Investigations of the Nonlinear LMC Cepheid Period-Luminosity Relation with Testimator and Schwarz Information Criterion Methods

S. M. Kanbur  
Department of Physics, State University of New York at Oswego, Oswego, NY; kanbur@oswego.edu  
C. Ngeow  
Department of Astronomy, University of Illinois, Urbana, IL  
AND  
A. Nanthakumar and R. Stevens  
Department of Mathematics, State University of New York at Oswego, Oswego, NY  

ABSTRACT. In this paper, we investigate the linearity versus nonlinearity of the Large Magellanic Cloud (LMC) Cepheid period-luminosity (P-L) relation using two statistical approaches not previously applied to this problem: the testimator method and the Schwarz Information Criterion (SIC). The testimator method is extended to multiple stages for the first time and is shown to be unbiased, and the variance of the estimated slope can be proved to be smaller than the standard slope estimated from linear regression theory. The Schwarz Information Criterion (also known as the Bayesian Information Criterion) is more conservative than the Akaike Information Criterion and tends to prefer lower order models. By using simulated data sets, we verify that these statistical techniques can be used to detect intrinsically linear and nonlinear P-L relations. These methods are then applied to independent LMC Cepheid data sets from the OGLE and MACHO projects. Our results imply that there is a change of slope in longer period ranges for all of the data sets. This strongly supports previous results obtained from independent statistical tests, which show that the observed LMC P-L relation is nonlinear, with a break period at or around 10 days.

1. INTRODUCTION

The cornerstone of the extragalactic distance scale is the Cepheid period-luminosity (P-L) relation defined by the Large Magellanic Cloud (LMC) Cepheids. The assumed linear relation of the LMC Cepheid P-L relation, which is linear in log P, where P is the pulsation period in days, has been under debate due to recent results showing that this relation could be nonlinear (Tammann & Reindl 2002; Kanbur & Ngeow 2004; Sandage et al. 2004; Ngeow et al. 2005). These authors contend that the existing Cepheid data in the LMC strongly suggest that the LMC P-L relation is consistent with two lines of significantly differing slopes, with a break at or around a period of 10 days. This is referred to in this paper as the nonlinearity of the Cepheid P-L relation. Arguments for choosing the fiducial period at 10 days can be found in Kanbur & Ngeow (2004), Sandage et al. (2004), Ngeow et al. (2005), and Ngeow & Kanbur (2006a). Furthermore, Kanbur & Ngeow (2004, 2006), Sandage et al. (2004), Ngeow et al. (2005), and Ngeow & Kanbur (2006c) examined various factors that may cause the nonlinearity of the LMC P-L relation, including the observing strategies, photometric errors, extinction errors, removal of outliers, sample selection, number of long-period Cepheids in the samples, and contamination of overtone Cepheids. They found that none of these factors or any combination of them could be responsible for the observed nonlinear LMC P-L relation. As argued in Ngeow & Kanbur (2006c), rigorous statistical tests are needed to test the linearity versus nonlinearity of the LMC P-L relation.

In our previous studies, the F-test (e.g., Weisberg 1980) has been applied to OGLE (Optical Gravitational Lensing Experiment) and MACHO (Massive Compact Halo Objects project) Cepheid data in Kanbur & Ngeow (2004) and Ngeow et al. (2005), respectively, to test for the nonlinearity of the LMC P-L relation. In such a formulation, the full and reduced models have four and two parameters, respectively. This test looks at the change in the mean residual sum of squares between the full and reduced model, divided by the mean residual sum of squares in the full model (see eq. [5] of Kanbur & Ngeow 2004). This test statistic can be formulated as the difference in slopes between short- and long-period slopes, divided by the standard error of that difference. Hence, if the number and nature of the long/short-period data are such that the long/short-period slope is estimated with a large error, then the F-value will be low and return a nonsignificant result. Thus, the F-test...
is sensitive to the number of data points on either side of the period cut at 10 days. The OGLE and MACHO data sets we used in Kanbur & Ngeow (2004) and Ngeow et al. (2005), respectively, do have an adequate number of long- and short-period Cepheids for the application of the F-test. The F-test has returned significant results when testing the nonlinearity of the P-L relation in both of the data sets.

Nevertheless, the results suggesting a nonlinear LMC P-L relation are still controversial. Even though we emphasize that statistical tests are needed, recent claims of linear LMC P-L relations in the literature lack rigorous statistical tests. In this paper, we apply two additional statistical tests, the testimator method and the Schwarz Information Criterion method, to examine the nonlinearity of the LMC Cepheid P-L relation. These tests are complementary to the F-test carried out in previous studies, since they serve to check and verify the results obtained from the F-test. In this way, previous conclusions about the nonlinear LMC P-L relation are considerably strengthened. Furthermore, both the testimator and Schwarz Information Criterion methods are also able to estimate the break period without any a priori assumption: recall that in previous work, a break period at 10 days is usually adopted. These two methods can be applied not only to Cepheid studies, as we do in this paper, but to other astronomical and astrophysical hypothesis-testing problems. We also emphasize that for the first time, our use of the testimator method and the Schwarz Information Criterion method, to examine the nonlinearity of the LMC Cepheid P-L relation. These statistical tests are needed, recent claims of linear LMC P-L relations in the literature lack rigorous statistical tests. In this paper, we apply two additional statistical tests, the testimator method and the Schwarz Information Criterion method, to examine the nonlinearity of the LMC Cepheid P-L relation as mentioned in the introduction (§ 1).

