Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

Bayesian Markov chain Monte Carlo imputation for the transiting exoplanets with an application in clustering analysis

Huei-Wen Teng, Wen-Liang Hung & Yen-Ju Chao

To cite this article: Huei-Wen Teng, Wen-Liang Hung & Yen-Ju Chao (2015) Bayesian Markov chain Monte Carlo imputation for the transiting exoplanets with an application in clustering analysis, Journal of Applied Statistics, 42:5, 1120-1132, DOI: 10.1080/02664763.2014.995609

To link to this article: http://dx.doi.org/10.1080/02664763.2014.995609

Journal of Applied Statistics, 2015
Vol. 42, No. 5, 1120–1132, http://dx.doi.org/10.1080/02664763.2014.995609

Bayesian Markov chain Monte Carlo imputation for the transiting exoplanets with an application in clustering analysis

Huei-Wen Teng[a], Wen-Liang Hung[b*] and Yen-Ju Chao[c]

[a]*Graduate Institute of Statistics, National Central University, Jhongli, Taiwan;* [b]*Department of Applied Mathematics, National Hsinchu University of Education, Hsin-Chu, Taiwan;* [c]*Cathy United Bank, Taipei, Taiwan*

(*Received 21 August 2013; accepted 3 December 2014*)

To impute the missing values of mass in the transiting exoplanet data, this paper uses the Frank copula to combine two Pareto marginal distributions. Next, a Bayesian Markov chain Monte Carlo (MCMC) imputation method is proposed. The proposed Bayesian MCMC imputation method is found to outperform the mean imputation method. Clustering analysis can shed light on the formation and evolution of exoplanets. After imputing the missing values of mass in the transiting exoplanet data using the proposed approach, the similarity-based clustering method (SCM) clustering algorithm is applied to the logarithm of mass and period for this complete data set. The SCM clustering result indicates two clusters. Furthermore, the intracluster Spearman rank-order correlation coefficients ρ_s for mass and period in these two clusters are 0.401 and −0.188, respectively, at a significance level of 0.01. This result illustrates that the mass and period correlate in an opposite way between the two different clusters. It implies that the formation and evolution processes of these two clusters are different.

Keywords: copula; Metropolis–Hastings algorithm; missing data; hot Jupiters; transiting exoplanets

1. Introduction

Transiting planets are presently our main source of information on the formation, structure and evolution of exoplanets. Because transit surveys are successful in the photometric, enough transiting planets are known to start more statistical methods to their ensemble features. However, there exists limited research investigating the missing data problem for expolanet data. Among them, Käärik [14] used the copula algorithm to impute the missing values. Boone *et al.* [1] utilized the Gibbs Sampler as a tool for incorporating the missing covariate structure.

*Corresponding author. Email: wlhung@mail.nhcue.edu.tw

Indeed, the expectation-maximization (EM) algorithm seeks maximum likelihood in statistics when the model encounters missing data problem and is a commonly used model-based imputation technique. Please refer Dempster *et al*. [2] and Little and Rubin [15]. The EM method alternates between the expectation (E) step and the maximization (M) step: the E step calculates the expectation of the log-likelihood evaluated using the current parameter estimate, whereas the M step maximizes the expected log-likelihood found on the E step. However, the EM algorithm may be inapplicable for less analytically tractable models. In this respect, Bayesian methods are useful because, armed with the Markov chain Monte Carlo (MCMC) simulation, they allow statistical inference for complicated models.

There have been successful Bayesian methods proposed for exoplanet data. Gregory [6] has recently modified the Bayesian MCMC to analyze the transiting exoplanet HD 73526. His result indicated three possible orbits with three different periods. In 2007, Gregory [7] used again a modified MCMC algorithm to detect a second planet in HD 208487. In the same year, he [8] also found evidences for three planets in HD 11964. Later, Gregory and Fischer [10] proposed a hybrid MCMC algorithm and obtained evidences for three planets in 47 Ursae Majoris. When the model parameters are highly correlated, the MCMC algorithm is not efficient enough for exploring the parameter space. To overcome this drawback, Gregory [9] developed an efficient MCMC algorithm in highly correlated model parameter space. However, few studies have been conducted on how to use the MCMC algorithm to impute the missing data in the transiting exoplanet data.

The Gibbs sampler and Metropolis algorithm are general MCMC algorithm. They yield a Markov chain whose equilibrium distribution (under some regularity conditions) is proportional to the likelihood function or the posterior density of interest. According to Sklar's theorem, copulas are employed to represent a multivariate distribution in terms of its marginal distributions [18, 20]. The parameter in the copula is called the dependence parameter, which measures dependence between the marginals. The Frank copula is often used in practical applications (see [17]), hence, in this paper, the Frank copula is employed to combine two marginal Pareto distributions. According to this copula model, a Bayesian MCMC imputation method is proposed for imputing missing values in mass of exoplanets discovered by the transiting method.

After imputing the missing values of the mass, we further consider clustering analysis for the transiting expolanets. Clustering is a powerful exploratory approach to finding groups in data and revealing the structure information of a given data set. The similarity-based clustering method (SCM) clustering algorithm [23] is applied to the complete data. The clustering results reveal two clusters. According to the intracluster correlations within a cluster, the formation and evolution processes of these two clusters are different.

The remainder of the paper is organized as follows. Section 2 presents the proposed methodology and a Metropolis–Hastings algorithm for numerical purposes. Section 3 analyses real data and runs the MCMC simulation to impute missing data. The details of clustering analysis are described in Section 4. The final section gives the conclusion and directions for future work.

2. The proposed methodology

In this section, the Pareto distribution is employed to model the mass and period of transiting exoplanets, and the popular Frank copula is utilized to bind two Pareto marginal distributions. Next, the proposed Bayesian imputation method is presented, and a Hasting-Metropolis algorithm for Bayesian inference is described.

