DRAFT VERSION FEBRUARY 19, 2010
Preprint typeset using LaTeX style emulateapj v. 2/16/10

QUANTIFYING BIMODALITY

OLEG Y. GNEDIN
ognedin@umich.edu
Draft version February 19, 2010

ABSTRACT

Quantifying whether a distribution is better described by one or two modes is still an unsolved problem in statistics. While there are algorithms that split the input distribution into two modes or assign probabilities that a given data point belongs to either of the two modes, there is no proper statistic that evaluates whether such a split is preferred to a unimodal distribution. In this Guide, we describe our improvement of a popular KMM algorithm, as well as an independent test of bimodality based on the Dip statistic. If you use this code for a publication, please acknowledge the original paper:

A. L. Muratov & O. Y. Gnedin, 2010, ApJ, submitted, arXiv:1002.1325
"Modeling the Metallicity Distribution of Globular Clusters".

Subject headings:

1. GMM – A BETTER VERSION OF KMM

Ashman et al. (1994) popularized a mixture modeling code KMM for detecting bimodality in astronomical applications. This code has been widely used for globular cluster studies and can be considered a standard method in the field. The KMM algorithm assumes that an input sample is described by a sum of two Gaussian modes and calculates the likelihood of a given data point belonging to either of the two modes. It also calculates the likelihood ratio test (LRT) as an estimate of the improvement in going from one Gaussian to two Gaussian distributions. However, the LRT obeys a standard χ^2 statistic only when the two modes have the same width (variance), which may not be satisfied by real datasets. Even though the probability of the LRT can be estimated using bootstrap in principle, in practice the use of the KMM code has been limited to common width modes (the so-called "homoscedastic" case). Brodie & Strader (2006) and Waters et al. (2009) provide further discussion of KMM.

The KMM method belongs to a general class of algorithms of Gaussian mixture modeling (GMM). GMM methods maximize the likelihood of the data set given all the fitted parameters, using the expectation-maximization (EM) algorithm (e.g., Press et al. 2007). A major simplification, which allows one to derive explicit equations for the maximum likelihood (ML) estimate of the parameters, is that each mode is described by a Gaussian distribution.

For simplicity, and as appropriate for the metallicity distribution, we consider a univariate input data set. However, the algorithm is fully scalable to multivariate distributions. The likelihood function of a univariate sample x_n is

$$\mathcal{L}_K = \prod_n \left(\sum_{k=1}^{K} p_k N(x_n|\mu_k,\sigma_k) \right), \qquad (1)$$

where

$$N(x|\mu,\sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right] \qquad (2)$$

is the Gaussian density. The modal fractions are normalized as $\sum_k p_k = 1$. A unimodal distribution ($K = 1$) has two independent parameters (μ and σ), whereas a bimodal distribution has five parameters ($p_1, \mu_1, \sigma_1, \mu_2, \sigma_2$), since $p_2 = 1 - p_1$.

The power of the GMM method lies in its ability to determine the ML values of the parameters (p_k, μ_k, σ_k). The disadvantage is that the method will always split the data set into the specified number of modes, K. In order to detect bimodality it is extremely important to be able to judge whether the bimodal fit is an improvement over the unimodal fit. For this purpose the KMM code uses the LRT test, which appears to be an approximation derived by Wolfe (1971) for the homoscedastic case ($\sigma_1 = \sigma_2$). Define the ratio of the maximum likelihoods as $\lambda \equiv \mathcal{L}_{1,\max}/\mathcal{L}_{2,\max}$. According to numerical Monte Carlo studies of Wolfe (1971), the statistic $-2\ln\lambda$ approximately obeys the χ^2 distribution with a number of degrees of freedom equal to "twice the difference between the number of parameters of the two models under comparison, not including the mixing proportions" (McLachlan 1987). This is the χ^2_2 distribution in our case. However, the statistic does not apply in the heteroscedastic case $\sigma_1 \neq \sigma_2$ (it would have been χ^2_4). Note that this unusual number of degrees of freedom was found as an empirical approximation. Unfortunately, no exact estimation exists for the goodness of modal split.

Several variations of the method have been suggested in the literature. McLachlan (1987) proposed a *parametric bootstrap* to test for the number of components. In this method, a test sample is drawn randomly from a unimodal Gaussian distribution with the parameters $\{\mu, \sigma\}$ best-fitting the input sample. The number of objects in the test sample is taken the same as in the input sample. The bimodal split is calculated for this test sample using the EM algorithm and the likelihood ratio $\lambda_{\rm boot}$ is saved. Repeating the bootstrap a large number of times, we obtain the probability of randomly drawing the ratio as large as that observed in the input sample, $\lambda_{\rm obs}$. If the probability is below a few percent, we reject the null hypothesis that the input sample belongs to a unimodal Gaussian.

