

**\*\*FULL TITLE\*\***

*ASP Conference Series, Vol. \*\*VOLUME\*\*, \*\*YEAR OF PUBLICATION\*\**

**\*\*NAMES OF EDITORS\*\***

## Statistical Models for Globular Cluster Luminosity Distribution

Max-Louis G. Buot

*Department of Mathematics & Computer Science, Xavier University,  
Cincinnati, Ohio 45207, USA*

Donald St. P. Richards

*Department of Statistics, The Pennsylvania State University, University  
Park, PA 16802; and SAMSI, Research Triangle Park, NC 27709, USA*

**Abstract.** We consider statistical models which have been proposed for luminosity distributions for the globular clusters in the Milky Way and M31. Although earlier research showed that the cluster luminosity functions in those two galaxies were well fit by Gaussian distributions, subsequent investigations suggested that their luminosities followed  $t$ - rather than Gaussian distributions. By applying the Bayesian Information Criterion, we find moderate statistical evidence that the  $t$ -distribution is superior to the Gaussian distribution as a model of luminosity distribution for the Milky Way. In the case of M31, we find moderate evidence that the Gaussian distribution is superior to the  $t$ -distribution. Our conclusion is that in neither case do we find strong evidence to support the use of one distribution over the other as a statistical model for the luminosities of the globular clusters in the Galaxy and M31. Moreover, we urge caution in the use of the Kolmogorov-Smirnov statistic to justify the choice of statistical models for globular cluster luminosity functions.

### 1. Introduction

Globular clusters are some of the oldest objects in the Universe, are innately luminous, and there is strong evidence that they are formed during periods of major star formation (Larsen & Richtler 1999). For these and other reasons, globular clusters are important in research on the formation of galaxies (Harris 2000; van den Bergh 2000).

The globular cluster luminosity function (GCLF) of a galaxy is the relative number of its globular clusters at a given luminosity. In the Milky Way, empirical evidence is that the GCLF is usually unimodal and nearly symmetric. It has also been observed that the peak of the luminosity function occurs at a magnitude which varies little from galaxy to galaxy (Harris 1998). For these reasons, it has been found that the GCLF peak is sometimes appropriate as a standard candle for distance measurement (Whitmore 1997).

In early work on GCLF, much attention was paid to plausible analytical forms of the GCLF. In several instances, including globular clusters in the Galaxy and M31, it was shown that a Gaussian distribution was a good analytical fit to the empirically observed distribution of luminosities (Racine and Shara, 1979; van den Bergh, 1985; Harris, *et al.*, 1991). Subsequently, a further analysis indicated that a  $t$ -distribution provided a better fit (Secker, 1992), and that

discovery has led to its wide acceptance in subsequent research on luminosity functions (Barmby, *et al.*, 2001).

As measured by maximum-likelihood procedures, the  $t$ -distribution has been shown to fit the empirical data more closely than the Gaussian distribution (Secker, 1992). However, this comparison is complicated by the fact that the analytical form of the Gaussian distribution is based on only two parameters, the mean and standard deviation, whereas the  $t$ -distribution has a third parameter, the index or shape parameter. This raises the issue of whether the increase in the the likelihood function for the  $t$ -distribution (over the Gaussian) may be caused by the presence of a third parameter.

The issue of goodness-of-fit of a hypothesized distribution *vis-a-vis* the number of parameters in the underlying analytical form of that distribution is precisely the *raison d'être* of the Bayesian Information Criterion (BIC). Simply put, the BIC is designed to ascertain the extent to which an improved fit is due to an increase in the number of parameters in the analytical form of the distribution.

In the sequel, we apply the BIC to the data provided by Secker (1992). We do not find overwhelming statistical evidence in support of the  $t$ -distribution over the Gaussian distribution, or *vice versa*, as a statistical model for the luminosity distribution of globular clusters in the Galaxy or M31. Moreover, we urge caution in the use of the Kolmogorov-Smirnov statistic as justification for the choice of statistical models for globular cluster luminosity distributions.

## 2. The Bayesian Information Criterion

Suppose that two statistical distributions are plausible models for an observed data set. These models may be fit to the data using a variety of statistical procedures, e.g., residual sums of squares, the method of moments, or the method of maximum likelihood.

