SF2A 2016
C. Reylé, J. Richard, L. Cambrésy, M. Deleuil, E. Pécontal, L. Tresse and I. Vauglin (eds)

IDENTIFYING THE NON-LINEARITIES IN THE PERIOD-LUMINOSITY RELATIONS OF CEPHEIDS

J. Dassa-Terrier[1]

Abstract. The period-luminosity relation of Cepheids is essential to characterize the second step of the cosmological distance ladder. However, the possible non-linearity and breaking points around a period of 10 days have important consequences for their use. We review and challenge four different statistical methods on our simulations. We constrain the conditions in which these methods can be applied.

Keywords: stars: variables: Cepheids

1 Introduction

It is possible to segment the cosmological distance ladder into three main steps. First, the measurement of parallax in the Milky Way, initiated by both von Struve and Bessel in 1837 and 1838 respectively. Then, the period of pulsation of Cepheids in the Local Universe, done by Leavitt for the first time in Leavitt & Pickering (1912). Finally, the magnitude of type Ia supernovae in distant galaxies.

Our work is focused on the second step, the Period-Luminosity relation, also known as Leavitt Law (LL):

$$\langle M \rangle = a \log P + b, \tag{1.1}$$

where $\langle M \rangle$ is the mean absolute magnitude, P the period, a and b the slope and zero point of the linear fit respectively. Cepheids being standard candles, the slope a of the LL can be found by observing the magnitude and period of a group of Cepheids while the zero point b must be found using galactic Cepheids with known distances.

The knowledge of the deduced absolute magnitude M combined with the observed apparent magnitude m allows to find the distance modulus:

$$\mu = \langle m \rangle - \langle M \rangle = 5 \log d - 5, \tag{1.2}$$

where d is the distance to the star. Applied to Cepheids in M31, this method allows to estimate its distance at $d_{M31} = 752 \pm 27\ kpc$ (Riess et al. 2012). This permits us to know distances to type Ia supernovae in the nearby galaxies, hence, calibrate their light curves which peaks at a standard absolute magnitude, leading to the Hubble constant H_0 (Sandage et al. 2006).

However, recent studies, including Tammann et al. (2002), Ngeow & Kanbur (2009) and Kodric et al. (2015), strongly imply that the LL shows a non-linear behaviour in the Large Magellanic Cloud (LMC), Small Magellanic Cloud (SMC) and the M31 galaxy with the presence of a breaking point at $P \simeq 10\ days$. This break in slope was already suggested by Karkurin in 1937 but still lacks a consensus for its physical explanation. Several solutions were suggested for this empirical observation such as: the dependence of the LL on metallicity, or the interaction of stellar photosphere with the hydrogen ionization front (Kanbur & Ngeow 2006). An example of LL with breaking point is shown in figure 1.

This breaking point impacts the distance modulus calculation with non-negligible consequences. Kodric et al. (2015), who uses a model with breaking point, estimates H_0 to be 3.2% larger than the one found in Riess et al. (2012), who uses a linear model.

We explore four different statistical techniques: the F-test, the testimator, the BP from Bai & Perron (2003) and the bcp from Barry & Hartigan (1993), to estimate rigorously the statistical significance of this non-linearity in the LL.

[1] Observatoire de Paris, Paris, France

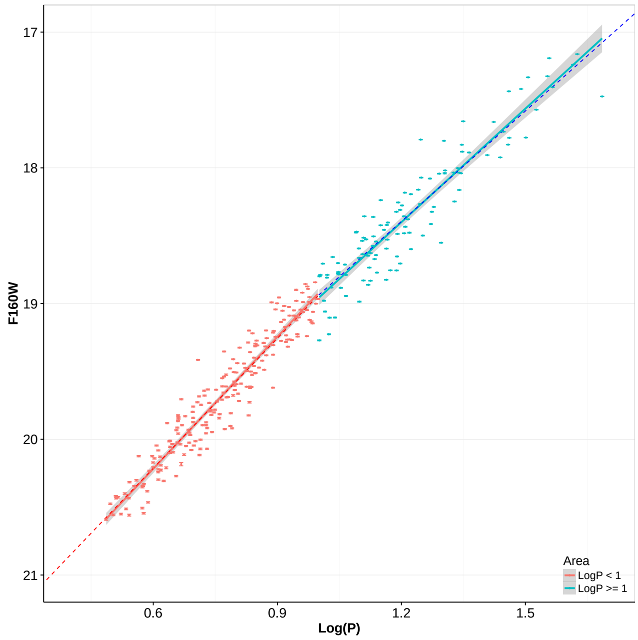

Fig. 1. The LL breaking point for 319 fundamental mode Cepheids in M31 as shown in Kodric et al. (2015).

