Mon. Not. R. Astron. Soc. **377,** 1281–1286 (2007) doi:10.1111/j.1365-2966.2007.11685.x

Multivariate comparisons of the period–light-curve shape distributions of Cepheids in five galaxies

C. Koen[★] and I. Siluyele

Department of Statistics, University of the Western Cape, Private Bag X17, Bellville, 7535 Cape, South Africa

Accepted 2007 February 28. Received 2007 February 28; in original form 2007 February 8

ABSTRACT

A number of published tests suitable for the comparison of multivariate distributions are described. The results of a small power study, based on realistic Cepheid log period – Fourier coefficient data, are presented. It is found that a statistic due to Henze has good general performance. The tests are applied to Cepheid observations in the Milky Way galaxy, Large Magellanic Cloud, Small Magellanic Cloud, IC 1613 and NGC 6822. The null hypothesis of equal populations is rejected for all pairs compared, except IC 1613 – NGC 6822.

Key words: methods: statistical – Cepheids – galaxies: stellar content.

1 INTRODUCTION

Following the pioneering work by Simon & Clement (1993), a number of authors (e.g. Buchler & Moskalik 1994; Antonello 2006) have compared the joint (i.e. multivariate) distributions of pulsation periods P and light-curve shapes (as measured by Fourier coefficients) of Cepheids in different galaxies. Commonly used dimensionless Fourier coefficients are

$$R_{j+1,j} = A_{j+1}/A_j \quad \text{and} \quad \phi_{j+1,j} = \phi_{j+1} - 2\phi_j, \tag{1}$$

where A_j is the amplitude, and ϕ_j the phase, of the jth term in the fit of a Fourier series to the light-curve shape. The intergalaxy comparisons have relied primarily on visual inspection of two-dimensional plots of various combinations of $\log P$, R_{21} and ϕ_{21}. In this paper more rigorous statistical procedures for the problem are introduced.

Formally the comparison of multidimensional scatterplots can be viewed as the comparison of samples drawn from multivariate distributions, with the aim of testing whether the two underlying populations are statistically the same. If no distributional assumptions are made, then in the case of univariate data, the χ^2 and Kolmogorov–Smirnov (KS) tests for equality of populations are very well known. For higher dimensional data there are a number of commonly used tests subject to assumptions about the family membership (usually Gaussian) of the populations. Otherwise, if no distributional assumptions are made, extensions of the KS statistic to higher dimensions have been discussed by a number of authors, both in the astronomy (e.g. Peacock 1983), and statistics (Justel, Peña & Zamar 1997) literature. However, this is by no means the only non-parametric statistic for the comparison of two multivariate samples. A number of other statistics will be described in the next section of the paper, and a limited simulation study of some of their properties will be presented in Section 3.

Fairly extensive observations of the Cepheids in the Milky Way galaxy and the two Magellanic Clouds are available. The recent compilation of periods and V- and I-band Fourier coefficients by Ngeow & Kanbur (2006) are used here. Details of the data selection criteria can be found in Kanbur & Ngeow (2004, 2006); relevant to this paper is the fact that $0.4 < \log P < 1.7$. Data for one star (AV Cir) are excluded: this object, which has the shortest period of the 154 MW Cepheids, appears to have anomalous Fourier parameters. The numbers of Cepheids are then 153, 390 and 641, for the MW, Small Magellanic Cloud (SMC) and Large Magellanic Cloud (LMC), respectively. To these data we add the ($\log P$, R_{21}, ϕ_{21}) compilations given by Antonello (2006) for two local group dwarf irregulars IC 1613 ($N = 18, 16$ for the V and I bands, respectively) and NGC 6822 ($N = 53, 50$ for V and I bands, respectively). The results of two-sample tests performed on pairs of these data are given in Section 4.

Conclusions are presented in Section 5.

2 THE TEST STATISTICS

The two samples are denoted by $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ and $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n\}$, where each of the $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ is in general D dimensional. The notation $x_{ik}(y_{jk})$ will be used for the kth component ($k = 1, 2, \ldots, D$) of $\boldsymbol{x}_i(\boldsymbol{y}_j)$. Pooling the two samples gives

$$\{z_1, z_2, \ldots, z_{m+n}\} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\} \cup \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n\}. \tag{2}$$

The first statistic is a multivariate extension of the one-dimensional KS statistic; the second is an extension of the Wald–Wolfowitz runs test (e.g. Conover 1971); while the remainder are all based on distributions of interpoint distances.