The description of the two-stage testimator method is summarized as follows. For a linear regression of the form \( y = \beta x + a \), the usual least-squares estimation of the slope to \( N \) data points is given as

\[
\hat{\beta} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2},
\]

where \( \bar{x} = N^{-1} \sum x_i \) and \( \bar{y} = N^{-1} \sum y_i \) are the mean values of \( x \) and \( y \), respectively. In the standard hypothesis-testing procedure, the null and alternate hypotheses are constructed as \( H_0 : \beta = \beta_0 \) and \( H_1 : \beta \neq \beta_0 \), respectively, where \( \beta_0 \) is the assumed value of (true) slope, given the prior knowledge on the slope. For example, \( \beta_0 \) can be predicted from theoretical calculations. In the case that the (true) variance of the slope is known, the \( z \)-statistical test (with normal distribution) can be applied; otherwise, the \( t \)-statistical test (with \( T \)-distribution) will be used for the hypothesis testing. In general, the variance is not known; therefore, we adopt the \( t \)-statistical test in this paper. If the null hypothesis is accepted from the hypothesis testing, the testimator (of the slope) \( \hat{\beta}_u \) is calculated as

\[
\hat{\beta}_u = k\hat{\beta} + (1 - k)\beta_0
\]

(Waikar et al. 1984).

The constant \( k \) in the above equation is defined as

\[
k = \frac{|t_{\text{observed}}|}{t_{\text{critical}}},
\]

where the mean square error \( \text{MSE} = (N-2)^{-1} \sum_{i=1}^{N} (y_i - \hat{a} - \hat{\beta}_u x_i)^2 \), \( S_{xx} = \sum_{i=1}^{N} (x_i - \bar{x})^2 \), and \( t_{\alpha/2, \nu} \) is the \( t \)-value for a \( 100(1 - \alpha/2)\% \) confidence interval obtained from the associated \( T \)-distribution table with \( \nu = N - 2 \) degrees of freedom. Note that the null hypothesis is rejected if \( k > 1 \). The properties of the testimator are such that

1. The testimator is an unbiased estimator under \( H_0 \).
2. The testimator has a smaller variance than the usual least-squares estimator; that is, \( \text{Var} \hat{\beta}_u < \text{Var} \hat{\beta} \).

The proofs for these two properties are given in the Appendix.
2.1.1. Application to the Cepheid P-L Relation

The motivation of this paper is to apply the testimator method to detect any nonlinearity in the LMC P-L relation; this has been detected using the $F$-test (Kanbur & Ngeow 2004; Ngeow et al. 2005). To study any possible variation in slope as the period increases through 10 days, we first sorted the data according to period, from shortest to longest period in $\log P$. The sorted sample was then divided into $m$ number of nonoverlapping and hence independent subsets, according to the Cepheid period. The purpose is to make the bivariate observations independent for each of the subsets. Each of the subsets will then contain $n$ numbers of Cepheids (if the number of data points in the last subset is small, then the last subset will be combined with the previous subset). This enables us to apply the testimator method in multiple stages, together with a conservative Bonferroni testing procedure,\(^1\) for detecting a slope variation in the sample. In essence, the line of attack is to compute the slope of the first subset and then compare with the slope of the next subset. If the two slopes are “similar,” we look at the slope of the third subset with the smoothed slope obtained from a combination of all the previous subsets. Hence, at the $i$th round, the slope of a given subset of the data is computed and compared with the smoothed slope from the testimator of all the previous data points. If the two slopes are statistically equivalent, then the current subset of data will be incorporated into the computation of the smoothed slope and compared with the slope of the next subset of data. This smoothness is an important feature, since it helps to alleviate, to some extent, the influence of outliers. However, if the slopes are “different” (i.e., a rejection of the null hypothesis), then there is an indication of slope change in the P-L relation. Therefore, there will be a total of $n_p = m - 1$ number of hypothesis testings in the multistage testimator procedures. In short, the algorithm of applying the testimator method in our case can be summarized as follows:

1. In the first round, the slope of the first subset, $\beta_1$, is calculated and denoted by $\hat{\beta}_1 = \beta_0$. The slope of the second subset is then compared to $\hat{\beta}_1$ under the null hypothesis $H_0: \beta_1 = \beta_0$ and alternate hypothesis $H_0^*: \beta_1 \neq \beta_0$. If $H_0$ is accepted, then the testimator in this round, $\hat{\beta}_2$, is calculated using equation (2).\(^2\)