2 Statistical methods

To test the ability of these methods to detect a breaking point in observations, we build simulations with similar ranges of period, magnitude and uncertainties as in the SMC, LMC and M31. These simulation can be fitted by two linear regressions with variable slopes and degrees of freedom.

2.1 F-test

Applied in Bhardwaj et al. (2016) and Kanbur & Ngeow (2004), the F-test is a ratio of the variance between the sample means and the variance within the samples. We choose two models:

$$M = a + b \log P, \tag{2.1}$$

and:

$$M = a_1 + b_1 \log P + \lambda(a_2 + b_2 \log P), \tag{2.2}$$

where $\lambda = 0$ if $P < 10$ days and $\lambda = 1$ if $P > 10$ days.

Our null hypothesis is the data being fit by a single linear regression. The F-test can be applied using the residues of the regressions as samples:

$$F = \frac{(RSS_1 - RSS_2)/[DF_1 - DF_2]}{RSS_2/DF_2}, \tag{2.3}$$

where RSS_1 and RSS_2 are the residual sums of squares in the 1-regression model and 2-regressions model respectively, DF_1 and DF_2 the degrees of freedom (DF) of the two models. RSS_1 and RSS_2 are found adaptating the general case: $RSS = \sum_{i=1}^{n}(y_i - f(x_i))^2$. We rely on the F-distribution for these DF to calculate the p-value associated with the value of F found.

The F-test needs to satisfy the following assumptions to be applied: homoscedasticity of the sample, normality of the residues and independent identically distributed observations. It is essential to check if these assumptions are met by the samples.

2.2 Testimator

The testimator (for test estimator) method is a variation on the estimator and was proposed by Bancroft (1944). In our case, the sample has to be divided in several independent subsets of increasing period. The slopes of the first two subsets will be compared with a t-test. If the null hypothesis, the two slopes are statistically identical, is accepted, then the two subsets are merged and its slope becomes the testimator. This new subset will be compared to the next adjacent subset under the same conditions and so on, until the null hypothesis is rejected (Bhardwaj et al. 2016; Kanbur et al. 2007).

In the case of a simple linear regression:

$$y = \beta_i x + a, \tag{2.4}$$

where β_i is the slope of the subset i, you can test the hypothesis $\beta_1 = \beta_0$, where β_0 is the prior (here the slope for the first subset). If successful, the testimator β_ω can be calculated:

$$\beta_\omega = k\beta_1 + (1-k)\beta_0, \tag{2.5}$$

where k is the ratio of the t-value found and the t-value which would be associated to a p-value of 0.05.

Kanbur et al. (2007) provides a detailed demonstration that the testimator is an unbiased estimator under the null hypothesis and that it has a smaller variance than the usual least-square estimator. It also has the interesting property of smoothing the impact of outliers. The main challenge of the testimator method is the choice of the size of the subsets. It also provides only an upper or lower limit to the value of the breaking point.

2.3 breakpoints

Bai & Perron (2003) describes the theoretical framework of the beakpoints (BP) algorithm in detail. Their dynamic programming algorithm segments the sample and look for its optimal partitioning, by minimizing the sum of squares in each segments. The implementation of this algorithm in the R package $STRUCCHANGE$ will find the optimal model based on the Bayesian Information Criterion (Erdman & Emerson 2007).

2.4 bcp

Barry & Hartigan (1993) gives a bayesian solution to the breakpoint identification problem. This method can be applied for observations independent under the normal assumption $N(\mu_i, \sigma^2)$ and it gives the probability p_i for a breaking point at position i. The full theoretical framework of this method being extensive, we will only give a brief description here.