2.1 The multivariate KS statistic

This is a generalization of the standard KS statistic used to compare two univariate samples. The key ingredient is the maximum

[★] E-mail: ckoen@uwc.ac.za

difference between the estimated cumulative distribution functions (EDFs) of the two samples. The standard definition of the EDF for D-dimensional data $\boldsymbol{x}_j$ is

$$F_{\mathrm{x}}(\boldsymbol{u}) = F_{\mathrm{x}}(u_1, u_2, \ldots, u_D) = \frac{1}{N}\sum_{j=1}^{N} I(\boldsymbol{x}_j; \boldsymbol{u}) \tag{3}$$

where the indicator function is defined by

$$I(\boldsymbol{x}_j; \boldsymbol{u}) = \begin{cases} 1 & x_{jk} \leqslant u_k \; (k = 1, 2, \ldots, D) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

The full multivariate KS statistic is computationally expensive, hence the simplified form

$$\mathrm{SKS} = \frac{mn}{m+n} \max_k |F_{\mathrm{x}}(z_k) - F_{\mathrm{y}}(z_k)| \qquad k = 1, 2, \ldots, m+n \tag{5}$$

(with z defined in equation 2) is used below – see remark (6) in Section 2.7.

2.2 The multivariate runs test of Friedman & Rafsky (1979)

The Wald–Wolfowitz runs test can be used to compare two univariate samples. The test consists of ordering the two combined data sets and then counting the number of runs (uninterrupted sequences of x or y values). Friedman & Rafsky (1979) devised a multivariate analogue to the univariate runs test. The minimal spanning tree of the combined data serves to order the data. All connecting lines ('edges') between points from unlike samples are then removed, leaving the multivariate 'runs'. The number R of disjoint runs is the test statistic: if the null hypothesis is true the two data sets will be well mixed in multidimensional space, and R will be large. The null hypothesis is therefore rejected for sufficiently small values of R.

2.3 The Baringhaus & Franz (2001) statistic

The test proceeds from the general observation that

$$\mathrm{E}|\boldsymbol{x}_i - \boldsymbol{y}_j| - \frac{1}{2}[\mathrm{E}|\boldsymbol{x}_i - \boldsymbol{x}_j| + \mathrm{E}|\boldsymbol{y}_i - \boldsymbol{y}_j|] \geqslant 0 \tag{6}$$

where E is the expectation operator, and $|\boldsymbol{u} - \boldsymbol{v}|$ is the Euclidean distance between $\boldsymbol{u}$ and $\boldsymbol{v}$. The equality in equation (6) applies if and only if the two samples are from the same population. Replacing population means by sample means, the test statistic

$$\mathrm{BF} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}|\boldsymbol{x}_i - \boldsymbol{y}_j| - \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}|\boldsymbol{x}_i - \boldsymbol{x}_j| - \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|\boldsymbol{y}_i - \boldsymbol{y}_j| \tag{7}$$

is expected to be small if the null hypothesis is true.

2.4 The Henze (1988) statistic

For each $\boldsymbol{x}_i$ its K (with K typically 1 or 2) nearest neighbours in multidimensional space are determined. The number of neighbours $N_x(i)$ which come from the same sample (i.e. the collection of $\boldsymbol{x}_j$ rather than $\boldsymbol{y}_j$) are then counted. The process is repeated with the members of the other sample, giving also $N_y(j)$. The statistic is

$$H(K) = \sum_{i=1}^{m} N_x(i) + \sum_{j=1}^{n} N_y(j). \tag{8}$$

Larger $H(K)$ are expected if the null hypothesis is false, since points from like samples are then expected to cluster.

2.5 The Hall & Tajvidi (2002) statistics

The calculation of this interpoint-distance-based statistic is as follows.