However, the parametric bootstrap is not a perfect solution. In the limit of a large number of objects in the input sample, the likelihood function is very sensitive to outliers far from the center of the distribution. Simple measurements errors in the wings of the Gaussian function may cause a unimodal distribution to be rejected, even if it is correct. In other words, GMM is more a test of Gaussianity than of unimodality (see Muthén 2003; Bauer 2007, for more discussion).

Lo et al. (2001) proposed a modified LRT method to test for the true number of components of a Gaussian mixture. The modified statistic must be evaluated numerically, but still does not address the problem with Gaussian wings. Subsequently, Lo (2008) suggested to use the standard LRT with the parametric bootstrap to test for heteroscedastic split, and also suggested restricting the ratio of the standard deviations of the two modes to be not less than 0.25, to avoid numerical artifacts. Such a method was recently implemented in globular cluster studies by Waters et al. (2009).

The sensitivity of LRT to the assumption of Gaussian distribution calls for additional, independent tests of bimodality. A useful and intuitive statistic is the separation of the means relative to their widths:

$$D \equiv \frac{|\mu_1 - \mu_2|}{\left[(\sigma_1^2 + \sigma_2^2)/2\right]^{1/2}}. \quad (3)$$

We use the factor $\sqrt{2}$ for consistency with the definition in Ashman et al. (1994), who noted that $D > 2$ is required for a clean separation between the modes. If the GMM method detects two modes but they are not separated enough ($D < 2$), then such a split is not meaningful. The power of GMM in this case is counterproductive. A histogram of such a distribution would show no more than two little bumps, which would not be recognized as distinct populations.

Another simple statistic is the kurtosis of the input distribution. A positive kurtosis corresponds to a sharply peaked distribution, such as the Eiffel Tower. A negative kurtosis corresponds to a flattened distribution, such as a top hat. A sum of two populations, not necessarily Gaussians, is broader than one population and therefore has a significantly negative kurtosis. However, $kurt < 0$ is a necessary but not sufficient condition of bimodality. A broad unimodal distribution, such as an actual top hat, also has negative kurtosis. Therefore, $kurt < 0$ is only useful as an additional check to support the results of LRT and the D-value.

In order to provide a more robust measure of the modal split, we have revised and implemented the GMM algorithm independently of the KMM code. We begin with a single run of the EM algorithm to calculate the means and standard deviations assuming a heteroscedastic bimodal distribution. Then we repeat the estimation assuming a unimodal Gaussian case. We take the ratio of the likelihoods λ, the separation D, and the kurtosis as the three statistics of interest. We then estimate the error distribution for the modal parameters using non-parametric bootstrap (drawing from the input sample with repetitions) of 100 realizations. We also run the parametric bootstrap to assess the confidence level at which a unimodal distribution can be rejected based on each of the three statistics. Its practical application for an input sample of more than 100 objects is limited to about 1000 bootstrap realizations, limiting the confidence level to $\sim 10^{-3}$. A sufficiently low probability of each statistic means a unimodal distribution can be rejected in favor of a bimodal distribution. The code also calculates the probability of each data point belonging to either mode.

A sum of two Gaussians with the same variance can sometimes be preferred to the case of different variances, because of the one fewer degree of freedom. The choice between the homoscedastic and heteroscedastic cases can be similarly made using LRT, but we feel that it is less important than choosing between a bimodal and unimodal distributions. For comparison with the KMM code, we calculate the homoscedastic split and its approximate probability using the χ^2_2 statistic. We also calculate an alternative split into two Gaussian modes with the same mean but different variance. This is a more extended distribution than a single Gaussian, which may be a better fit for a unimodal but non-Gaussian sample.

The steps of our algorithm are summarized below:

1. Calculate $\{\mu, \sigma\}$ and $\{p_1, \mu_1, \sigma_1, \mu_2, \sigma_2\}$ using a single EM run.

2. Form three statistics: λ, D, and $kurt$.

3. Run non-parametric bootstrap to estimate the errors: $\Delta\mu_k$, $\Delta\sigma_k$, Δp_1, and ΔD.

4. Run parametric bootstrap to estimate the probability of a unimodal distribution, according to λ, D, and $kurt$.

As a first test of the algorithm, we verified that it reproduces exactly the test output of the KMM code, given the test input provided with the code.