2. In the second round, the slope of the third subset, $\beta_3$, is calculated and denoted by $\hat{\beta}_3$. The testimator from the first round, represented as $\hat{\beta}_2 = \beta_0$, is used in the hypothesis testing for this round. The null and alternate hypotheses in this round become $H_0: \beta_3 = \beta_0$ and $H_0^*: \beta_3 \neq \beta_0$. If $H_0$ is accepted, a new testimator, $\hat{\beta}_4$, is calculated using equation (2).\(^2\)

3. In the $i$th round, the slope of the $(i+1)$th subset, $\beta_{i+1}$, estimated by $\hat{\beta}_{i+1}$, is calculated. The testimator from the previous $(i-1)$ round is denoted as $\hat{\beta}_{i-1} = \beta_0$. The null and alternative hypothesis in this round become $H_0: \beta_{i+1} = \beta_0$ and $H_0^*: \beta_{i+1} \neq \beta_0$, respectively. If $H_0$ is accepted, then $\hat{\beta}_w = k\hat{\beta}_{i+1} + (1-k)\beta_0$, with $k$ refined from equation (3).\(^2\)

4. This is repeated until the $i = n_p$ round or the null hypothesis is rejected in the $i$th round, which indicates a change in slope for the $(i+1)$th subset.

5. Since in principle there will be a total of $n_p$ hypothesis testings, the Bonferroni testing procedure requires that $t_{critical} = t_{\alpha/(2n_p)}$ in each round.

Throughout this paper, we adopt $\alpha = 0.05$ to ensure that the overall confidence level is more than 95% in our test. The first two rounds of our testimator procedures to study the possible nonlinear LMC P-L relation is illustrated in Figure 1.

In order to demonstrate the reliability of this procedure, we apply the testimator method to two simulated data sets: one built from a nonlinear P-L relation with a break at 10 days, and another developed from a linear P-L relation. For demonstration purpose, we select one set of the simulated data (out of many simulations) in each case as a representation for testing the testimator method. The plots of these two fake data sets, each of them containing 1500 data points, can be found in Figure 1 of Ngeow & Kanbur (2006c). Full details of our procedure for developing these two “fake” data sets are described in Ngeow & Kanbur (2006c). The results of applying the testimator procedures as described above to these two fake data sets are illustrated in Figure 1.

\(^1\) The Bonferroni testing procedure states that for testing $n_p$ number of hypotheses, the confidence coefficient $(1 - \alpha/2)$ is replaced by $(1 - \alpha/(2n_p))$ in each of the hypothesis testings. This is to ensure that the overall confidence coefficient will not be less than the original desired value of $(1 - \alpha/2)$.
data sets are given in Table 1. In the table, column (1) denotes the subset; column (2) gives the range of the period in each subset; column (3) lists the number of data points, \( n \), in each subset; columns (4) and (5) are respectively the fitted slopes in each subset and the assigned values of \( \hat{\beta} \), that were used in the hypothesis testing; columns (6) and (7) are the respective observed and critical \( t \)-values for each of the hypothesis testing runs; columns (8) and (9) are the corresponding \( k \)-value and the outcome of the hypothesis testing, respectively; and finally, column (10) is the value of the estimator if the null hypothesis is accepted. Since we know which fake data set is intrinsically linear or nonlinear when constructing the P-L relation, we can verify the results found in Table 1. For the fake data with linear P-L relations, our estimator results show that the slopes for each subset are consistent with the smoothed slopes given from the previous subsets, and the hypothesis testings correctly indicate that there is no change in slope across all the period ranges. In contrast, the hypothesis testings for the fake data with nonlinear P-L relations show that subset 7 has a different slope than the previous subsets, which indicates a change of slope in this subset. Furthermore, the estimator procedures also correctly pick up the “break period” in subset 7, which brackets the input break period at 10 days, from the outcome of hypothesis testing. Therefore, the estimator method can pick up a P-L relation that is intrinsically nonlinear.

### 2.2. The Schwarz Information Criterion

The problem of deciding whether the LMC Cepheid data are more consistent with two lines of significantly different slopes rather than a single line is exactly analogous to deciding the dimensionality of the model that will fit the given LMC Cepheid data. The method of maximizing the likelihood tends to choose the highest possible dimension. Akaike (1974) suggested maximizing the likelihood subject to a penalty depending on the dimensionality of the model under consideration (Akaike Information Criterion, AIC): 

\[
\text{AIC} = -2 \ln L + 2k_p
\]

where \( L \) is the likelihood function of the model of dimension \( k_p \) (see Takeuchi 2000 as an example in the application of astronomy). However, Schwarz (1978) showed that maximum likelihood estimators can be obtained from large sample limits of Bayes estimates for certain classes of a priori distributions. These distributions only put positive probability on the subspaces of the parameter space corresponding to the competing models. Schwarz (1978) derived the following criterion (Schwarz Information Criterion, SIC; sometimes also referred to as the Bayesian Information Criterion [BIC] in the literature): choose the model for which

\[
\text{SIC} = -2 \ln L + k_p \ln N
\]

is a minimum, where \( N \) is the total number of data points and \( k_p = p + 1 \) (where \( p \) is the number of fitted parameters; see Schwarz 1978). Some examples of model selection using the BIC can be found in astronomical and astrophysical literature, including Arensloft et al. (2001), Handler et al. (2000, 2002), Koen (1996, 1999, 2006), Koen & Schumann (1999), Koen & Laney (2000), Koen & Lombard (1993, 2003), Liddle (2004, 2007), Mukherjee et al. (1998), Porciani & Norberg (2006), and Sterken et al. (1999).