(i) For $\boldsymbol{x}_i$ from the first sample, calculate the $m - 1$ distances

$|\boldsymbol{x}_i - \boldsymbol{x}_k| \quad k = 1, 2, \ldots, i-1, i+1, \ldots, m$

and the n distances

$|\boldsymbol{x}_i - \boldsymbol{y}_k| \quad k = 1, 2, \ldots, n.$

(ii) Order the $n + m - 1$ interpoint distances calculated in (i). Let $M_i(j)$ be the number of $\boldsymbol{y}_k$ amongst the j nearest neighbours of $\boldsymbol{x}_i$. Under the null hypothesis the expected value of $M_i(j)$ is $nj/(m + n - 1)$, that is,

$$DM_i(j) = |M_i(j)\text{-}n_j/(m + n\text{-}1)| \tag{9}$$

is expected to be large if the samples are from different populations.

(iii) Interchanging the roles of the two samples in (i) and (ii) leads also to the measure

$$DN_\ell(j) = |N_\ell(j) - mj/(m + n - 1)| \tag{10}$$

based on $N_\ell(j)$, the number of $\boldsymbol{x}_k$ amongst the j nearest neighbours of $\boldsymbol{y}_\ell$.

(iv) The values of $DM_i(j)$ (i = 1, 2, … , m) and $DN_\ell(j)$ (ℓ = 1, 2, … , n) can then be combined to give a single test statistic. Hall & Tajvidi (2002) presented the two forms

$$\text{HT-T} = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n} DM_i(j) + \frac{1}{n}\sum_{\ell=1}^{n}\sum_{j=1}^{m} DN_\ell(j)$$

$$\text{HT-S} = \sum_{j=1}^{n}\max_i DM_i(j) + \sum_{j=1}^{m}\max_\ell DN_\ell(j). \tag{11}$$

[The formulae given by Hall & Tajvidi (2002) allow for weighting of the terms $DM_i(j)$ and $DN_\ell(j)$, but this will not be pursued here.]

Clearly, the statistic is related to that of Henze (1988).

2.6 A statistic based on *all* interpoint distances

Maa, Pearl & Bartoszynski (1996) showed that the distributions of the three sets of interpoint distances

$$\begin{aligned} &|\boldsymbol{x}_i - \boldsymbol{x}_j| \quad (i = 1, 2, \ldots, m-1; j = i+1, \ldots, m) \\ &|\boldsymbol{y}_i - \boldsymbol{y}_j| \quad (i = 1, 2, \ldots, n-1; j = i+1, \ldots, n) \\ &|\boldsymbol{x}_i - \boldsymbol{y}_j| \quad (i = 1, 2, \ldots, m; j = 1, \ldots, n) \end{aligned} \tag{12}$$

are all identical if, and only if, the $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ are drawn from the same two populations. Equivalence of the three distributions of distances can be tested by using three-sample univariate tests: see point (5) in Section 2.7. The acronym IPDD ('interpoint distance distribution') will be used for this statistic.

2.7 A few remarks

(1) The exact distributions of all of these statistics are unknown. In most cases this does not matter as significance levels can be determined by permutation. The recipe used is:

(i) Calculate the statistic of interest.

(ii) Combine the two samples; then randomly divide the pooled data into two samples of sizes m and n.

(iii) Calculate the statistic of interest for the two new samples.

(iv) Repeat steps (ii)–(iv) many (at least a few hundred) times, noting the value of the statistic for each repetition.

(v) The significance level (p-value) of the statistic in step (i) is determined by its ranking with respect to the values obtained in steps (ii)–(iv).

The rationale is the following: if the two original samples are indeed from the same distribution, then the value of the statistic will be unremarkable. This will be revealed if it is compared to values obtained from artificial samples generated by randomly re-assigning observations to the two samples – which is perfectly legitimate under the null hypothesis of equal populations. However, if the two samples are from different populations, then the randomized samples will differ fundamentally from the original two samples: the latter will reflect purely the distinct populations, while the former will consist of a mixture of values from the two populations. Therefore, the statistic calculated in step (i) is expected to have a value which is radically different from that calculated for the mixed samples in steps (ii)–(iv).