In polynomial regression, for example, the residual sum of squares can be reduced simply by increasing the degree of the polynomial regression function. In general, the more complex the mathematical form of a model, the better a model will be seen to fit the data. Therefore, it is clear that the choice of a statistical model should not be assessed entirely by measures such as residual sums of squares or likelihood function values for, by increasing the number of parameters in the hypothesized model, we can obtain a reduction in the residual sums of squares or an increase in the values of the likelihood function.

The Bayesian Information Criterion (BIC) constitutes a standard approach to assessing the relative plausibility of two competing statistical models which are being fit to data with *large* sample sizes. To balance any difference in the number of free parameters between two statistical models, the BIC penalizes a model which has a larger number of free parameters. To illustrate this approach, consider the situation of two competing models,  $f_1(x; \theta_1, \dots, \theta_{m_1})$  and  $f_2(x; \phi_1, \dots, \phi_{m_2})$ . Here,  $\theta_1, \dots, \theta_{m_1}$  and  $\phi_1, \dots, \phi_{m_2}$  are parameters of the corresponding density functions  $f_1$  and  $f_2$ , respectively. On being given a random sample  $X_1, \dots, X_n$ , we construct the likelihood functions

$$L_1(\theta_1, \dots, \theta_{m_1}) = \prod_{i=1}^n f_1(x_i; \theta_1, \dots, \theta_{m_1})$$

and

$$L_2(\phi_1, \dots, \phi_{m_2}) = \prod_{i=1}^n f_2(x_i; \phi_1, \dots, \phi_{m_2})$$

Given the parameters  $\theta_1, \dots, \theta_{m_1}$  and  $\phi_1, \dots, \phi_{m_2}$ , and explicit formulas for the density functions  $f_1$  and  $f_2$ , the relative superiority of the first model over the second is measured by the

$$\text{BIC} = 2[\ln L_1(\theta_1, \dots, \theta_{m_1}) - \ln L_2(\phi_1, \dots, \phi_{m_2})] - (m_1 - m_2) \ln n.$$

The first term in this expression is a measure of the increase in the likelihood function values of the first model over the second, and the second term is a penalty term reflecting any difference in the numbers of parameters in the models. Thus, the BIC assesses any increase in the likelihood in light of the additional number of parameters necessary to achieve any such increase.

In practice, the values of  $\theta_1, \dots, \theta_{m_1}$  and  $\phi_1, \dots, \phi_{m_2}$  are unknown and need to be estimated from the data. Thus, we calculate the corresponding maximum likelihood estimates  $\hat{\theta}_1, \dots, \hat{\theta}_{m_1}$  and  $\hat{\phi}_1, \dots, \hat{\phi}_{m_2}$  and the estimated BIC,

$$\widehat{\text{BIC}} = 2[\ln L_1(\hat{\theta}_1, \dots, \hat{\theta}_{m_1}) - \ln L_2(\hat{\phi}_1, \dots, \hat{\phi}_{m_2})] - (m_1 - m_2) \ln n. \quad (1)$$

General rules for assessing the relative goodness-of-fit of the models  $f_1$  and  $f_2$  using the BIC are as follows (Jeffreys, 1961, Appendix B; Kass & Raftery, 1995; Mukherjee, *et al.*, 1998):

- $\widehat{\text{BIC}} < 2$ : Weak evidence that Model 1 is superior to Model 2
- $2 \leq \widehat{\text{BIC}} \leq 6$ : Moderate evidence that Model 1 is superior to Model 2
- $6 < \widehat{\text{BIC}} \leq 10$ : Strong evidence that Model 1 is superior to Model 2
- $\widehat{\text{BIC}} > 10$ : Very strong evidence that Model 1 is superior to Model 2

We now apply these procedures to compare three statistical models for GCLF data.

### 3. Models for GCLF distribution in the Milky Way and in M31

Consider the following competing models for GCLF in the Galaxy: A Gaussian model (van den Bergh, 1985),

$$f_1(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right], \quad (2)$$

and a  $t$ -distribution model (Secker, 1992),

$$f_2(x; \mu, \sigma, \delta) = \frac{\Gamma((\delta + 1)/2)}{\sqrt{\pi\delta}\sigma\Gamma(\delta/2)} \left[1 + \frac{(x - \mu)^2}{\delta\sigma^2}\right]^{-(\delta+1)/2}, \quad (3)$$

and the allowable ranges of the parameters are  $-\infty < \mu < \infty$ ,  $\sigma > 0$ , and  $\delta > 0$ . In each model,  $\mu$  represents the population mean and  $\sigma$  the population standard deviation. In the case of the  $t$ -distribution model,  $\delta$  is a shape parameter.