### 2.2.1. Application to the Cepheid P-L Relation

To test the nonlinearity of the Cepheid P-L relation with the SIC method, we consider the models in this paper that have a linear P-L relation (the null hypothesis) and a nonlinear P-L relation (the alternative hypothesis).
relation with a break period (in days) at $P_0$ (the alternate hypothesis). For the former case, we have

$$H_0 : m = \hat{m} = \hat{\beta} \log P + \hat{a},$$

with

$$\hat{\sigma}^2 = \frac{1}{N - 2} \sum_{i=1}^{N} (m_i - \hat{m}_i),$$

$$L = \frac{1}{(2\pi\hat{\sigma}^2)^{N/2}} \exp \left[ -\frac{1}{2\hat{\sigma}^2} \sum_{i=1}^{N} (m_i - \hat{m}_i)^2 \right],$$

and

$$\text{SIC}(H_0) = -2 \ln L + 3 \ln N.$$  

Similarly, for the alternate model, we have

$$H_A : m = \hat{m}$$

\[
\begin{align*}
&\hat{\beta}_s \log P + \hat{a}_s, & \log P < \log P_0, \\
&\text{with } \hat{\sigma}_s^2 = \frac{1}{N_s - 2} \sum_{i=1}^{N_s} (m_i - \hat{m}_i), \\
&\hat{\beta}_L \log P + \hat{a}_L, & \log P \geq \log P_0, \\
&\text{with } \hat{\sigma}_L^2 = \frac{1}{N_L - 2} \sum_{i=1}^{N_L} (m_i - \hat{m}_i),
\end{align*}
\]

$$L = \frac{1}{(2\pi\hat{\sigma}_s^2)^{N_s/2}} \frac{1}{(2\pi\hat{\sigma}_L^2)^{N_L/2}} \times \exp \left[ -\frac{1}{2\hat{\sigma}_s^2} \sum_{i=1}^{N_s} (m_i - \hat{m}_i)^2 \right]$$

$$- \frac{1}{2\hat{\sigma}_L^2} \sum_{i=1}^{N_L} (m_i - \hat{m}_i)^2,$$

and

$$\text{SIC}(H_A) = -2 \ln L + 5 \ln N.$$  

In these expressions, $N_s + N_L = N$, and $m$ is the observed magnitude after correcting for extinction. The parameters of slope $\beta$ and zero point $a$ in the above models are obtained from the maximum likelihood estimation (MLE; which is equivalent to standard least-squares estimation in our case). Note that the sample variance $\sigma^2$ from the MLE is a biased estimate. We corrected the bias with $N_{(L,S)} - 2$ degrees of freedom. For the alternate models, $\text{SIC}(H_A)$ is calculated with a range of $\log P_0$ in increments, for example, of 0.001. Therefore, a model with a linear P-L relation, and a range of models with nonlinear P-L relations at different break periods, are tested with the SIC method. The model with smallest SIC value is the preferred model. In case of $\text{SIC}(H_A) < \text{SIC}(H_0)$, the minimum value of $\text{SIC}(H_A)$ not only suggests that the P-L relation is nonlinear, but also gives an estimate of the break period.

To test the SIC method, we applied the same simulated data sets that were used to test the estimator method in § 2.1.1. For the “fake” data set with a linear P-L relation, the values of $\text{SIC}(H_0)$ and $\text{SIC}(H_A)$ are $-164.65$ and $-161.21$, respectively, while for the “fake” data set with a nonlinear P-L relation, we found that $\text{SIC}(H_0) = -100.56$ and the minimum value of $\text{SIC}(H_A) = -154.62$ occurs at $\log P_0 = 0.983$. We then further tested the SIC method for our application, using various simulations. We first ran two sets of simulations: one set used the linear P-L relation as the input P-L relation, and another set of simulations included the nonlinear P-L relation, with a break at $\log P_0 = 1.0$. These simulations mimic the period distribution and the observed dispersions along the P-L relation from the real data. The details for constructing these simulations can be found in Ngeow & Kanbur (2006c). For each set of simulations, a large number of simulations is run (typically 1000), and the break period (in $\log P_0$) is searched for with the SIC method. If the break period cannot be found, then this implies that the linear P-L relation is the preferred model, and vice versa. The top panels of Figure 2 display the distributions of the break periods from these two sets of data. For the case of a linear P-L relation, the SIC method did not find any break period $\sim 90\%$ of the time, while for the case of no linear P-L relation, the SIC method detected a range of break periods with a peak at $\log P_0 \sim 1.0$. Therefore, the SIC method can be used to correctly identify the P-L relation that is either intrinsically linear or nonlinear at a given break period.