(2) Large-sample results are available for some of the statistics. In the case of the Friedman & Rafsky (1979) statistic R, the standardized form

$$\begin{aligned} FR &= [(R-1)(m+n)-2mn]\sqrt{\frac{m+n-1}{2mn(m+n)Q}} \\ Q &= \frac{2mn}{m+n}+\frac{[C-(m+n)+2]}{(m+n-2)(m+n-3)} \\ &\quad \times [(m+n)(m+n-1)-4mn+2]-1 \\ C &= \frac{1}{2}\left\{\sum_{i=1}^{m} c(\boldsymbol{x}_i)[c(\boldsymbol{x}_i)-1]+\sum_{j=1}^{n} c(\boldsymbol{y}_j)[c(\boldsymbol{y}_j)-1]\right\} \end{aligned} \qquad (13)$$

has an asymptotic standard normal distribution. The notation $c(z)$ indicates the number of points in the minimal spanning tree which are connected to z. The asymptotic distribution of FR is used below because determination of the minimum spanning tree is currently computationally too expensive for randomization to be viable.

(3) The KS statistic depends only on the ordering of the data, whereas all five the other statistics rely in some way on interpoint distances. The KS statistic is therefore impervious to the units in which the components of the vectors $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ are measured, whereas the same will not be true for the other statistics. Standardization such as

$$x'_{ij}=\frac{x_{ij}}{S_j} \quad y'_{kj}=\frac{y_{kj}}{S_j}$$
$$i=1,2,\ldots,m \quad k=1,2,\ldots,n \quad j=1,2,\ldots,D \qquad (14)$$

where S_j is some measure of the scale of the jth component of $\boldsymbol{x}$ and $\boldsymbol{y}$, for example, the s.d. of $x_{1j}, x_{2j}, \ldots, x_{mj}, y_{1j}, \ldots, y_{nj}$, or

$$S_j=\max_{k,\ell}(x_{kj}, y_{\ell j})-\min_{k,\ell}(x_{kj}, y_{\ell j}). \qquad (15)$$

is therefore used.

(4) A variation on the theme of the IPDD statistic is to replace the full IPDDs by the distributions of distances to nearest neighbours only, for example, the $m(m-1)/2$ distances $|\boldsymbol{x}_i-\boldsymbol{x}_j|$ in equation (12) are replaced by the m distances

$$\min_{j\neq i}|\boldsymbol{x}_i-\boldsymbol{x}_j| \quad i=1,2,\ldots,m. \qquad (16)$$

The $n(n-1)/2$ distances in the second equation in equation (12) are reduced to n, and the mn distances in the third equation are reduced to $m+n$. Although less information is used the number of calculations and computer memory requirements are substantially reduced. The result statistic will be denoted ‘NNDD’ (‘nearest neighbour distance distribution’) in what follows.

(5) A number of statistics for testing for the statistical equivalence of three cumulative distribution functions (i.e. extensions of the classical univariate KS statistic to the three-sample case) are available. Two of these, described in Fisz (1963, p. 408f), were evaluated as part of the power studies reported below, and found to have closely similar performance. Results are therefore only quoted for the simpler of the two (see also Kiefer 1959):

$$D^2=\max_u \sum_{j=1}^{3} N_j[S_j(u)-\overline{S}(u)]^2, \qquad (17)$$

where S_j is the EDF, and N_j the size, of sample j ($j=1, 2, 3$). The statistic $\overline{S}$ is the EDF of the three pooled samples.

(6) Two points of considerable relevance to the KS statistic, both of which have been dealt with in the literature, are (i) the computational expense of a complete evaluation of the multivariate EDF and (ii) the effect of the non-uniqueness of the multivariate cumulative distribution function. In order to keep the discussion which follows as transparent as possible $D=2$ is assumed; the essence is unchanged for larger D.

(i) The EDF defined in equation (3) is a step function, with changes not only in the observed data points (x_{j1}, x_{j2}), but also in some $(x_{j1}, x_{\ell 2})$ with $j\neq \ell$. The implication is that F needs to be evaluated in all $(x_{j1}, x_{\ell 2})$ $(j,\ \ell=1, 2, \ldots, N)$ – see Gosset (1987), fig. A1. A consequence is that evaluation of F is computationally very expensive. It has been suggested in both the astronomy and statistics literature that the calculations be restricted to the observed data points only (i.e. the case $j=\ell$) – see Fasano & Franceschini (1987) and Justel et al. (1997). Further justification for the simplified statistic in equation (5) can be found in Greenberg (2006), where it is shown that the power of two-sample tests based on the full and simplified forms of F are very similar.