Under the Gaussian model (2), the likelihood function corresponding to a random sample  $X_1, \dots, X_n$  is

$$\begin{aligned} L_1(\mu, \sigma) &= \prod_{i=1}^n f(X_i; \mu, \sigma) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \right]. \end{aligned}$$

It is well-known that the maximum likelihood estimator of the parameter  $\mu$  is  $\hat{\mu} = \bar{X}$ , the sample mean. Further, denoting by  $S$  the sample standard deviation, it is well-known that the maximum likelihood estimator of  $\sigma^2$  is  $\hat{\sigma}^2 = (n-1)S^2/n$ . Since  $(n-1)/n \approx 1$  for large values of  $n$  and because the estimator  $S^2$  has other desirable statistical properties, it is common practice to estimate  $\sigma^2$  by  $S^2$ . Using the 100 observations comprising the Milky Way data (Secker, 1992, Table 1), it is found that the likelihood function  $L_1$  for the Gaussian model is maximized at  $\mu = -7.14$ ,  $\sigma = 1.41$  (*loc. cit.*, p. 1475), and then we obtain (*loc. cit.*, p. 1476)

$$\ln L_1(-7.14, 1.41) = -176.4 \quad (4)$$

In the case of the  $t$ -distribution model (3), the likelihood function is

$$\begin{aligned} L_2(\mu, \sigma, \delta) &= \prod_{i=1}^n f_2(X_i; \mu, \sigma, \delta) \\ &= \prod_{i=1}^n \frac{\Gamma(\frac{\delta+1}{2})}{\sqrt{\pi\delta} \sigma \Gamma(\frac{\delta}{2})} \left[ 1 + \frac{(X_i - \mu)^2}{\delta\sigma^2} \right]^{-\frac{\delta+1}{2}} \end{aligned}$$

Unlike the Gaussian case, no algebraic formulas for  $\hat{\mu}$ ,  $\hat{\sigma}$ , or  $\hat{\delta}$  are available in this case, and therefore the maximization of  $L_2$  and the corresponding likelihood estimates must be obtained numerically. Again calculating with the Milky Way data, Secker (1992, p. 1476) reports the maximum likelihood estimates  $\hat{\mu} = -7.31$ ,  $\hat{\sigma} = 1.03$ ,  $\hat{\delta} = 3.55$ , and the corresponding value of  $L_2$  is given by

$$\ln L_2(-7.31, 1.03, 3.55) = -173.0 \quad (5)$$

The maximum likelihood calculations in (4) and (5) suggest that the Gaussian model (2) is inferior to the  $t$ -model as a distribution for GCLF data in the Galaxy. However, there remains the issue of whether the increase in likelihood in (5) over (4) is due to the larger number of parameters in (3).

On applying the estimated BIC in (1) with the values in (4) and (5) and with  $m_1 = 2$ ,  $m_2 = 3$ , and  $n = 100$ , we obtain

$$\begin{aligned} \widehat{\text{BIC}} &= 2[\ln L_1(-7.14, 1.41) - \ln L_2(-7.31, 1.03, 3.55)] + \ln 100 \\ &= 2[-176.4 + 173.0] + 4.6 \\ &= -2.2 \end{aligned}$$

Since  $\widehat{\text{BIC}} < 2$  then, by the general rules on applications of the BIC for model comparison, we have moderate evidence that the  $t$ -distribution is superior to the Gaussian distribution as a model for the Galactic data. In particular, we do not find overwhelming evidence in support of one model over the other.