2.1 *A model for transiting exoplanets*

To impute data for transiting exoplanets, let X denote the mass, and Y denote the period, of an observed planet. For transiting expolanets, Pareto distributions, also known as variants of power

law distributions, are commonly used as marginal distribution of mass and period. Following Jiang *et al*. [11] and Tabachnik and Tremaine [21], we assume that both X and Y follow Pareto distributions.

Specifically, X has a probability density function (pdf),

$$f(x) = \frac{\alpha x_{\min}^{\alpha}}{x^{\alpha+1}} \quad \text{for } x \geq x_{\min},$$

where $\alpha > 0$ and $x_{\min}$ is the minimum of observed masses. Y has a pdf

$$g(y) = \frac{\beta y_{\min}^{\beta}}{y^{\beta+1}} \quad \text{for } y \geq y_{\min},$$

where $\beta > 0$ and $y_{\min}$ is the minimum of observed periods.

Combining these two marginal distributions with a Frank copula yields the joint cumulative distribution function (cdf) of (X, Y) as

$$H(x,y) = C(F(x), G(y); \theta), \tag{1}$$

where $C(u_1, u_2; \theta)$ is the Frank copula with dependence parameter θ, and $F(x)$ and $G(y)$ are the cdf of X and Y , respectively. Recall that the Frank copula is of the form,

$$C(u_1, u_2; \theta) = \frac{-1}{\theta} \ln\left[1 + \frac{(\mathrm{e}^{-\theta u_1} - 1)(\mathrm{e}^{-\theta u_2} - 1)}{\mathrm{e}^{-\theta} - 1}\right],$$

where $0 \leq u_1, u_2 \leq 1$ and $\theta \in (-\infty, \infty) \setminus 0$. Differentiating Equation (1) with respect to the first and second arguments gives the joint pdf of (X, Y),

$$h(x,y) = \left[\frac{\theta(1 - \mathrm{e}^{-\theta})\,\mathrm{e}^{-\theta[F(x)+G(y)]}}{(1 - \mathrm{e}^{-\theta} - (1 - \mathrm{e}^{-\theta F(x)})(1 - \mathrm{e}^{-\theta G(y)}))^2}\right] f(x)g(y), \tag{2}$$

which allows the setting of the likelihood function to be completed, as described in the following subsection.

2.2 *A Bayesian imputation method*

Let $(\mathbf{x}, \mathbf{y})$ denote the whole data set. To indicate the missing data with an asterisk in the superscript, the data $(\mathbf{x}, \mathbf{y})$ are relabeled and partitioned into two categories $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2^*, \mathbf{y}_2)$ as follows. Here, the first category consists of complete data of size n_1:

$$(\mathbf{x}_1, \mathbf{y}_1) = ((x_{1,1}, \ldots, x_{1,n_1}), (y_{1,1}, \ldots, y_{1,n_1})),$$

The second category consists of missing data in mass of size n_2:

$$(\mathbf{x}_2^*, \mathbf{y}_2) = ((x_{2,1}^*, \ldots, x_{2,n_2}^*), (y_{2,1}, \cdots, y_{2,n_2})),$$

where $n_1 + n_2 = n$.

Intuitively, $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2^*, \mathbf{y}_2)$ are independent because these observations are independent no matter the data are observed or unobserved. Similar to Tabachnik and Tremaine [21], we assume

independence between the complete data and the missing data for simplicity. By Equation (2), the likelihood function is

$$L(\alpha, \beta, \theta, \mathbf{x}_2^*|\mathbf{x}_1, \mathbf{y}_1, \mathbf{y}_2) = \prod_{i=1}^{n_1} h(x_{1,i}, y_{1,i}) \prod_{j=1}^{n_2} h(x_{2,j}^*, y_{2,j}). \tag{3}$$

Furthermore, let $\pi(\alpha), \pi(\beta)$ and $\pi(\theta)$ be prior distributions of the parameters α, β and θ, respectively. In this subsection, the prior distributions are assumed to be gamma distributions; that is,

$$\begin{aligned}
\pi(\alpha) &= \frac{\lambda_1^{\kappa_1}}{\Gamma(\kappa_1)} \alpha^{\kappa_1 - 1} \mathrm{e}^{-\lambda_1 \alpha}, \\
\pi(\beta) &= \frac{\lambda_2^{\kappa_2}}{\Gamma(\kappa_2)} \beta^{\kappa_2 - 1} \mathrm{e}^{-\lambda_2 \beta}, \\
\pi(\theta) &= \lambda \, \mathrm{e}^{-\lambda \theta},
\end{aligned}$$

where $\kappa_1, \lambda_1, \kappa_2, \lambda_2$, and λ are positive hyperparameters.

For simplicity, we assume that the prior distribution of the missing data is assumed to have the same distribution as the marginal distributions:

$$\pi(\mathbf{x}_2^*) = \prod_{j=1}^{n_2} f(x_{2,j}^*).$$

Therefore, the posterior distribution of parameters, $\pi(\alpha, \beta, \theta, \mathbf{x}_2^*, |\mathbf{x}_1, \mathbf{y}_1, \mathbf{y}_2)$, is

$$\pi(\alpha, \beta, \theta, \mathbf{x}_2^*, |\mathbf{x}_1, \mathbf{y}_1, \mathbf{y}_2) \propto L(\alpha, \beta, \theta, \mathbf{x}_2^*, |\mathbf{x}_1, \mathbf{y}_1, \mathbf{y}_2) \times \pi(\alpha) \times \pi(\beta) \times \pi(\theta) \times \pi(\mathbf{x}_2^*). \tag{4}$$

2.3 *A Metropolis–Hastings algorithm*

Bayesian inference focuses on the posterior distribution Equation (4), which does not have a closed-form formula. To tackle this problem, a Metropolis–Hastings algorithm is implemented to generate samples having a target distribution in Equation (4). A Metropolis–Hastings algorithm is a method of the MCMC algorithm, and it constructs a Markov chain with stationary distribution equal to the target distribution.