The relatively large dispersion around $\log P_0 \sim 1.0$, and the long tail toward the shorter period, which are exhibited in the top panels of Figure 2, could be due to a combination of two effects: (1) the existence of the intrinsic dispersion along the P-L relation, and (2) the nonuniform distribution of the periods in the data (see Ngeow et al. 2005 and Ngeow & Kanbur 2006c for more discussion about the period distribution for Cepheid variables). To portray the impact of these effects on the application of the SIC method for detecting the break period, we ran two additional experiments. The first retained the original period distribution, but the intrinsic dispersion of the P-L relation was excluded (however, the random photometric errors still persist in the simulation), while the second simulation used a uniform period distribution (in $\log P$), and the intrinsic dispersion of the P-L relation was not excluded. The resulting distributions of the break period from the SIC method are presented in the bottom panels of Figure 2. It can be seen that the long tail of the distribution present in the top panels is reduced when a uniform period distribution is assumed. Furthermore, if the intrinsic dispersion does not exist in the Cepheid P-L relation, then the SIC method is very efficient for detecting the
intrinsic break period (at log $P_o = 1.0$ in our case). In reality, the intrinsic dispersion along the P-L relation cannot be eliminated or reduced (at least in the optical bands), and the period distribution of the Cepheid variables will not be uniform (for the reasons discussed in Ngeow & Kanbur 2006c). We emphasize that the theoretical pulsation models are needed to identify the location of the break period or to confirm the break period at log $P_o \sim 1.0$ (Ngeow et al. 2005).

### 3. DATA AND RESULTS

In this section, we apply both the testimator and SIC methods to the real LMC Cepheid data in order to investigate whether or not the V-band Cepheid P-L relation at mean light is nonlinear. We concentrate on V-band mean light data because this is the most prevalent type available in the literature, and also because the evidence for nonlinearity as a function of phase is clear (Ngeow & Kanbur 2006a). The data sets we used in this study include the OGLE data adopted from Kanbur & Ngeow (2006) and the MACHO data adopted from Ngeow et al. (2005). They are referred to as the “OGLE” sample (with 641 Cepheids) and the “MACHO” data (with 1216 Cepheids), respectively. Note that both data sets have been corrected for extinction using the method described in the corresponding papers. It is also important to point out that these are two independent data sets. To investigate the influence of longer period stars in our testing, and to increase the number of Cepheids in the OGLE sample, we append the data from Sebo et al. (2002) to the OGLE sample after properly removing duplicate Cepheids in both samples and correcting for extinction. This third data set is called the “OGLE+Sebo” sample (with 723 Cepheids), and it extends to log $P \sim 2.0$.

The results from applying the testimator method to these three LMC Cepheid data sets are summarized in Table 2, using the same layout from Table 1. In the case of the OGLE and OGLE+Sebo data sets, we have tried different sample subset sizes by dividing the samples into and , which are respectively referred as Test 1 and Test 2 in the table. In all cases, the testimator method implies that there is a change of slope in the last subset of the samples. Similar results found from Test 1 and Test 2 suggest that our results are not affected by the size of each subset. This indicates that the LMC P-L relation becomes nonlinear as the period increases through 10 days to longer periods. The last subset also brackets the fiducial break period at or around 10 days: this is consistent with previous work of Ngeow et al. (2005).