The sample is subdivided into blocks, for each block from position $i+1$ to j, a prior distribution $\mu_i j$ with normal assumption $N(\mu_0, \sigma^2/(j-i))$ is defined. At each position i, the probability for a break point at position $i+1$ is calculated from the ratio:

$$\frac{p_i}{1-p_i} = \frac{P(U_i = 1|X, U_j, j \neq i)}{P(U_i = 0|X, U_j, j \neq i)}, \tag{2.6}$$

where U_i is an indicator of the presence or absence of change point at position $i+1$. If $U_i = 1$, a break point is identified at $i+1$.

This method was implemented in R in the package *bcp* by Erdman & Emerson (2007)

3 Simulations

In order to test these methods, we build simulations consisting of N observations with x-axis values ranging from 0 to 1.7 and y-axis values calculated with the model from equation 2.2 to which we add a normal perturbation σ:

$$y = a_1 + b_1\ x + \lambda(a_2 + b_2\ x) + \sigma, \tag{3.1}$$

where $\lambda = 0$ if $x < \tau$ days and $\lambda = 1$ if $x > \tau$ days. τ is the breaking point we are choosing.

We know the methods to be impacted by the number of observations, the noise and the difference between b_1 and b_2. Two different sets of data are built. In the first set, N will vary from 50 to 500 on a log scale, σ will vary from 0.05 to 0.5 by steps of 0.05 and Δ will vary from 0 to 0.9 by steps of 0.1. This data will serve for the F-test, the Testimator and BP. In the second set, N will vary from 50 to 235 on a log scale, σ will vary from 0.05 to 1 and Δ will vary from 0 to 0.3. This data will serve for the bcp method.

Each combination will be used to produce 200 samples (first set) and 27 samples (second set).

4 Results

We apply the four methods described to our simulations. In figure 2, we display the p-value found by the F-test and observe for high Δ, low σ and high N a clear trend toward low p-values. For no difference between the two slopes, we have no false identification, which confirms the F-test as a reliable tool for false positive checks.