In a Metropolis–Hastings algorithm, choosing appropriate proposal distributions that allow the MCMC samples to fully explore the target distribution is a challenge in practice (see [19]). For simplicity, a truncated random walk Metropolis sampling is applied, where (truncated) normal distributions are used as proposal distributions [5]. Here, the parameter space is $\{\alpha, \beta, \theta, \mathbf{x}_2^*\}$. Let $\alpha^{(t)}, \beta^{(t)}, \theta^{(t)}, \mathbf{x_2}^{*(t)}$ be the tth interaction in the MCMC algorithm. The following proposal distributions for each parameter in the MCMC algorithm are used.

$$\begin{aligned}
\alpha^{(t)}|\alpha^{(t-1)} &\sim N(\alpha^{(t-1)}, \tau_\alpha^2) I_{\{\alpha^{(t)} > 0\}}, \\
\beta^{(t)}|\beta^{(t-1)} &\sim N(\beta^{(t-1)}, \tau_\beta^2) I_{\{\beta^{(t)} > 0\}}, \\
\theta^{(t)}|\theta^{(t-1)} &\sim N(\theta^{(t-1)}, \tau_\theta^2), \\
x^{(t)}|x^{(t-1)} &\sim N(x^{(t-1)}, \tau_x^2) I_{\{x^{(t)} > x_{\min}\}},
\end{aligned}$$

where $\tau_\alpha, \tau_\beta, \tau_\theta, \tau_x$ are positive real numbers used as tuning parameters, $I_{\{\cdot\}}$ is the indicator function, and $N(\cdot, \cdot)$ is a normal distribution. For each parameter, the full conditional distributions are

as follows:

$$\pi_f(\alpha) \equiv \pi(\alpha|\beta,\theta,\mathbf{x}_1,\mathbf{y}_1,\mathbf{x}_2) \propto L(\alpha,\beta,\theta,\mathbf{x}_2^*|\mathbf{x}_1,\mathbf{y}_1,\mathbf{y}_2) \times \pi(\alpha),$$
$$\pi_f(\beta) \equiv \pi(\beta|\alpha,\theta,\mathbf{x}_1,\mathbf{y}_1,\mathbf{y}_2) \propto L(\alpha,\beta,\theta,\mathbf{x}_2^*|\mathbf{x}_1,\mathbf{y}_1,\mathbf{y}_2) \times \pi(\beta),$$
$$\pi_f(\theta) \equiv \pi(\theta|\alpha,\beta,\mathbf{x}_1,\mathbf{y}_1,\mathbf{y}_2) \propto L(\alpha,\beta,\theta,\mathbf{x}_2^*|\mathbf{x}_1,\mathbf{y}_1,\mathbf{y}_2) \times \pi(\theta),$$
$$\pi_f(x) \equiv \pi(\mathbf{x}_2^*|\alpha,\beta,\theta,\mathbf{x}_1,\mathbf{y}_1,\mathbf{y}_2) \propto L(\alpha,\beta,\theta,\mathbf{x}_2^*|\mathbf{x}_1,\mathbf{y}_1,\mathbf{y}_2) \times \pi(\mathbf{x}_2^*).$$

Let $\Phi(\cdot)$ denote the cdf of the standard normal distribution. The sampling algorithm is as follows:

(1) Start with suitable initial values $\alpha^{(0)}, \beta^{(0)}, \theta^{(0)}, \mathbf{x}_2^{*(0)}$. Set $t = 1$.
(2) Perform following steps:
   (a) Update $\alpha^* \sim N(\alpha^{(t)}, \tau_\alpha^2)1_{\{\alpha>0\}}$ and set $\alpha^{(t+1)} = \alpha^*$ with probability

$$a(\alpha^*, \alpha^{(t)}) = \min\left(\frac{\pi_f(\alpha^*)\Phi(\alpha^{(t)}/\tau_\alpha)}{\pi_f(\alpha^{(t)})\Phi(\alpha^*/\tau_\alpha)}, 1\right).$$

   (b) Update $\beta^* \sim N(\beta^{(t)}, \tau_\beta^2)1_{\{\beta>0\}}$ and set $\beta^{(t+1)} = \beta^*$ with probability

$$a(\beta^*, \beta^{(t)}) = \min\left(\frac{\pi_f(\beta^*)\Phi(\beta^{(t)}/\tau_\beta)}{\pi_f(\beta^{(t)})\Phi(\beta^*/\tau_\beta)}, 1\right).$$

   (c) Update $\theta^* \sim N(\theta^{(t)}, \tau_\theta^2)$ and set $\theta^{(t+1)} = \theta^*$ with probability

$$a(\theta^*, \theta^{(t)}) = \min\left(\frac{\pi_f(\theta^*)}{\pi_f(\theta^{(t)})}, 1\right).$$

   (d) Update $x^* \sim N(x^{(t)}, \tau_x^2)$ and set $x^{(t+1)} = x^*$ with probability

$$a(x^*, x^{(t)}) = \min\left(\frac{\pi_f(x^*)\Phi((x^{(t)} - x_{\min})/\tau_x)}{\pi_f(x^{(t)}\Phi((x^* - x_{\min})/\tau_x)}, 1\right).$$

   (e) Set $t = t + 1$.
(3) Repeat Step (2) until convergence.

3. Numerical results

To start with, this section describes the exoplanet data. Section 3.2 describes strategies in the MCMC simulation and shows convergence of the simulation. Section 3.3 gives statistical inference on the parameters and imputation results.