The results from using the SIC method are presented in Figure 3 and Table 3 for the same data sets. In Figure 3, the values of SIC for both SIC($H_o$) and SIC($H_o'$) are plotted as a function of the chosen break period, log $P_o$. Since SIC($H_o$) is independent of log $P_o$, this represents a straight horizontal line in the figure, and the values of SIC($H_o'$) for these three data sets are given in Table 3. For the case of SIC($H_o'$) as a function of log $P_o$, we have used the same method described in the corresponding paper. It is also important to point out that these are two independent data sets. To investigate the influence of longer period stars in our testing, and to increase the number of Cepheids in the OGLE sample, we append the data from Sebo et al. (2002) to the OGLE sample after properly removing duplicate Cepheids in both samples and correcting for extinction. This third data set is called the “OGLE+Sebo” sample (with 723 Cepheids), and it extends to log $P \sim 2.0$.
of log $P_0$, the figure bears witness to the fact that there is a range of log $P_0$ for which the values of SIC($H_0$) are smaller than SIC($H_1$) in all three data sets. This implies that the nonlinear P-L relation is preferred within these period ranges. This result also reinforces the findings of Figure 2 that it is difficult to determine the exact break period of the P-L relation with the SIC method (see § 2.2.1 as well), if it is present. The minimum values for SIC($H_0$) found from the figure, and the corresponding log $P_0$, are also summarized in Table 3. The confidence intervals for the break period can be estimated using bootstrap resampling methods. For the model with given log $P_0$ in Table 3, the errors of the regression, $e_i = m_i - \hat{m}_i$, are randomly drawn (with replacement) to construct a “new” data set, and a new break period is estimated. This is repeated many times to build up the distribution of the break periods. The resulting histograms for these three sets of data are presented in Figure 4. From these distributions, the fifth, 25th, 75th, and 95th percentiles are estimated for each of the data sets. The results are given in the last four columns of Table 3. At first glance, the break period found from the MACHO data seems to be inconsistent with the OGLE and OGLE+Sebo results. This is due to the difficulty of accurately estimating the break period, due to the existence of the instability strip. To demonstrate this, we use the exact periods in MACHO data as input periods to our simulations, and generate three different sets of simulations: (1) a simulation with an intrinsic nonlinear P-L relation, (2) a simulation with a linear P-L relation, and (3) a simulation with an intrinsic nonlinear P-L relation, but without the intrinsic dispersion. The resulting histograms for these three sets of simulations are displayed in Figure 5. From this figure, it is clear that our result for the break period from MACHO data does not imply an inconsistency with the OGLE and OGLE+Sebo results. The break period found in the data, log $P_0 = 0.833$, is within the range of the break periods found...
from the simulations. This figure also portrays the difficulty of estimating the break period from real data when the intrinsic dispersion along the P-L relation is present. Therefore, the break periods given in Table 3 are consistent with the results from testimator (Table 2), the result from the nonlinear estimation procedure applied in Ngeow et al. (2005; log $P_0 = 0.934$, with upper and lower 95% confidence levels of 1.089 and 0.778, respectively), and the adopted log $P_0 = 1.0$ in the literature. Note that in previous studies (e.g., Tammann & Reindl 2002; Kanbur & Ngeow 2004, 2006; Sandage et al. 2004; Ngeow et al. 2005; Ngeow & Kanbur 2006b), the break period is conveniently chosen to be at 10 days, which is rep-

### Table 3

| Data Set             | SIC($H_a$) | SIC($H_o$) | log $P_0$ | Fifth Percentile | 25th Percentile | 75th Percentile | 95th Percentile |
|----------------------|------------|------------|-----------|------------------|-----------------|-----------------|-----------------|
| OGLE sample          | $-179.29$  | $-188.86$  | $1.041$   | $0.550$          | $1.002$         | $1.041$         | $1.101$         |
| OGLE+Sebo sample     | $-296.43$  | $-304.28$  | $1.041$   | $0.560$          | $0.922$         | $1.052$         | $1.131$         |
| MACHO sample         | $182.61$   | $156.20$   | $0.833$   | $0.806$          | $0.826$         | $0.838$         | $0.936$         |
Fig. 4.—Histograms resulting from the bootstrap resampling at the break period given in Table 3 for the three LMC Cepheid data sets. See text for details.

Fig. 5.—Comparisons of histograms from three sets of simulations for the MACHO data: (1) a simulation that takes a nonlinear P-L relation with a break at log $P_0 = 1.0$ as the input P-L relation, and the intrinsic dispersion is included; (2) a simulation with a linear P-L relation as the input P-L relation, and the intrinsic dispersion is included; and (3) a simulation that takes a nonlinear P-L relation with a break at log $P_0 = 1.0$ as the input P-L relation, but without the intrinsic dispersion. Unlike other simulations done in this paper, the periods that go into the simulations are from the actual MACHO data.

represented as a dotted vertical line in Figure 3. The SIC results also supported the nonlinear P-L relation as the preferred model at log $P_0 = 1.0$.

4. CONCLUSION AND DISCUSSION

By applying two additional statistical approaches, the testimator method and the Schwarz Information Criterion, to the OGLE, OGLE+Sebo, and MACHO Cepheid data, we have found strong statistical evidence for a change of slope in the Cepheid P-L relation in the LMC for a longer period range. These results also strongly support previous results obtained from the F-test. Therefore, the observed LMC P-L relation is nonlinear, based on these rigorous statistical tests. This implies that either the LMC P-L relation is indeed nonlinear, or else there are some hidden factors in the analysis (see Ngeow & Kanbur 2006c for more discussion on this). Furthermore, the break periods, or the range of permissible break periods found from this study, are consistent with the conveniently chosen break period at 10 days in previous studies. However, our study, using both real and fake data, implies that it is difficult to accurately estimate the break period with both the testimator and SIC methods. This is due to the existence of the intrinsic dispersion along the P-L relation. The confirmation of the break period at or around 10 days has to be achieved from stellar pulsation modeling studies.