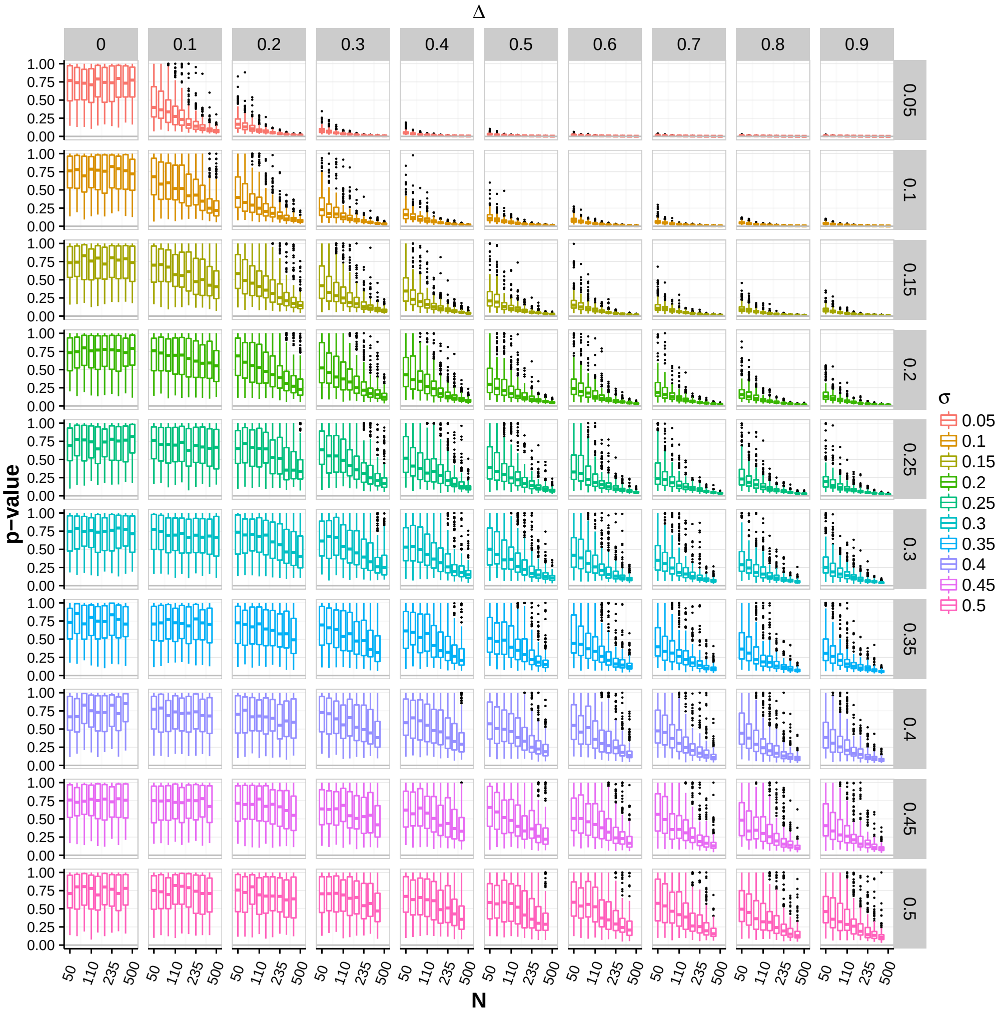

Fig. 2. Evolution of the p-value found by the F-test with the number N of observations given for each combination of added noise σ and slope difference Δ. Outliers are shown in red.

In figure 3, we show the difference between the real breaking point τ and the breaking point found by Testimator. Though the trend is not as strong as in the F-test, it is important to note that the Testimator can only provide a lower limit to the value of the breaking point. We see for $\Delta = 0$ that the testimator will go through all the subsets without detecting any breakpoint in most of the samples.

The BP method is tested in figure 4. Again we show the difference between the predefined τ and the one found by the method. Promising results are found even for high values of σ where the average value is close to 0. It is even more remarkable given the fact that, contrarily to the F-test, the BP algorithm does not receive any input else than the data itself.

The bcp method requires a few more steps. In the first place, a probability density function (PDF) is produced from each combination, such as in figure 5. Then the maximum value of the PDF is used to choose a breaking point and applied the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) to the reduced and full regression model (see equations 2.1 and 2.2). We will note $AIC = AIC_r - AIC_f$ and $BIC = BIC_r - BIC_f$. This leads to figure 6 and figure 7, in which we show the difference between the BIC and the AIC found from the reduced and full model, respectively.