3.1 *Data descriptions*

The first discovery of a transiting exoplanet was confirmed by the known radial velocity discovered exoplanet HD209458. This paper aims to impute the missing values of mass in the transiting exoplanet data from the Extrasolar Encyclopedia (http://exoplanet.eu/catalog/) on 16 March 2013.

There are 294 planets available for this work and each of them has the values of the projected (minimum) mass M, which is $M = M_{\text{real}} \sin i$, where M_{real} is the real physical mass of the exoplanet and i is the orbital inclination angle and P is the orbital period. In this data set, there are 39 missing observations of mass. That is, the missing proportion is 13.3%. Therefore, the missing information should not be discarded.

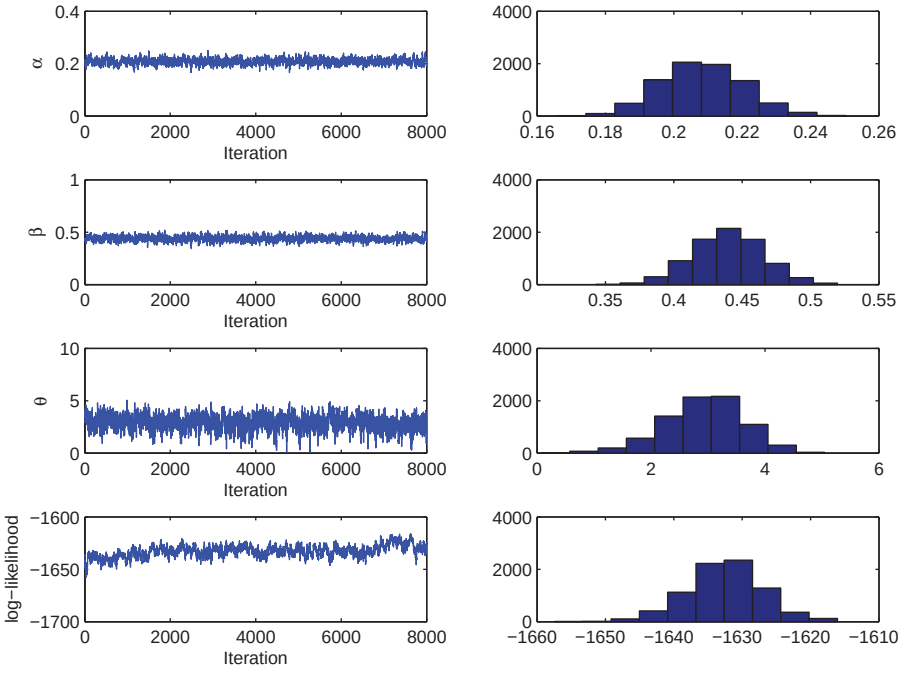

Figure 1. The trace plot and the histogram of α, β, θ, and the log-likelihood of one MCMC simulation of 10,000 iterations with the first 2000 samples are discarded.

3.2 *Convergence of the MCMC simulation*

In a pilot study, tuning parameters are adjusted for possible values until the acceptance rate for each parameter falls between the range of 30% to 70 %. Afterwards, we run the MCMC simulation using these tuning parameters, and the acceptance rates using the selected tuning parameters of α, β, and τ, are 71%, 60% and 56%, respectively.

For convergence diagnostics, we conduct some strategies in [3, 16]. In our implementation, we run a few parallel chains with random starting states and investigate these chains by comparing their trace plots, histograms, and other aspects. Each simulated Markov chain is a run of 10,000 iterations, and the first 2000 samples are discarded as burn-in samples. In unreported tables, statistics (such as posterior means) calculated from different parallel Markov chains seem to be quite close. Figure 1 depicts trace plots and histograms of parameters of interest and log-likelihood as a general diagnostic benchmark after the burn-in period of one Markov chain Monte Carlo simulation. Based on these plots, we conclude that the simulation shows good mixing and appears to be convergent.

The Bayesian imputation is carried out in R program, and running one single Markov chain of 10,000 iterations takes 2.74 h in Notebook TOSHIBA with CPU Intel Core i3-350M 2.27 GHz. As a remark, the computational time of running the MCMC algorithm can be for sure reduced, for example using other program languages such as Matlab or C++, but it is beyond the scope of this paper.

3.3 *Imputation results*

By using the proposed approach, the posterior means of the Pareto parameters of the mass and period are obtained to be 0.199 and 0.437, respectively. The posterior mean of dependence

parameter of the Frank copula, θ equal to 3.624, indicates a positive correlation between mass and period.

According to Genest [4], the parameter θ is related to the Spearman rank-order correlation coefficient ρ_s through the following formula:

$$\rho_s \approx \frac{1 - \theta\, e^{-\theta/2} - e^{-\theta}}{(e^{-\theta/2} - 1)^2}.$$

Thus, the Spearman rank-order correlation coefficient between mass and period is 0.545 indicating a strong positive correlation between mass and period. Furthermore, Figure 1 shows histograms of the posterior distributions of the Pareto parameters α and β of the mass and period, and the dependence parameter θ.

Finally, Table 1 shows the imputation results of the missing values of mass. To evaluate the performance of the proposed imputation method, comparison is made between the proposed method and the mean imputation method in terms of the squared imputation error (IE) (see [22]).

The IE is defined as follows:

$$\text{IE} = \frac{\sum_i (O_i - I_i)^2}{\sum_i O_i^2},$$

where O_i and I_i are the ith observation and imputed value, respectively. Since there are 294 planets in this work and there are 39 missing values of mass, the number of complete data is 255.