(ii) For univariate x the usual definition of the CDF is

$$F(u)=Pr(x\leqslant u). \qquad (18)$$

In principle, an alternative is to define

$$F(u)=Pr(x\geqslant u), \qquad (19)$$

but since

$$Pr(x\leqslant u)+Pr(x>u)=1 \qquad (20)$$

the second definition (19) is perfectly equivalent to (18). In the bivariate case $\boldsymbol{x}=(x_1, x_2)$, equation (20) is replaced by

$$\begin{aligned} &Pr(x_1\leqslant u_1, x_2\leqslant u_2)+Pr(x_1\leqslant u_1, x_2>u_2) \\ &\quad +Pr(x_1>u_1, x_2\leqslant u_2)+Pr(x_1>u_1, x_2>u_2)=1 \end{aligned}$$

so that the four definitions

$$\begin{aligned} F(u_1, u_2) &= Pr(x_1\leqslant u_1, x_2\leqslant u_2) \\ F(u_1, u_2) &= Pr(x_1\leqslant u_1, x_2\geqslant u_2) \\ F(u_1, u_2) &= Pr(x_1\geqslant u_1, x_2\leqslant u_2) \\ F(u_1, u_2) &= Pr(x_1\geqslant u_1, x_2\geqslant u_2) \end{aligned}$$

are interrelated, but *not* equivalent. Put differently, each of the four CDFs captures information not fully contained in the other three. For this reason Peacock (1983) suggested the incorporation of all four definitions into the multidimensional KS statistic.

The expense of calculating the KS statistic in this fashion increases approximately as D^2 with increasing dimensionality of the problem; only the form based on the definition in equation (4) is therefore used here.

(7) In the study reported below, and in other related work, it was found that the power (i.e. ability to reject the null hypothesis when it is false) of the Henze (1988) statistic improved steadily as the number of nearest neighbours K was increased from one to four. Little improvement for $K > 4$ was seen. Results will therefore be given only for $K = 1$ and 4; the statistics are denoted H1 and H4, respectively.

3 A POWER STUDY

The data derived from the V-band observations of SMC Cepheids (table 5 in Ngeow & Kanbur 2006) were used as a 'population'. One outlying data point was excluded, leaving 389 trivariate ($\log P$, ϕ_{21}, R_{21}) points – see Figs 1–3. The power experiments proceeded as follows.

(i) Two samples of size 70 were drawn without replacement (i.e. each datum in the samples was unique) from the population.

(ii) The multivariate mean of one of the samples was changed by adding a fixed number to one of the components of the trivariate observations.

(iii) The statistics discussed in Section 2 were calculated.

(iv) Significance levels of the statistics were evaluated by randomization, as described in Section 2.7, point (1). Five hundred randomizations were used throughout.

(v) Steps (i)–(iv) were repeated 300 times, and the fraction of rejections (at the 5 per cent level) of the null hypothesis were noted. This completes one power study.

(vi) Since the statistics depend on the specifics of the univariate distributions of each of the three components ($\log P$, R_{21} and ϕ_{21}), the scheme (i)–(v) was repeated three times, each time for a different component in step (ii).

(vii) Steps (i)–(vi) shed light on the relative performance of each statistic in detecting changes in mean values. All the steps were repeated, substituting a scale change in one of the coordinates for the mean change in step (ii).

(viii) Power against a covariance change was also studied: the dependence structure of one sample was changed by adding the points shown as squares in Figs 1–3 to it.

The specific mean shift used in step (ii) of the outline above was 5 per cent of the range of the population of 389 values (of the specific component). The scale change was

$$x_{ik} = \mu_k + 1.5(u_{ik} - \mu_k),$$

where $\mu_k (k = 1, 2 \text{ or } 3)$ is the population mean value of the coordinate, and the $u_{ik} (i = 1, 2, \ldots, 70)$ are the members of one of the two samples drawn in step (i).

The results are reported in Table 1. As explained above, the mean shifts or scale changes were applied separately to each of the three components ($\log P$, R_{21} and ϕ_{21}); the outcomes are given in the first six lines of the table. The results of comparing two samples with different covariances are in line 7.

Inspection of Figs 2 and 3 shows two isolated data elements (plotted as stars). In order to evaluate their influence, the mean shift and scale change experiments were repeated with these two points excluded. The results are in the last six lines of Table 1.