The implications of a nonlinear LMC P-L relation on the extragalactic distance scale and the Hubble constant have been discussed in Ngeow & Kanbur (2005, 2006b) and are not repeated here. A number of authors, including Spergel et al. (2007), Tegmark et al. (2006), Macri et al. (2006), and Olling (2007), and references therein, have commented on how an independent estimate of the Hubble constant that is accurate to 1% can significantly reduce the error bars on $\Omega$, the total density of the universe. Applying the correct form of the Cepheid P-L relation will help to reduce the systematic error of the Hubble constant in future studies (Ngeow & Kanbur 2006b, 2006c). Over and above this, if the Cepheid P-L relation does indeed have a change of slope at 10 days, it is important to understand this from a stellar pulsation and evolutionary point of view, and to investigate fully the ramifications of this new feature (Kanbur & Ngeow 2006; Marconi et al. 2005).

Ngeow & Kanbur (2006c) have investigated various factors that may cause the observed nonlinear LMC P-L relation, including the influence of outliers and a lack of longer period Cepheids in the sample. However, the results from that study suggest that none of these factors are responsible for the observed nonlinear LMC P-L relation. We emphasize that the samples we used in our studies have been cleaned of obvious outliers. Furthermore, the testimator approach estimates the slope with a variance that is smaller than the standard formula (property 2 in § 2.1.1) and is able to minimize the effect of (additional) outliers by smoothing. Regarding the lack of longer
period Cepheids in the sample, we have used the OGLE+Sebo combination as an expansion to the OGLE sample, with an increased period coverage. Both samples have shown the same results using the testimator and SIC methods. Therefore, we believe these should not be causes for the observed nonlinear LMC P-L relation.

The authors would like to thank the referee for useful suggestions. This research was supported in part by NASA through the American Astronomical Society’s Small Research Grant Program. C. N. acknowledges financial support from NSF award OPP-0130612 and a University of Illinois seed funding award to the Dark Energy Survey.

APPENDIX A

PROOF FOR THE PROPERTIES OF THE TESTIMATOR

Here we prove the two properties of the testimator as described in § 2.1. To prove that the testimator is an unbiased estimator under $H_0$, we note that the testimator from equation (1) is

$$\hat{\beta}_o = k(\hat{\beta} - \beta_o) + \beta_o,$$

where $k$ is defined in equation (3). Therefore, the above expression can be rewritten as

$$\hat{\beta}_o = \frac{\hat{\beta} - \beta_o}{t_{w_2,r}/\text{MSE}}(\hat{\beta} - \beta_o) + \beta_o. \quad (A1)$$

This implies that

$$E(\hat{\beta}_o) = \frac{\sqrt{S_{xx}}}{t_{w_2,r}}E\left(\frac{1}{\text{MSE}}\right)[\hat{\beta} - \beta_o][\hat{\beta} - \beta_o] + \beta_o.$$ 

Since $E(|z|z) = 0$ for variable $z = \hat{\beta} - \beta_o$ with a standard normal distribution, and from the above expression, we obtain

$$E(\hat{\beta}_o) = \beta_o,$$

as desired. The second assertion states that $\text{Var} \hat{\beta}_o < \text{Var} \hat{\beta}$. To prove this, we first rearrange equation (A1) such that

$$(\hat{\beta}_o - \beta_o)^2 = \frac{\hat{\beta} - \beta_o}{t_{w_2,r}/\text{MSE}}^2 S_{xx}.$$ 

Assuming $\hat{\beta}$ is normally distributed with $N(\beta_o, \sigma^2)$, and define $Z = (\hat{\beta} - \beta_o)/\sigma_\beta$, then $Z$ has a standard normal distribution with $N(0, 1)$. Note that $\sigma_\beta^2 = \text{Var} \hat{\beta} = \sigma^2/S_{xx}$, where $\sigma^2$ is the variance of the linear regression $y = \beta x + a$, and the above expression is reduced to

$$(\hat{\beta}_o - \beta_o)^2 = Z^2 \frac{\sigma^2}{\text{MSE}} \frac{1}{t_{w_2,r}} S_{xx}.$$ 

Hence, we have

$$\text{Var} \hat{\beta}_o = E(Z^2)E\left(\frac{\sigma^2}{\text{MSE}}\right) \frac{1}{t_{w_2,r}} S_{xx}, \quad (A2)$$

as the last two terms are constants and $\text{Var} \hat{\beta}_o = E[(\hat{\beta}_o - \beta_o)^2]$. For $E(Z^2)$, since the fourth moment of a standard normal distribution (the kurtosis) is 3, then $E(Z^2) = 3$. For $E(\sigma^2/\text{MSE})$, we observe that $\sigma^2/\text{MSE} = 1/(\text{MSE}/\sigma^2) = (N-2)/\Sigma [(y_i - \hat{\alpha} - \hat{\beta} x_i)/a]^2$. Therefore, $(N-2)\text{MSE}/\sigma^2$ is $\chi^2$ distributed with $\nu = N-2$ degrees of freedom. It is well known that if $X$ is $\chi^2$ distributed with $\nu$ degrees of freedom, then $E(1/X) = 1/(\nu-2)$; hence, $E(\sigma^2/\text{MSE}) = (N-2)/N-4$. Recall that $\sigma^2/S_{xx} = \text{Var} \hat{\beta}$; thus, equation (A2) is reduced to

$$\text{Var} \hat{\beta}_o = 3 \frac{(N-2)}{(N-4)} \frac{1}{t_{w_2,r}} \text{Var} \hat{\beta}.$$

If $t_{w_2,r} > [3(N-2)(N-4)]^{1/2}$, we then have

$$\text{Var} \hat{\beta}_o < \text{Var} \hat{\beta},$$

as the assertion states. Due to the Bonferroni testing procedure, the condition $t_{w_2,r} > [3(N-2)(N-4)]^{1/2}$ is satisfied when $N > 5$ and $\alpha < 0.1$.