Furthermore, the missing proportion is 13.3%. A total of 34 ($255 \times 13.3\%$) observations are randomly chosen to be regarded as missing in mass and both the proposed imputation and mean imputation methods are used to impute these 34 missing values. According to 50 replications, the average IEs of the proposed imputation and mean imputation are 26.08% and 67.27%, respectively; indicating that the proposed imputation approach outperforms the mean imputation method.

Jiang *et al.* [12] took all planets from The Extrasolar Encyclopedia on 10 April 2008 and they excluded planets that have missing values in massed or period. Furthermore, the outlier, PSR B1620-26b, with a huge period (100 years), is also excluded. They obtained that the estimate of dependence parameter θ in the Frank copula is $\hat{\theta} = 2.383$, revealing a positive correlation between mass and period in all the planets. From the above results, the positive correlation between mass and period between transiting exoplanets is larger than all the exoplanets. This is because that Jiang *et al.* [12] discarded observations with any missing values.

4. Clustering analysis

According to Jiang *et al.* [13], the clustering analysis would give hints about the formation and evolution of exoplanets. Jiang *et al.* [13] applied logarithmic transformation to the projected mass M and the orbit period P, and the SCM clustering algorithm [23] to the transformed variables, $\ln M$ and $\ln P$, for the exoplanets. In this section, the SCM is used to analyze the transiting exoplanets using these two logarithmic transformation variables.

First, the missing values of mass are ignored. That is, the SCM algorithm is applied only to the 255 observed data. The corresponding dendrogram is shown in Figure 2. As can be seen, there are three well-separated clusters, say C1, C2 and C3. Table 2 presents the detailed description about the cluster centers and the intracluster correlations ρ_s.

Figure 3 shows the scatter plot of the complete data obtained by using the proposed approach to impute the missing values of mass. According to Figure 3, the transiting probability small explains the absence in the top left-hand corner. After implementing the SCM clustering algorithm, the dendrogram shown in Figure 4 clearly indicates two well-separated clusters. We find that two clusters denoted as '1' and '2' in Figure 5. Table 3 also shows the cluster centers

Table 1. The imputation results of the the missing values of mass.

Data no.	Planet	M	P
1	KIC 10905746 b	0.016875	9.8844
2	KIC 4862625 b	0.061013	138
3	KIC 6185331 b	0.022357	49.76971
4	KIC-8552719 b	0.016927	88.4075
5	KOI-500 b	0.013142	7.05
6	KOI-500 c	0.034111	9.52
7	KOI-500 d	0.011723	3.07
8	KOI-500 e	0.156693	4.64
9	KOI-500 f	0.007660	0.99
10	KOI-730 b	0.020073	14.7903
11	KOI-730 c	0.013986	9.8499
12	KOI-730 d	0.028863	19.7216
13	KOI-730 e	0.014214	7.3831
14	KOI-94 b	0.010632	3.7432451
15	KOI-94 c	0.013120	10.423707
16	KOI-94 d	0.017242	22.3430004
17	KOI-94 e	0.163122	54.319931
18	Kepler-33 b	0.017178	5.66793
19	Kepler-33 c	0.017789	13.17562
20	Kepler-33 d	0.017372	21.77596
21	Kepler-33 e	0.016244	31.7844
22	Kepler-33 f	0.021397	41.02902
23	Kepler-47(AB) b	0.013797	49.514
24	Kepler-47(AB) c	0.622300	303.148
25	Kepler-49 b	0.013044	7.2037945
26	Kepler-49 c	0.012388	10.9129343
27	Kepler-51 b	0.015519	45.1555023
28	Kepler-51 c	1.785766	85.3128662
29	Kepler-54 b	0.020348	8.0109434
30	Kepler-54 c	0.030138	12.0717249
31	Kepler-55 b	0.432675	27.94811449
32	Kepler-55 c	0.653032	42.1516418
33	Kepler-56 b	0.017357	10.5034294
34	Kepler-56 c	0.014381	21.4050484
35	Kepler-59 b	0.007624	11.8681707
36	Kepler-59 c	0.011987	17.9801235
37	Kepler-60 b	0.004664	7.1316185
38	Kepler-60 c	0.005835	8.9193459
39	Kepler-60 d	0.007959	11.9016171

and the intracluster correlations ρ_s for the complete data. Comparing Table 2 with Table 3, shows that ignoring the missing values of mass would lead to overestimation of the number of clusters and the intracluster correlations ρ_s. Note that $0.605 > 0.401, -0.160 > -0.188$.

The cluster MP1 represents the lower mass planets (below 0.5 M_J) and the period is about 11 d. The cluster MP2 represents the hot Jupiters. The most notable of these are HD 209458 b

Table 2. The cluster centers and the intracluster correlations ρ_s for clustering 255 observed transiting exoplanets.

Cluster	Cluster center $(\ln M, \ln P)$	Cluster center (M, P)	Members	ρ_s
C1	$(-1.9327, 5.1170)$	$(0.1448, 166.8341)$	2	None
C2	$(-3.3415, 2.3598)$	$(0.0354, 10.5888)$	44	0.605
C3	$(-0.1976, 1.2111)$	$(0.8207, 3.3572)$	209	-0.160

Table 3. The cluster centers and the intracluster correlations ρ_S for clustering the complete data.

Cluster	Cluster center $(\ln M, \ln P)$	Cluster center (M, P)	Members	ρ_S
MP1	(−3.9394, 2.4111)	(0.0195, 11.1462)	77	0.401
MP2	(−0.1584, 1.2173)	(0.8535, 3.3781)	217	− 0.188

which was the first transiting hot Jupiter found; HD 189733 b which was first mapped in 2007 by the Spitzer Space Telescope and HAT-P-7b which was recently observed by the Kepler mission. Figure 5 shows the members of these two clusters and Table 4 presents the corresponding transiting exoplanets of these two clusters.