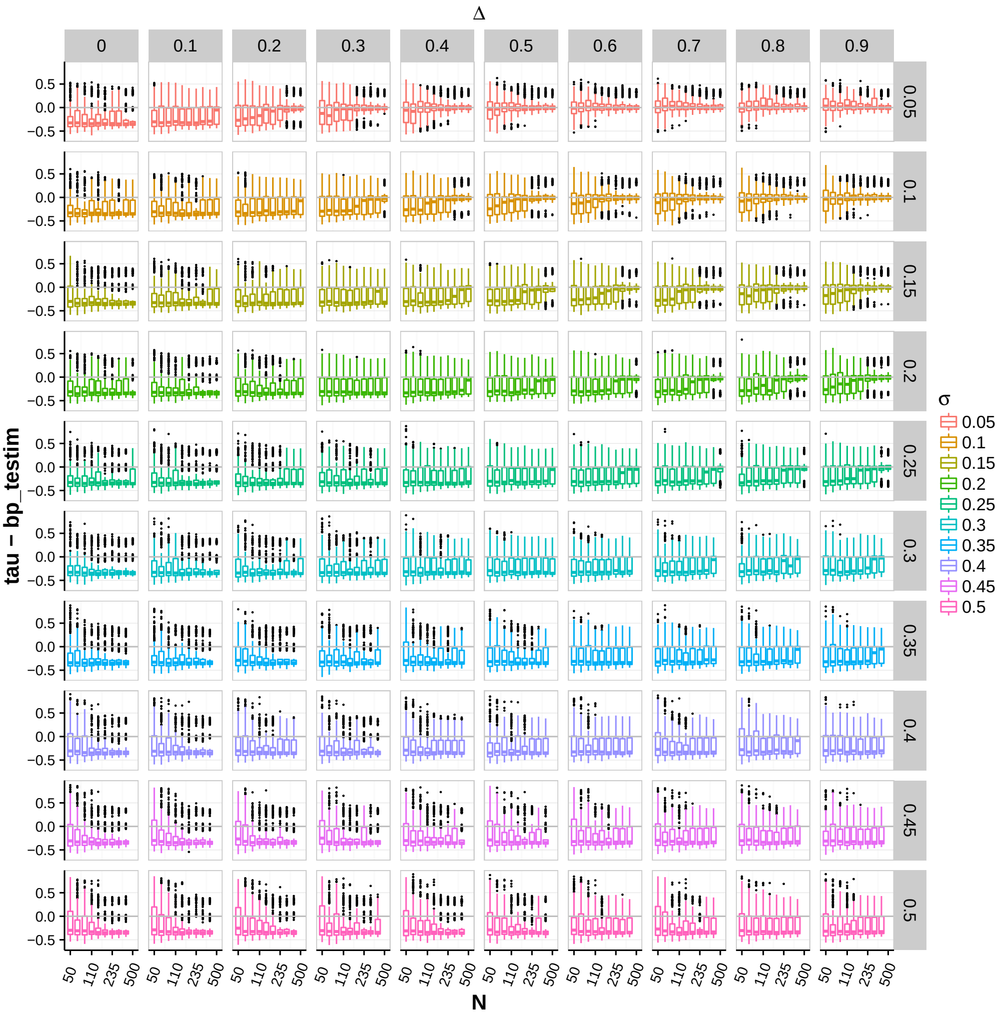

Fig. 3. Evolution of the breaking point value found by Testimator with the number N of observations given for each combination of added noise σ and slope difference Δ.

Values of AIC and BIC tend to get higher following the same trend as the other methods.

5 Discussion and Conclusions

Breaking points can be identified by all four methods. A first estimation would suggest that, for samples of approximatively 200 or more observations, Δ is required to be twice as large as σ in order to have an identification occurring.

The F-test never gave a false positive identification.

Testimator has not shown a very strong resolving power but is strongly dependant of the way the bins are chosen. Defining the optimum number of subsets according to N, σ and Δ should be investigated in further works.

The BP algorithm has strong results but finds a non-negligible number of false positive for $\Delta = 0$. Hence, a verification with the F-test is advised, as it showed that the probability of a false positive with this method is extremely small.

The bcp method shows promising results and requires further investigation with a higher number of simulations and wider range of values for N, σ and Δ. We will also adapt our simulations in order to make them closer to real observations (particularly populations of Cepheids). More statistical methods will be challenged and be applied to real data after similar analysis are completed.