REFERENCES

Akaike, H. 1974, IEEE Trans. Automatic Control, 19, 716
Arentoft, T., Sterken, C., & Handler, G. 2001, MNRAS, 326, 192
Arnold, J. C., & Katti, S. K. 1972, J. Multivariate Anal., 2, 236
Bancroft, T.A. 1944, Ann. Math. Stat., 15, 190
———. 1964, Biometrics, 20, 427
Bock, M. E., Yancey, T. A., & Judge, G. G. 1973, J. Am. Stat. Assoc., 68, 109

Ghosh, M., & Sinha, B. K. 1988, J. Multivariate Anal., 27, 206
Han, C. P. 1978, J. Am. Stat. Assoc., 73, 855
Handler, G., Romero-Colmenero, E., & Montgomery, M. H. 2002, MNRAS, 335, 399
Handler, G., et al. 2000, MNRAS, 318, 511
Huntsberger, D. V. 1955, Ann. Math. Stat., 26, 734
Kanbur, S., & Ngeow, C. 2004, MNRAS, 350, 962

2007 PASP, 119:512–522
Kanbur, S., & Ngeow, C. 2006, MNRAS, 369, 705
Koen, C. 1996, MNRAS, 283, 471
———. 1999, MNRAS, 309, 769
———. 2006, MNRAS, 365, 489
Koen, C., & Laney, D. 2000, MNRAS, 311, 636
Koen, C., & Lombard, F. 1993, MNRAS, 263, 287
———. 2003, MNRAS, 343, 241
Koen, C., & Schumann, R. 1999, MNRAS, 310, 618
Liddle, A. 2004, MNRAS, 351, L49
———. 2007, MNRAS, 377, L74
Macri, L., Stanek, K., Bersier, D., Greenhill, L., & Reid, M. 2006, ApJ, 652, 1133
Marconi, M., Musella, I., & Fiorentino, G. 2005, ApJ, 632, 590
Mukherjee, S., Feigelson, E. D., Jogesh Babu, G., Murtagh, F., Fraley, C., & Raftery, A. 1998, ApJ, 508, 314
Ngeow, C., & Kanbur, S. 2005, MNRAS, 360, 1033
———. 2006a, MNRAS, 369, 723
———. 2006b, ApJ, 642, L29
———. 2006c, ApJ, 650, 180
Ngeow, C., Kanbur, S., Nikolaev, S., Buonaccorsi, J., Cook, K., & Welch, D. 2005, MNRAS, 363, 831
Olling, R. 2007, MNRAS, in press (astro-ph/0607607)
Pandey, B. N. 1997, Commun. Stat.—Theory and Methods, 26, 2191
Pandey, B. N., & Malik, H. J. 1990, Commun. Stat.—Theory and Methods, 19, 629
Pandey, B. N., Malik, H. J., & Dubey, P. K. 1995, Commun. Stat.—Theory and Methods, 24, 2261
Pandey, B. N., & Srivastava, A. K. 2001, IAPQR Trans., 26, 109
Paul, A. E. 1950, Ann. Math. Stat., 21, 539
Porciani, C., & Norberg, P. 2006, MNRAS, 371, 1824
Sandage, A., Tammann, G. A., & Reindl, B. 2004, A&A, 424, 43
Schwarz, G. 1978, Ann. Stat., 6, 461
Sebo, K., et al. 2002, ApJS, 142, 71
Spergel, D., et al. 2007, ApJ, in press (astro-ph/0603449)
Sterken, C., Broens, E., & Koen, C. 1999, A&A, 342, 167
Takeuchi, T. 2000, Ap&SS, 271, 213
Tammann, G. A., & Reindl, B. 2002, Ap&SS, 280, 165
Tegmark, M., et al. 2006, Phys. Rev. D, 74, 123507
Waikar, V. B., Ratnaparkhi, M. V., & Schuurmann, F. J. 2002, in Monte Carlo and Quasi-Monte Carlo Methods 2000, ed. K.-T. Fang, F. Hickernell, & H. Niederreiter (Berlin: Springer), 487
Waikar, V. B., Schuurmann, F. J., & Raghunathan, T. E. 1984, Commun. Stat.—Theory and Methods, 13, 1901
Weisberg, S. 1980, Applied Linear Regression (1st ed.; New York: Wiley)
Yancey, T. A., Judge, G. G., & Bohrer, R. 1989, Econometrica, 57, 1221