We also made two random realizations of a unimodal Gaussian distribution, $N(0,1)$, with 150 objects and 1500 objects, respectively. The smaller sample has the mean and standard deviation of $\mu = -0.088$ and $\sigma = 0.987$, within the intended target given the sample size. Indeed, a non-parametric bootstrap gives $\Delta\mu = 0.084$, $\Delta\sigma = 0.063$. The kurtosis of the input sample is $kurt = 0.104$. A heteroscedastic split gives two peaks, by construction, with $\mu_1 = -0.483$, $\sigma_1 = 1.206$ and $\mu_2 = -0.032$, $\sigma_2 = 0.938$. However, the split is not statistically significant. The likelihood is improved only by $-2\ln\lambda = 0.26$ relative to the unimodal case, which gives the probability better than 99% that the input sample is unimodal. The parametric bootstrap gives a similar probability of 96%. The separation of the peaks also leads to the same conclusion: $D = 0.42 \pm 1.46$. The parametric bootstrap probability of drawing such value of D randomly from a unimodal distribution is 87%. The probability of drawing the measured kurtosis is 74%. Thus, all three statistics show correctly that the input distribution is not bimodal. The larger test sample has smaller parameter errors, as expected, but similar significance levels from the parametric bootstrap.

We then apply the GMM algorithm to the sample of observed metallicities of the Galactic globular clusters. A unimodal fit gives $\mu = -1.298 \pm 0.049$ and $\sigma = 0.562 \pm 0.028$, where the errors are calculated with the non-parametric bootstrap. A heteroscedastic split gives $\mu_1 = -1.608 \pm 0.064$, $\sigma_1 = 0.317 \pm 0.051$ and $\mu_2 = -0.583 \pm 0.074$, $\sigma_2 = 0.281 \pm 0.075$. Of the total number of 148 clusters, 103 (or 70%) are in the metal-poor group and 45 (or 30%) are in the metal-rich group. A homoscedastic split gives $\mu_1 = -1.620 \pm 0.037$, $\mu_2 = -0.608 \pm 0.055$, and $\sigma_1 = \sigma_2 = 0.303 \pm 0.026$. In this case, there are 101 metal-poor clusters and 47 metal-rich clusters. In either case, the likelihood improvement in the 1000 parametric bootstrap realizations is never as high as observed, $-2\ln\lambda = 27.5$. That is, a unimodal distribution is rejected at a confidence level better than 0.1%. The separation of the peaks is also very clear, $D = 3.42 \pm 0.47$. The observed cluster distribution is indeed bimodal!

To install GMM, untar the distribution and type `make`. It will create an executable `gmm`. To run GMM on the Galactic sample, included in the file `obs.in`, type

```
gmm obs.in 0 -1 0
```

The first command line argument is the file name, the second is 0 for different variances or 1 for same variances, and the other arguments are the approximate means of the modes/peaks. The number of arguments determines precisely how many modes GMM will calculate. In this case, it will look for two modes, centered near -1 and 0. The algorithm automatically finds the means, so their initial values are not important.

Given the input, GMM will produce the following output:

```
Gaussian Mixture Model of a univariate sample
looking for 2 peaks with different variances
number of data points = 148  kurtosis = -0.691
...running unimodal Gaussian
iter=2 err=0.0e+00:  peak=-1.298 (n=148.0 sig=0.562)  logL1=-124.725
...running Gaussian mixture with different variances
iter=45 err=7.0e-07: peak1=-1.608 (n=103.3 sig=0.317) peak2=-0.583 (n=44.7 sig=0.281) logL=-110.967
Chi-square statistic (null=unimodal): chi2=27.52 Ndof=4 p=1.56e-05
Peak separation DD = 3.42
...running Gaussian mixture with same variances
iter=24 err=9.9e-07: peak1=-1.620 (n=100.9 sig=0.303) peak2=-0.608 (n=47.1 sig=0.303) logL=-111.115
Chi-square statistic (null=unimodal): chi2=27.22 Ndof=2 p=1.23e-06
Peak separation DD = 3.34
...running Gaussian mixture with same means and different variances
iter=35 err=8.5e-07: peak1=-1.298 (n=145.4 sig=0.562) peak2=-1.298 (n= 2.6 sig=0.562) logL=-124.725
Chi-square statistic (null=equal means): chi2=27.52 Ndof=1 p=1.56e-07
Chi-square statistic (null=equal variances): chi2=0.30 Ndof=1 p=5.86e-01
...running bootstrap to estimate errors of best-fit parameters
Bootstrap unimodal: mean = -1.297 +- 0.049  sig = 0.562 +- 0.028
Bootstrap peak1: mean = -1.596 +- 0.064  sig = 0.329 +- 0.051  n = 104.5 +- 9.3
Bootstrap peak2: mean = -0.581 +- 0.074  sig = 0.268 +- 0.075  n = 43.5 +- 9.3
Bootstrap DD = 3.37 +- 0.47
...running parametric bootstrap to rule out unimodal distribution
Parametric bootstrap: p(chi2) < 0.001
Parametric bootstrap: p(DD) = 0.199
Parametric bootstrap: p(kurt) = 0.02
summary -1.608 0.064 -0.583 0.074 0.317 0.051 0.281 0.075 148 0.302 0.063 3.42 0.47 0.001 0.199 0.020 obs.in
```