According to Table 4, the second cluster MP2 includes the transiting exoplanets from OGLE, HAT, WASP and TrES surveys. Furthermore, the cluster MP2 also includes several planets on

Figure 2. The dendrogram of 255 observed exoplanets in the logarithmic transformation of the projected mass M and the orbit period P. There are too many data points, so the data identity numbers below the horizontal axis are not clear.

Table 4. The clustering results in the $\ln M - \ln P$ plane.

Cluster	Planet
MP1	55 Cnc e, CoRoT-7 b, GJ 1214 b, GJ 3470 b, HAT-P-26 b, HD 97658 b, KIC 10905746 b, KIC 12557548 b, KIC 4862625 b, KIC 6185331 b, KIC-8852719 b, KOI-500 b, KOI-500 c, KOI-500 d, KOI-500 f, KOI-730 b, KOI-730 c, KOI-730 d, KOI-730 e, KOI-94 b, KOI-94 c, KOI-94 d, Kepler-10 b, Kepler-10 c, Kepler-11 b, Kepler-11 c, Kepler-11 d, Kepler-11 e, Kepler-11 f, Kepler-16 (AB) b, Kepler-18 b, Kepler-18 c, Kepler-18 d, Kepler-19 b, Kepler-20 b, Kepler-20 c, Kepler-20 d, Kepler-20 e, Kepler-20 f, Kepler-21 b, Kepler-22 b, Kepler-30 b, Kepler-30 d, Kepler-33 b, Kepler-33 c, Kepler-33 d, Kepler-33 e, Kepler-33 f, Kepler-34(AB) b, Kepler-35(AB) b, Kepler-36 b, Kepler-36 c, Kepler-42 b, Kepler-42 c, Kepler-42 d, Kepler-47(AB) b, Kepler-48 c, Kepler-49 b, Kepler-49 c, Kepler-50 b, Kepler-50 c, Kepler-51 b, Kepler-53 c, Kepler-54 b, Kepler-54 c, Kepler-56 b, Kepler-56 c, Kepler-57 c, Kepler-58 b, Kepler-59 b, Kepler-59 c, Kepler-60 b, Kepler-60 c, Kepler-60 d, Kepler-68 b, Kepler-68 c, Kepler-9 d

(*Continued*).

Table 4. Continued

MP2	55 Cnc e, CoRoT-1 b, CoRoT-10 b, CoRoT-11 b, CoRoT-12 b, CoRoT-13 b, CoRoT-14 b, CoRoT-16 b, CoRoT-17 b, CoRoT-18 b, CoRoT-19 b, CoRoT-2 b, CoRoT-20 b, CoRoT-21 b, CoRoT-23 b, CoRoT-3 b, CoRoT-4 b, CoRoT-5 b, CoRoT-6 b, CoRoT-7 b, CoRoT-8 b, CoRoT-9 b, GJ 1214 b, GJ 3470 b, GJ 436 b, HAT-P-1 b, HAT-P-11 b, HAT-P-12 b, HAT-P-13 b, HAT-P-14 b, HAT-P-15 b, HAT-P-16 b, HAT-P-17 b, HAT-P-18 b, HAT-P-19 b, HAT-P-2 b, HAT-P-20 b, HAT-P-21 b, HAT-P-22 b, HAT-P-23 b, HAT-P-24 b, HAT-P-25 b, HAT-P-26 b, HAT-P-27-WASP-40 b, HAT-P-28 b, HAT-P-29 b, HAT-P-3 b, HAT-P-30-WASP-51 b, HAT-P-31 b, HAT-P-32 b, HAT-P-33 b, HAT-P-34 b, HAT-P-35 b, HAT-P-36 b, HAT-P-37 b, HAT-P-38 b, HAT-P-39 b, HAT-P-4 b, HAT-P-40 b, HAT-P-41 b, HAT-P-5 b, HAT-P-6 b, HAT-P-7 b, HAT-P-8 b, HAT-P-9 b, HATS-1 b, HD 149026 b, HD 17156 b, HD 189733 b, HD 209458 b, HD 80606 b, HD 97658 b, KELT-2A b, KELT-3 b, KIC 10905746 b, KIC 12557548 b, KIC 4862625 b, KIC 6185331 b, KIC-8852719 b, KOI-13 b, KOI-135 b, KOI-196 b, KOI-200 b, KOI-202 b, KOI-204 b, KOI-206 b, KOI-254 b, KOI-423 b, KOI-428 b, KOI-500 b, KOI-500 c, KOI-500 d, KOI-500 e, KOI-500 f, KOI-680 b, KOI-730 b, KOI-730 c, KOI-730 d, KOI-730 e, KOI-872 b, KOI-94 b, KOI-94 c, KOI-94 d, KOI-94 e, Kepler-10 b, Kepler-10 c, Kepler-11 b, Kepler-11 c, Kepler-11 d, Kepler-11 e, Kepler-11 f, Kepler-11 g, Kepler-12 b, Kepler-14 b, Kepler-15 b, Kepler-16 (AB) b, Kepler-17 b, Kepler-18 b, Kepler-18 c, Kepler-18 d, Kepler-19 b, Kepler-20 b, Kepler-20 c, Kepler-20 d, Kepler-20 e,
MP2	Kepler-20 f, Kepler-21 b, Kepler-22 b, Kepler-23 b, Kepler-23 c, Kepler-24 b, Kepler-24 c, Kepler-25 b, Kepler-25 c, Kepler-26 b, Kepler-26 c, Kepler-27 b, Kepler-27 c, Kepler-28 b, Kepler-28 c, Kepler-29 b, Kepler-29 c, Kepler-30 b, Kepler-30 c, Kepler-30 d, Kepler-31 b, Kepler-31 c, Kepler-32 b, Kepler-32 c, Kepler-33 b, Kepler-33 c, Kepler-33 d, Kepler-33 e, Kepler-33 f, Kepler-34(AB) b, Kepler-35(AB) b, Kepler-36 b, Kepler-36 c, Kepler-38(AB) b, Kepler-4 b, Kepler-42 b, Kepler-42 c, Kepler-42 d, Kepler-47(AB) b, Kepler-47(AB) c, Kepler-48 b, Kepler-48 c, Kepler-49 b, Kepler-49 c, Kepler-5 b, Kepler-50 b, Kepler-50 c, Kepler-51 b, Kepler-51 c, Kepler-52 b, Kepler-52 c, Kepler-53 b, Kepler-53 c, Kepler-54 b, Kepler-54 c, Kepler-55 b, Kepler-55 c, Kepler-56 b, Kepler-56 c, Kepler-57 b, Kepler-57 c, Kepler-58 b, Kepler-58 c, Kepler-59 b, Kepler-59 c, Kepler-6 b, Kepler-60 b, Kepler-60 c, Kepler-60 d, Kepler-68 b, Kepler-68 c, Kepler-7 b, Kepler-8 b, Kepler-9 b, Kepler-9 c, Kepler-9 d, OGLE-TR-10 b, OGLE-TR-111 b, OGLE-TR-113 b, OGLE-TR-132 b, OGLE-TR-182 b, OGLE-TR-211 b, OGLE-TR-56 b, OGLE2-TR-L9 b, Qatar-1 b, Qatar-2 b, SWEEPS-04, SWEEPS-11, TrES-1, TrES-2, TrES-3, TrES-4, TrES-5, WASP-1 b, WASP-10 b, WASP-11-HAT-P-10 b, WASP-12 b, WASP-13 b, WASP-14 b, WASP-15 b, WASP-16 b, WASP-17 b, WASP-18 b, WASP-19 b, WASP-2 b, WASP-20 b, WASP-21 b, WASP-22 b, WASP-23 b, WASP-24 b, WASP-25 b, WASP-26 b, WASP-28 b, WASP-29 b, WASP-3 b, WASP-31 b, WASP-32 b, WASP-33 b, WASP-34 b, WASP-35 b, WASP-36 b, WASP-37 b, WASP-38 b, WASP-39 b, WASP-4 b, WASP-41 b, WASP-42 b, WASP-43 b, WASP-44 b, WASP-45 b, WASP-46 b, WASP-47 b, WASP-48 b, WASP-49 b, WASP-5 b, WASP-50 b, WASP-52 b, WASP-53 b, WASP-54 b, WASP-55 b, WASP-56 b, WASP-57 b, WASP-58 b, WASP-59 b, WASP-6 b, WASP-60 b, WASP-61 b, WASP-62 b, WASP-63 b, WASP-64 b, WASP-65 b, WASP-66 b, WASP-67 b, WASP-68 b, WASP-69 b, WASP-7 b, WASP-70 b, WASP-71 b, WASP-77A b, WASP-78 b, WASP-79 b, WASP-8 b, WASP-80 b, WTS-1 b, XO-1 b, XO-2 b, XO-3 b, XO-4 b, XO-5 b