In the case of 82 globular clusters from M31 (Secker, 1992, Table 2), similar maximum likelihood calculations (*loc. cit.*, p. 1478) for the Gaussian model lead to the estimates  $(\widehat{\mu}, \widehat{\sigma}) = (16.98, 0.99)$  and the corresponding likelihood function value is given by

$$\ln L_1(16.98, 0.99) = -115.4$$

In the case of the  $t$  model, the parameters are found to have maximum likelihood estimates  $(\widehat{\mu}, \widehat{\sigma}, \widehat{\delta}) = (17.0, 0.90, 11.02)$  (*loc. cit.*, 1992, p. 1478), and the corresponding likelihood function value is

$$\ln L_2(17.0, 0.90, 11.02) = -115.1$$

With  $m_1 = 2$ ,  $m_2 = 3$ , and  $n = 82$ , the outcome of the BIC calculation is

$$\begin{aligned} \widehat{\text{BIC}} &= 2[\ln L_1(16.98, 0.99) - \ln L_2(17.0, 0.90, 11.02)] + \ln 82 \\ &= 2[-115.4 + 115.1] + 4.4 \\ &= 3.8 \end{aligned}$$

Thus, we have moderate evidence that the Gaussian distribution is superior to the  $t$ -distribution as a statistical model for the luminosity distribution of globular clusters in M31. Again, we find no overwhelming evidence in support of one distribution over the other.

A Cauchy model was also examined as a possible statistical model for GCLF in the Galaxy and M31. Here, the analytical form of the density function is (Secker, 1992, p. 1474)

$$f_3(x; \mu, \sigma) = \frac{\sigma}{\pi[\sigma^2 + (x - \mu)^2]}, \quad (6)$$

where the parameter ranges are  $-\infty < \mu < \infty$  and  $\sigma > 0$ . The parameter  $\mu$  represents the median and  $\sigma$  a measure of the spread of the distribution. Here again, we apply maximum likelihood calculations (*loc. cit.*, pp. 1476, 1478) to compare the Cauchy with the Gaussian or  $t$ -distribution models. When the Gaussian or  $t$  models are compared to the Cauchy model as a fit to the Galaxy GCLF data, we obtain  $\widehat{\text{BIC}} > 20$  in both cases. When the Gaussian or  $t$  models are compared to the Cauchy model for the M31 data, we again obtain  $\widehat{\text{BIC}} = 17.4$  in the Gaussian case and 24.2 in the  $t$  case. In all instances, there is strong or very strong evidence that the Gaussian or  $t$  distributions are superior to the Cauchy distribution as a fit to the GCLF data.

#### 4. Concluding Remarks

The BIC is only one of many procedures for testing the goodness-of-fit of a statistical model and, as with any procedure, it should be used carefully. In

particular, it is not a panacea, and perhaps should be applied in conjunction with other information criteria.

There are also drawbacks with the BIC. Findley (1991) notes that under specific circumstances, the BIC will tend to select the model which has fewer parameters. Precisely, Findley proves that “if the log-likelihood-ratio sequence of two models with different numbers of estimated parameters is bounded in probability then the BIC will, with asymptotic probability 1, select the model having fewer parameters.” At least heuristically, this result can be deduced *via* (1), as follows: Suppose that  $m_1 < m_2$  and suppose also that the first term in (1) is bounded above and below by universal constants, i.e., constants which do not depend on  $n$ . Then, as  $n \rightarrow \infty$ , the value of  $\widehat{\text{BIC}}$  in (1) is asymptotically equal to  $(m_2 - m_1) \ln n$ , a number which will be large and positive. By the general rules for application of the BIC, we would infer strong evidence in favor of  $f_1(x; \theta_1, \dots, \theta_{m_1})$ , the model having the fewer number of parameters.

In the context of astrophysical applications of comparing two statistical models,  $f_1(x; \theta_1, \dots, \theta_{m_1})$  and  $f_2(x; \phi_1, \dots, \phi_{m_2})$ , for globular cluster luminosity functions, an application of Findley’s theorem requires a sequential calculation of the corresponding log-likelihood ratios

$$\ln \frac{L_1(\hat{\theta}_1, \dots, \hat{\theta}_{m_1})}{L_2(\hat{\phi}_1, \dots, \hat{\phi}_{m_2})}$$

for successively increasing sample sizes  $n$ . If it is *believed* that the observed sequence of log-likelihood ratios will be bounded between two universal constants, and such a belief necessarily will have to be justified on intrinsic astrophysical arguments, then it be necessary to look for alternatives to the BIC in measuring the relative plausibility of a statistical model.