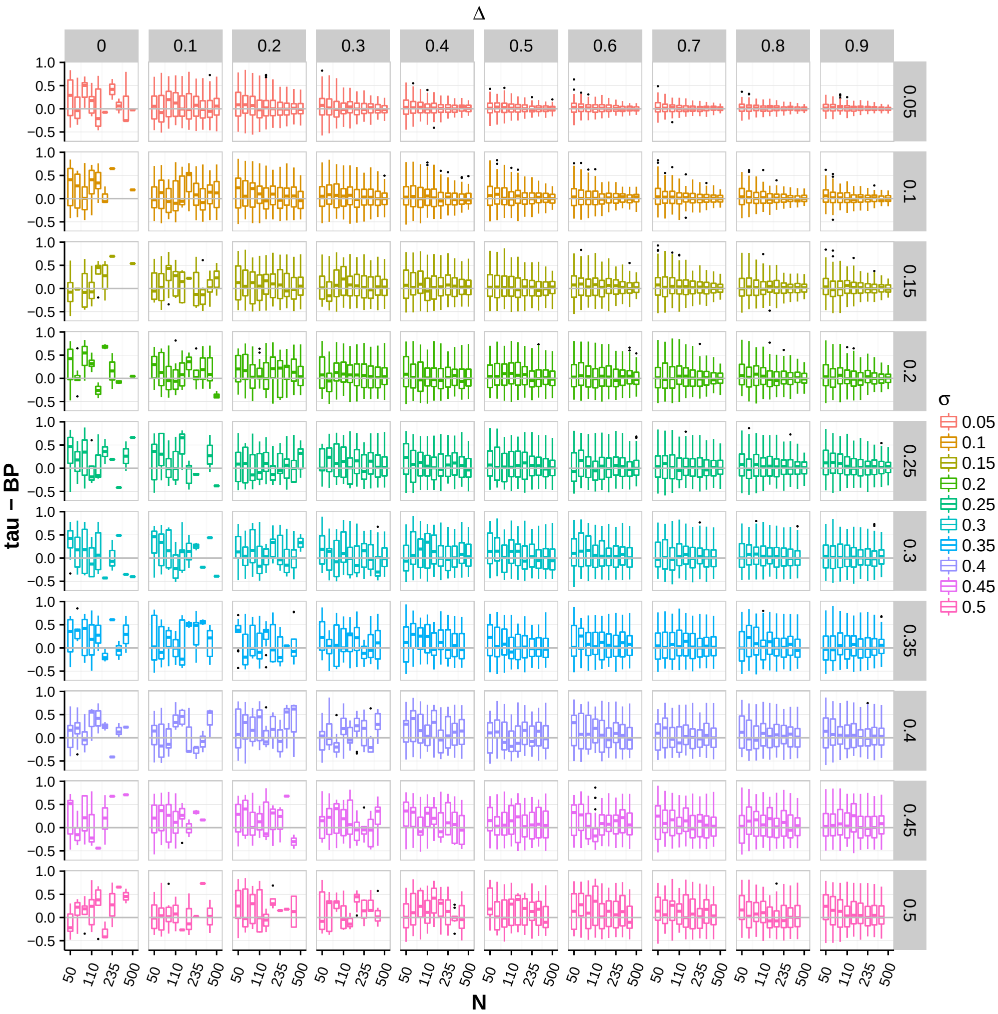

Fig. 4. Evolution of the breakpoint τ value found by BP with the number N of observations given for each combination of added noise σ and slope difference Δ. BP returns no value if no breakpoint is identified.

We would like to thank David Valls-Gabaud, Anne-Laure Melchior and Gérard Grégoire.

References

Bai, J. & Perron, P. 2003, Journal of Applied Econometrics, 18, 1

Bancroft, T. A. 1944, Ann. Math. Statist., 15, 190

Barry, D. & Hartigan, J. 1993, Journal of the American Statistical Association, 88, 309

Bhardwaj, A., Kanbur, S. M., Macri, L. M., et al. 2016, MNRAS, 457, 1644

Erdman, C. & Emerson, J. 2007, Journal of Statistical Software, 23, 1

Kanbur, S. M., Ngeow, C., Nanthakumar, A., & Stevens, R. 2007, pasp, 119, 512

Kanbur, S. M. & Ngeow, C.-C. 2004, MNRAS, 350, 962

Kanbur, S. M. & Ngeow, C.-C. 2006, MNRAS, 369, 705

Kodric, M., Riffeser, A., Seitz, S., et al. 2015, apj, 799, 144

Leavitt, H. S. & Pickering, E. C. 1912, Harvard College Observatory Circular, 173, 1

Ngeow, C. & Kanbur, S. M. 2009, in Astronomical Society of the Pacific Conference Series, Vol. 404, The Eighth Pacific Rim Conference on Stellar Astrophysics: A Tribute to Kam-Ching Leung, ed. B. Soonthornthum, S. Komonjinda, K. S. Cheng, & K. C. Leung, 262