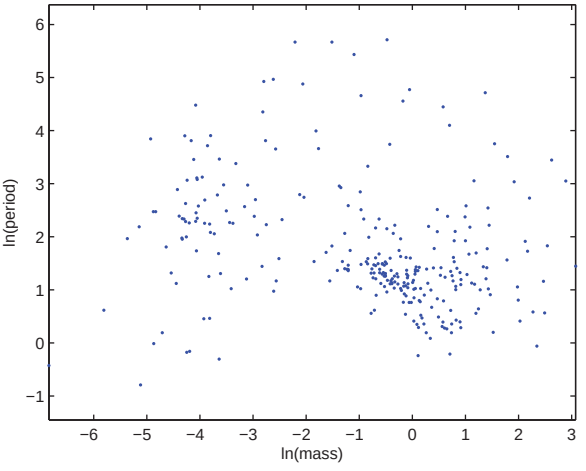

Figure 3. The scatter plot of the complete data in $\ln M - \ln P$ plane.

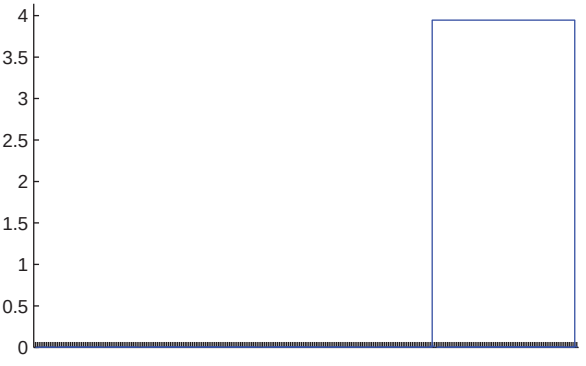

Figure 4. The dendrogram of exoplanets in the $\ln M - \ln P$ plane. There are too many data points, so the data identity numbers below the horizontal axis are not clear.

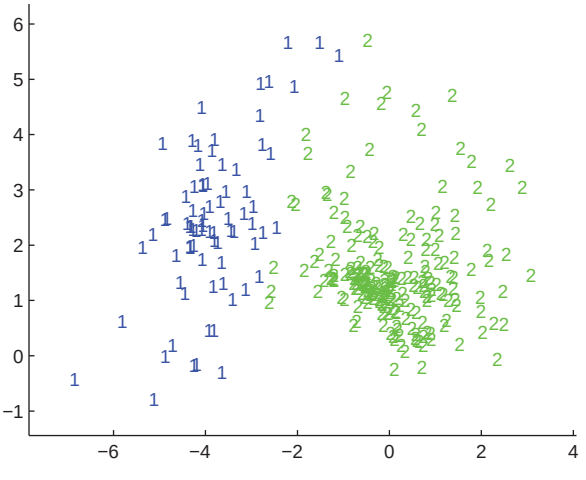

Figure 5. The symbol '1' indicates the members of Cluster MP1, the symbol '2' indicates the members of Cluster MP2.

extremely close orbits (below 3 d), such as, WASP-12 b, WASP-14 b, WASP-18 b, WASP-19 b and GJ 436 b.