In the case of GCLF model fitting, it is important to bear in mind the fact that the BIC is *consistent*: If a given statistical model is the true model for a given data set then, as the sample size  $n \rightarrow \infty$ , the BIC will determine that the given model is correct (Jeffreys, 1961).

There is also the matter of the Kolmogorov-Smirnov statistic as a procedure for measuring the goodness-of-fit of a statistical model. In the case of the Galaxy and M31 data considered in this paper, the Kolmogorov-Smirnov statistic was also used (Secker, 1992, p. 1467 *ff.*) to support the choice of the  $t$  distribution over the Gaussian and Cauchy models. However, here again, there is cause for concern. The importance of statistics based on empirical distribution functions, such as the Kolmogorov-Smirnov statistic, stems from the fact that they are distribution-free in the case of continuous data, such as luminosity measurements. However, *these statistics are no longer distribution-free when the model parameters need to be estimated from the data*. Consequently, in model fitting contexts, goodness-of-fit levels of significance derived from the Kolmogorov-Smirnov statistic are usually incorrect when applied with estimated parameters (Babu and Rao, 2004; Babu and Feigelson, 2006).

We urge that caution be exercised in using the Kolmogorov-Smirnov statistic to justify the fitting of the models (2), (3), or (6) to GCLF data.

**Acknowledgments.** This work was supported by NSF grant AST-0434234. We are grateful Mercedes Richards for her comments on earlier drafts of the manuscript.

## References

- Babu, G. J. & Rao, C. R. (2004). Goodness-of-fit tests when parameters are estimated. *Sankhyā*, **66**:63-74
- Babu, G. J. & Feigelson, E. D. (2006). Astrostatistics: Goodness-of-fit and all that! In: ASP Conf. Ser. 351, *Astronomical Data Analysis Software and Systems XV*, ed. Gabriel, C., Arviset, C., Ponz, D. & Enrique, S. (San Francisco: ASP), 127-136
- Barmby, P., Huchra, J. P., & Brodie J. P. (2001). The M31 globular cluster luminosity function. *Astronom. J.*, **121**:1482-1496
- Findley, D. F. (1991). Counterexamples to parsimony and BIC. *Ann. Inst. Statist. Math.*, **43**:505-514
- Harris, W. E., Allwright, J. W. B., Pritchett, C. J., & van den Bergh, S. (1991). The luminosity distribution of globular clusters in three giant Virgo ellipticals. *Astrophys. J. Suppl.*, **76**:115-151
- Harris, W. E. (198). Globular cluster systems as distance indicators. In: ASP Conf. Ser. 4, *The Extragalactic Distance Scale*, ed. C. Pritchett & S. van den Bergh (San Francisco: ASP), 231-254
- Harris, W. E. (2000). Globular cluster systems. *Star Clusters, Saas-Fee Advanced Course 28*. Lecture Notes 1998, (Berlin: Springer-Verlag)
- Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. (Oxford: Clarendon Press)
- Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.*, **90**:773-795
- Larsen, S. S. & Richtler, T. (1999). Young massive star clusters in nearby galaxies . I. Identification and general properties of the cluster systems. *Astron. & Astrophys.*, **345**:59-72
- Mukherjee, S., Feigelson, E. D., Babu, G. J., Murtagh, F., Fraley, C., & Raftery, A. (1998). Three types of gamma-ray bursts. *Astrophys. J.*, **508**:314-327
- Racine, R. & Shara, M. (1979). The luminosity distribution of globular clusters in M31. *Astronom. J.*, **84**:1694-1696
- Secker, J. (1992). A statistical investigation into the shape of the globular cluster luminosity distribution. *Astronom. J.*, **104**:1472-1481
- van den Bergh, S. (1985). The luminosity function of globular clusters. *Astrophys. J.*, **297**:361-364
- van den Bergh, S. (2000). Some musings on globular cluster systems. *Publ. Astronom. Soc. Pacific*, **112**:932-941
- Whitmore, B. C. (1997). Globular clusters as distance indicators. In: *The Extragalactic Distance Scale*, ed. M. Livio, M. Donahue, & N. Panagia. (Cambridge: Cambridge Univ. Press), 254-272