Riess, A. G., Fliri, J., & Valls-Gabaud, D. 2012, apj, 745, 156

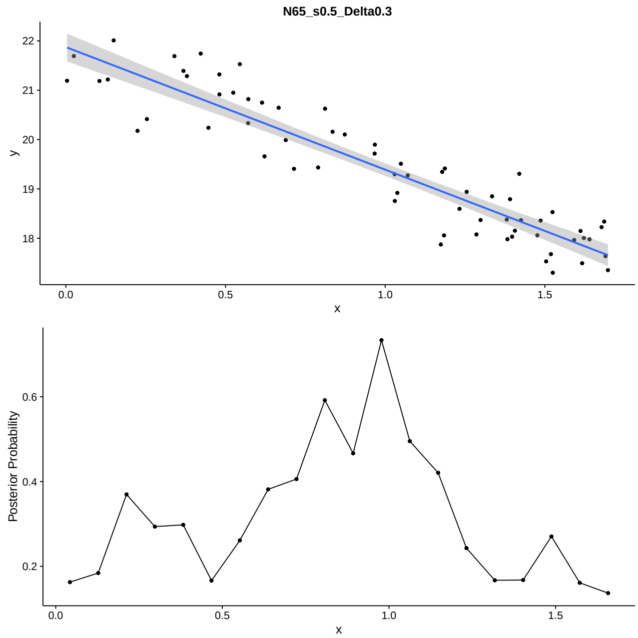

Fig. 5. Simulation for 65 observations with $\sigma = 0.3$ and $\Delta = 0.3$ (up) and associated PDF for a breaking point (down).

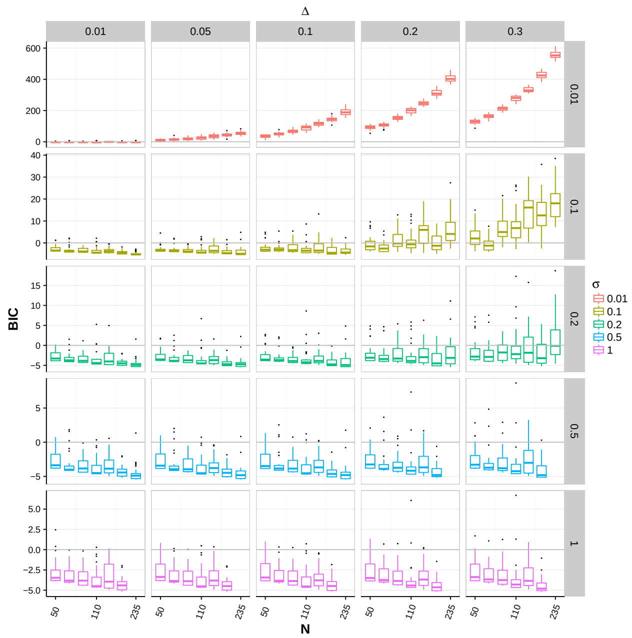

Fig. 6. Evolution of $BIC = BIC_r - BIC_f$ with the number N of observations given for each combination of added noise σ and slope difference Δ.

Sandage, A., Tammann, G. A., Saha, A., et al. 2006, apj, 653, 843

Tammann, G. A., Reindl, B., Thim, F., Saha, A., & Sandage, A. 2002, in Astronomical Society of the Pacific Conference Series, Vol. 283, A New Era in Cosmology, ed. N. Metcalfe & T. Shanks, 258

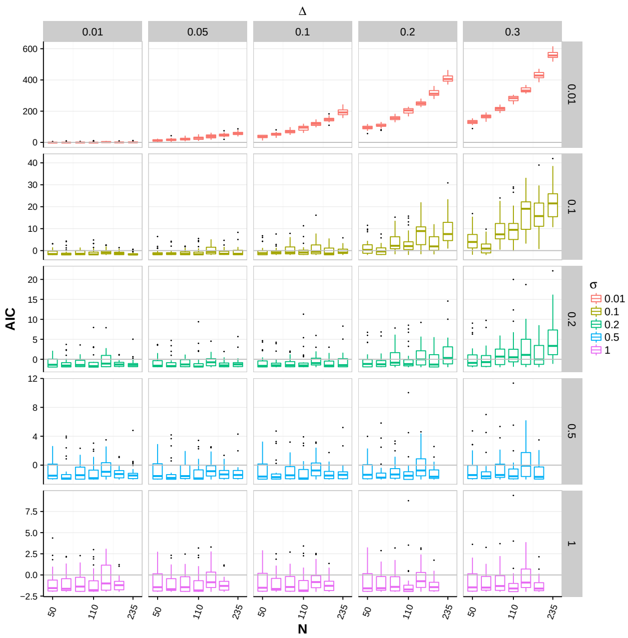

Fig. 7. Evolution of $AIC = AIC_r - AIC_f$ with the number N of observations given for each combination of added noise σ and slope difference Δ.