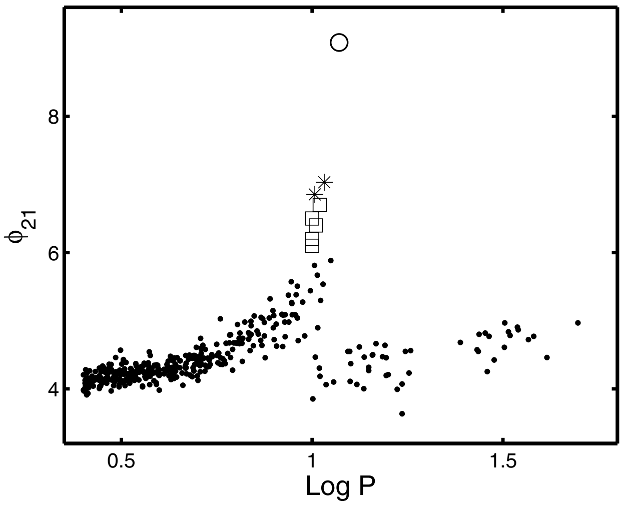

Figure 1. Periods and Fourier phase parameters for the SMC data (all symbols except squares). The outlying point shown by the open circle was excluded in the power studies. The two points shown by stars were also excluded in some of the studies, in order to evaluate the sensitivity of the results to their presence. Artificial points shown by squares were added in one power study in order to evaluate sensitivity to a small covariance change.

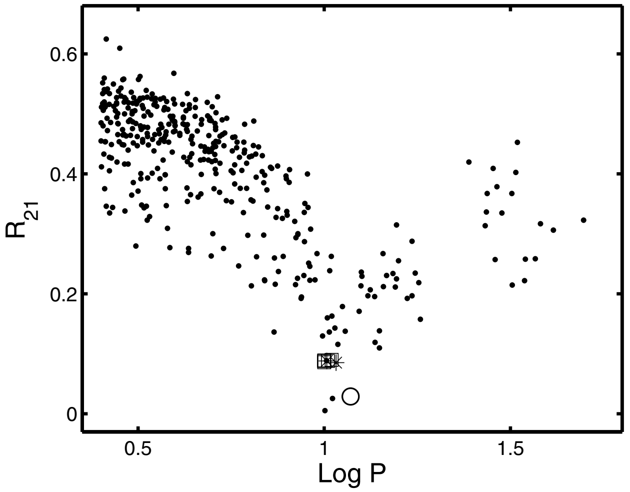

Figure 2. As for Fig. 1, but showing the distributions of periods and Fourier amplitude components.

Study of Table 1 leads to the following conclusions:

(i) The H statistic is the most powerful against shifts in the mean of the first ($\log P$) and third (ϕ_{21}) coordinates; the FR statistic is almost as good at detecting shifts in the ϕ_{21} distribution. Changes in the mean of the second coordinate (R_{21}) are best detected by the SKS and HT-S statistics; H4 is next best.

(ii) The H statistic is the most powerful detector of scale changes in all three coordinates.

(iii) None of the statistics has high rejection rates when the dependence structures of the two samples are subtly different; the H and FR statistics fared slightly better than the rest.

(iv) From (i)–(iii) it follows that, for general alternatives to the null hypothesis, H4 is the statistic of choice.

(v) It is interesting that the SKS statistic traditionally used by astronomers generally falls far short of the H statistics, and also

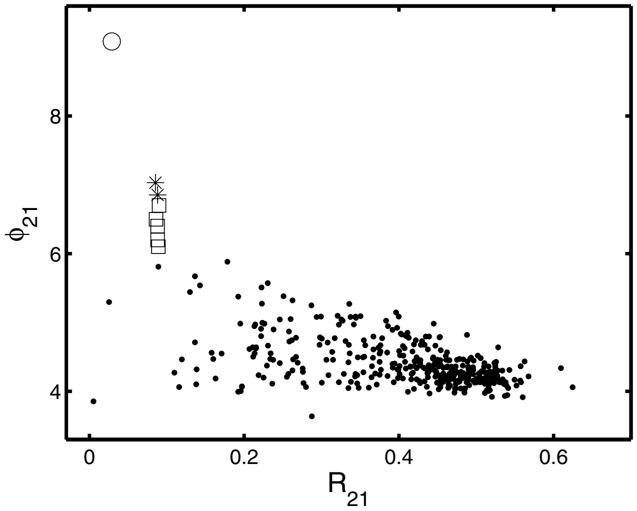

Figure 3. As for Fig. 1, but showing the distributions of Fourier amplitude and phase components.