The intracluster correlations give important information about the formation and evolution of exoplanets within a cluster. From Table 3, the intracluster correlations ρ_s for mass and period in Clusters MP1 and MP2 are 0.401 and -0.188, respectively at significance level 0.01. This implies that the mass and period correlate in an opposite way between two different clusters. It reflects differences in the formation and evolution processes of these two clusters.

The clustering analysis is implemented in Matlab program, and it takes 4.09 s in Notebook ASUS A43S with CPU Intel Core i5-2430M 2.4 GHz.

5. Conclusions

This paper employs the Frank copula to combine the Pareto marginal distributions, and use a Bayesian method to estimate the parameters of interest and to impute missing values of mass in the transiting exoplanet data.

After imputing the missing data, the SCM algorithm gives very good and reasonable clustering results. Two clusters are found in the logarithm mass and period of the transiting exoplanets illustrating differences in the formation and evolution processes of these two clusters according to the intracluster correlations ρ_s.

This paper considers only the Pareto distribution. It is interesting to investigate whether other marginal distributions provide better results, and whether other copula functions perform better. Finally, because the Metropolis–Hastings algorithm is computationally demanding, a more efficient algorithm for Bayesian inference should be developed.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive comments to improve the presentation of the paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work is supported in part by the National Science Council, Taiwan, under Huei-Wen Teng's Grants [NSC 100-2118-M-008-002] and Wen-Liang Hung's Grants [NSC 100-2118-M-134-001].

References

[1] E.L. Boone, K. Ye, and E.P. Smith, *Using data augmentation via the Gibbs sampler to incorporate missing covariate structure in linear models for ecological assessments*, Environ. Ecol. Statist. 16 (2009), pp. 75–87.
[2] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. R. Stat. Soc. Ser. B 39 (1977), pp. 1–38.
[3] A. Gelman and D.B. Rubin, *Inference from iterative simulation using multiple sequences*, Statist. Sci. 7 (1992), pp. 457–472.
[4] C. Genest, *Frank's family of bivariate distributions*, Biometrika 74 (1987), pp. 549–555.
[5] W.R. Gilksm, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Chapman & Hall, New York, 1996.
[6] P.C. Gregory, *A Bayesian analysis of extrasolar planet data for HD 73526*, Astrophys. J. 631 (2005), pp. 1198–1214.

[7] P.C. Gregory, *A Bayesian Kepler periodogram detects a second planet in HD 208487*, Mon. Not. R. Astron. Soc. 374 (2007), pp. 1321–1333.
[8] P.C. Gregory, *A Bayesian periodogram finds evidence for three planets in HD 11964*, Mon. Not. R. Astron. Soc. 381 (2007), pp. 1607–1616.
[9] P.C. Gregory, *Bayesian exoplanet tests of a new method for MCMC sampling in highly correlated model parameter spaces*, Mon. Not. R. Astron. Soc. 410 (2011), pp. 94–110.
[10] P.C. Gregory and D.A. Fischer, *A Bayesian periodogram finds evidence for three planets in 47 Ursae Majoris*, Mon. Not. R. Astron. Soc. 403 (2010), pp. 731–747.
[11] I.G. Jiang, L.C. Yeh, Y.C. Chang, and W.L. Hung, *On the mass-period distributions and correlations of extrasolar planets*, Astron. J. 134 (2007), pp. 329–336.
[12] I.G. Jiang, L.C. Yeh, Y.C. Chang, and W.L. Hung, *Construction of coupled period-mass functions in extrasolar planets through a nonparametric approach*, Astron. J. 137 (2009), pp. 329–336.
[13] I.G. Jiang, L.C. Yeh, W.L. Hung, and M.S. Yang, *Data analysis on the extrasolar planets using robust clustering*, Mon. Not. R. Astron. Soc. 370 (2006), pp. 1379–1392.
[14] E. Käärik, *Modeling dropouts by conditional distribution, a copula based approach*, Metodološki zvezki 3 (2006), pp. 109–120.
[15] R.J.A. Little and D.B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed., John Wiley, New York, 2002.
[16] L.S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag, New York, 2001.
[17] S.G. Meester and J. MacKay, *A parametric model for cluster correlated categorical data*, Biometrics 50 (1994), pp. 954–963.
[18] R.B. Nelsen, *An Introduction to Copulas*, Springer, New York, 1999.
[19] W. Shao, G. Guo, F. Meng, and S. Jia, *An efficient proposal distribution for Metropolis–Hastings using a B-splines technique*, Comput. Statist. Data Anal. 57 (2013), pp. 465–478.
[20] A. Sklar, *Fonctions de répartition à n dimensions et leurs marges*, Publ. Inst. Statist. Univ. Paris 8 (1959), pp. 229–231.
[21] S. Tabachnik and S. Tremaine, *Maximum-likelihood method for estimating the mass and period distributions of extrasolar planets*, Mon. Not. R. Astron. Soc. 335 (2002), pp. 151–158.
[22] I. Wasito and B. Mirkin, *Nearest neighbours in least-squares data imputation algorithms with different missing patterns*, Comput. Statist. Data Anal. 50 (2006), pp. 926–949.
[23] M.S. Yang and K.L. Wu, *A similarity-based robust clustering method*, IEEE Trans. Pattern Anal. Mach. Intell. 26 (2004), pp. 434–448.