the paucity of massive close-in planets, there are other fine structures in the data distribution to be investigated in the future.

REFERENCES

DiCiccio, T. J., & Efron, B. 1996, Statist. Sci., 11, 189
Efron, B. 1979, Ann. Statist., 7, 1
Efron, B., & Tibshirani, R. J. 1993, An Introduction to the Bootstrap (New York: Chapman & Hall)
Ettoumi, F. Y., Mefti, A., Adane, A., & Bouroubi, M. Y. 2002, Renewable Energy, 26, 47
Flynn, M. R. 2004, Stochastic Environ. Res. Risk Assessment, 18, 306
Gozdziewski, K., & Maciejewski, A. J. 2001, ApJ, 563, L81
Hogg, R. V., & Craig, A. T. 1989, Introduction to Mathematical Statistics (New York: Macmillan)
Ji, J., Kinoshita, H., Liu, L., & Li, G. 2003, ApJ, 585, L139
Ji, J., Li, G., & Liu, L. 2002, ApJ, 572, 1041
Ji, Y., Wu, C., Liu, P., Wang, J., & Combes, K. R. 2005, Bioinformatics, 21, 2118
Jiang, I.-G., & Ip, W.-H. 2001, A&A, 367, 943
Jiang, I.-G., Ip, W.-H., & Yeh, L.-C. 2003, ApJ, 582, 449
Jiang, I.-G., & Yeh, L.-C. 2004, MNRAS, 355, L29
Jiang, I.-G., Yeh, L.-C., Hung, W.-L., & Yang, M.-S. 2006, MNRAS, 370, 1379
Johnson, M. E. 1987, Multivariate Statistical Simulation (New York: Wiley)
Kinoshita, H., & Nakai, H. 2001, PASJ, 53, L25
Laughlin, G., & Chambers, J. 2001, ApJ, 551, L109
Loukas, S. 1984, J. Statist. Comput. Simul., 20, 145
Magnussen, S. 2004, Comput. Statist. Data Anal., 46, 397
Marchi, S. 2007, ApJ, 666, 475
Michael, J. R., & Schucany, W. R. 2002, Am. Statist., 56, 48
Papaloizou, J. C. B., & Terquem, C. 2006, Rep. Prog. Phys., 69, 119
Pätzold, M., & Rauer, H. 2002, ApJ, 568, L117
Tabachnik, S., & Tremaine, S. 2002, MNRAS, 335, 151
Wall, C., Boersma, B. J., & Moin, P. 2000, Phys. Fluids, 12, 2522
Zucker, S., & Mazeh, T. 2002, ApJ, 568, L113