Table 1. Results of limited power studies: fraction of rejections of the null hypothesis for various true alternatives, for each of nine statistics. One parameter, given in the first column of each line, differs for the two samples. The last six lines report on samples drawn from a slightly different population (obtained by excluding the two starred data points in Figs 1–3). See the text for further details.

Difference	BF	SKS	H1	H4	NNDD	IPDD	HT-T	HT-S	FR
E($\log P$)	0.18	0.22	0.36	0.49	0.15	0.04	0.14	0.20	0.41
E(R_{21})	0.25	0.47	0.26	0.38	0.11	0.06	0.21	0.46	0.30
E(ϕ_{21})	0.72	0.56	0.88	0.97	0.58	0.07	0.49	0.84	0.92
var($\log P$)	0.14	0.47	0.69	0.84	0.41	0.11	0.24	0.37	0.71
var(R_{21})	0.11	0.46	0.46	0.68	0.20	0.13	0.26	0.50	0.52
var(ϕ_{21})	0.14	0.15	0.44	0.54	0.24	0.13	0.15	0.26	0.44
covariance	0.05	0.08	0.09	0.13	0.05	0.07	0.05	0.06	0.11
E($\log P$)	0.15	0.22	0.42	0.52	0.14	0.04	0.12	0.19	0.44
E(R_{21})	0.23	0.46	0.26	0.44	0.12	0.05	0.19	0.42	0.32
E(ϕ_{21})	0.31	0.31	0.40	0.57	0.16	0.07	0.21	0.44	0.42
var($\log P$)	0.11	0.45	0.69	0.74	0.34	0.12	0.15	0.30	0.68
var(R_{21})	0.16	0.48	0.47	0.68	0.21	0.15	0.28	0.52	0.52
var(ϕ_{21})	0.10	0.19	0.40	0.53	0.18	0.14	0.19	0.33	0.40

fares worse than the FR statistic (excepting the case of a change in the mean of R_{21}).

(vi) Generally

power(*IPDD*) < power(HT-T) < power(HT-S) < power(FR)

and

power(BF) < power(SKS)

hold. The simpler NNDD consistently performs better than the IPDD.

(vii) Excluding the two data points marked by stars in the figures did not lead to changes in the performance rankings of the different statistics, except in the case of mean changes in ϕ_{21}. The power against scale changes is generally not affected much, although the power of the best statistic (H4) for discerning changes in the variance of $\log P$ is somewhat reduced. The power against mean shifts in ϕ_{21} is substantially reduced: this means that isolated data points could have a substantial influence on the outcome of hypothesis tests.

(viii) Study of the figures shows that the two starred points have extreme values of ϕ_{21}, are amongst the five smallest R_{21}, but have pedestrian values of $\log P$. The results listed in (viii) can be ascribed to these properties.

4 APPLICATION TO THE FOURIER COEFFICIENT DATA

An intercomparison of the Cepheid data for the MW, LMC and SMC Cepheids found the difference between any pair to be significant with $p < 0.001$, according to *all* tests. This pronouncement is based on the result of 1000 permutations – in no case was the statistic calculated from the permutation samples more extreme than the value calculated from the observed data. The result applies to both *V*- and *I*-band data.

By contrast no significant differences were found between the IC 1613 and NGC 6822 data sets: the smallest level attained over all tests was 0.10 for the *V* data, and 0.58 for the *I* data (based on 5000 permutations, as are the rest of the results reported below).

The significance levels of statistics comparing the NGC 6822 data with, respectively, the MW, LMC and SMC measurements were almost all below 1 per cent. Only two exceptions were encountered, both for comparison of NGC 6822 data with MW data (*V* band): significance levels of 2 and 14 per cent were obtained for the HT-T and IPDD statistics, respectively.

Highly significant ($p <$ 1 per cent) differences were found between the IC 1613 and LMC data (both *I* and *V*). Comparisons of the IC 1613 data with those for the MW and SMC gave more varied results – see Table 2 – but overall it appears that the null hypothesis of equal populations can safely be rejected. The reader's attention is drawn to some apparently discrepant results in the table, such as the 18 per cent levels for the IPDD and HT-T statistics (MW *I*-band comparison) and the 84 per cent level for the FR statistic (SMC *V*-band comparison).