The last line summarizes the important statistics for convenient parsing by a shell script.

The code also writes a file `peakprob.out`, which contains the assigned probabilities of each data point belonging to either of the modes. The first line repeats the best-fit parameters $\mu_1, \sigma_1, p_1, \mu_2, \sigma_2, p_2$ (and so on if more modes are requested). The other n lines give the probabilities $p_{1,n}$, $p_{2,n}$, etc. ($\sum_{k=1}^{K} p_{k,n} = 1$) and the value of x_n. These data can be used to separate the objects into their most likely mode.

2. DIP TEST

A completely independent test of unimodality was proposed by Hartigan & Hartigan (1985). It was first used for globular cluster studies by Gebhardt & Kissler-Patig (1999). The Dip test is based on the cumulative distribution of the input sample. The Dip statistic is the maximum distance between the cumulative input distribution and the best-fitting unimodal distribution. In some sense, this test is similar to KS test but the Dip test searches specifically for a flat step in the cumulative distribution function, which corresponds to a "dip" in the histogram representation. The probability of rejecting a unimodal distribution is calculated empirically and tabulated as a function of sample size. We obtained an updated table of the probabilities, `dip_tab.txt`, calculated recently by Martin Maechler (www.cran.r-project.org/web/packages/diptest).

We have added a driver routine to the original Fortran code of Hartigan & Hartigan (1985). Our code interpolates the probability table for any input sample size up to 5000 objects. Looking just at the significance levels, the Dip test appears less powerful than GMM. The Dip probability of the observed Galactic sample being bimodal is 90%, whereas the LRT probability is 99.998% and the parametric bootstrap probability is 99.9%. However, the Dip test has the benefit of being insensitive to the assumption of Gaussianity and is therefore a true test of modality. It is also much faster to run than the GMM code.

To install Dip, untar the distribution and type `make dip`. It will create an executable `dip`. To run Dip test on the Galactic sample, included in the file `obs.in`, type

```
dip 148 obs.in
```

The first command line argument is the number of objects in the file, the second is the file name. The input sample must be sorted in the increasing order. Given the input, Dip will produce the following output:

```
148  0.0390332602  0.896628618
```

The first number repeats the number of objects, the second is the value of Dip statistic (which you can ignore), the third is the significance level with which a unimodal distribution can be rejected.

REFERENCES

Ashman, K. M., Bird, C. M., & Zepf, S. E. 1994, AJ, 108, 2348
Bauer, D. J. 2007, Multivariate Behavioral Research, 42, 757
Brodie, J. P. & Strader, J. 2006, ARA&A, 44, 193
Gebhardt, K. & Kissler-Patig, M. 1999, AJ, 118, 1526
Hartigan, J. A. & Hartigan, P. M. 1985, The Annals of Statistics, 13, 70
Liddle, A. R. 2009, Annual Review of Nuclear and Particle Science, 59, 95
Lo, Y. 2008, Statistics and Computing, 18, 233
Lo, Y., Mendell, N. R., & Rubin, D. B. 2001, Biometrika, 88, 767
McLachlan, G. J. 1987, Journal of the Royal Statistical Society. Series C (Applied Statistics), 36, 318
Muthén, B. 2003, Psychological Methods, 8, 369
Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 2007, Numerical recipes. The art of scientific computing, 3rd ed. (Cambridge: University Press)
Waters, C. Z., Zepf, S. E., Lauer, T. R., & Baltz, E. A. 2009, ApJ, 693, 463
Wolfe, J. H. 1971, in Technical Bulletin STB 72-2 (San Diego: U.S. Naval Personnel and Training Research Laboratory)