Finally, as a check, the MW *V*-band data were compared to an earlier compilation by Antonello & Morelli (1996) of data for Galactic Cepheids with periods longer than 8 d. In order that the two data sets be compatible, the same restriction was imposed on the Ngeow & Kanbur (2006) data. Significance levels for all statistics except SKS ($p = 0.83$) and NNDD ($p = 0.67$) were larger than 0.90.

Table 2. The significance levels of the various two-sample statistics, calculated for comparisons of data for IC 1613 with the MW and SMC data, respectively.

Data set	BF	SKS	H1	H4	NNDD	IPDD	HT-T	HT-S	FR
MW (*V*)	0.03	0.04	0	0	0	0.03	0.10	0.007	0
MW (*I*)	0.03	0.03	0.04	0.002	0.005	0.18	0.18	0.04	0.001
SMC (*V*)	0	0	0.86	0.02	0	0.002	0	0	0.84
SMC (*I*)	0	0.003	0.004	0	0	0.002	0.001	0	0.009

5 CONCLUSIONS

The results of the power study in Section 3 suggest that the H4 statistic is generally particularly useful. It is also simple to program, and computationally fast. The FR statistic also has good power; since there is downloadable software for calculating minimal spanning trees available on the internet, it is also easy to program. Calculation of the FR statistic is computationally expensive, but preliminary studies suggest that the asymptotic results (point 2) in Section 2.7) may be used for fairly small ($N < 100$) samples. The SKS statistic outperforms all other statistics for some data sets. It is not difficult to program, and is fast. The IPDD statistic is demanding of computer resources, and has poor power. All the other statistics have intermediate performance.

Intercomparison of the trivariate ($\log P$, ϕ_{21}, R_{21}) data for the various galaxies leads to the rejection of the null hypothesis of equal populations in all cases except IC 1613 compared with NGC 6822. These results are based on formal statistical procedures, in contrast with previous studies in which qualitative assessments were made.

Finally, we note that although the analysis in the paper has been restricted to three-dimensional variables, the extension to higher dimensions (by the addition of higher order Fourier coefficients) is trivial.

ACKNOWLEDGMENTS

CK is grateful to Prof Fred Lombard (University of Johannesburg) for many discussions of this statistical problem, and to Dr Mohammed Abdel-Sabour Abdel-Latif (NRIAG, Egypt) for first bringing this problem to his attention. IS received financial support from the African Institute of Mathematical Sciences (Muizenberg, South Africa).

REFERENCES

Antonello E., 2006, A&A, 449, 569
Antonello E., Morelli P. L., 1996, A&A, 314, 541
Baringhaus L., Franz C., 2001, J. Multivariate Anal., 88, 190
Buchler J. R., Moskalik P., 1994, A&A, 292, 450
Conover W. J., 1971, Practical Nonparametric Statistics. Wiley, New York
Fasano G., Franceschini A., 1987, A&A, 225, 155
Fisz M., 1963, Probability Theory and Mathematical Statistics, 3rd edn. Robert E. Krieger Publishing Company, Malabar (Florida)
Friedman J. H., Rafsky L. C., 1979, Ann. Stat., 7, 697
Gosset E., 1987, A&A, 188, 258
Greenberg S., 2006, MSc thesis, Univ. Johannesburg
Hall P., Tajvidi N., 2002, Biometrika, 89, 359
Henze N., 1988, Ann. Stat., 16, 772
Justel A., Peña D., Zamar R., 1997, Stat. Probab. Lett., 35, 251
Kanbur S., Ngeow C.-C., 2004, MNRAS, 350, 962
Kanbur S., Ngeow C.-C., 2006, MNRAS, 369, 705
Kiefer J., 1959, Ann. Math. Stat., 30, 420
Maa J.-F., Pearl D. K., Bartoszynski R., 1996, Ann. Stat., 24, 1069
Ngeow C.-C., Kanbur S., 2006, MNRAS, 369, 723
Peacock J. A., 1983, MNRAS, 202, 615
Simon N. R., Clement C. M., 1993, ApJ, 419, L21

This paper has been typeset from a TEX/LATEX file prepared by the author.