New Methods of Mixture Model Cluster Analysis

James Wicker
Editor/Researcher

National Astronomical Observatory of China

A little about me

I came to Beijing in 2007 to work as a postdoctoral researcher at the National Astronomical Observatory of China, part of the Chinese Academy of Sciences.

My research focused on developing new methods in statistical analysis which can overcome problems inherent in traditional methods.

Since 2009, my main job has been to be an editor for a research journal about astronomy, but I am continuing the research on new statistical methodology.

Plan for this talk

Introduce what mixture model cluster analysis is and how it is useful

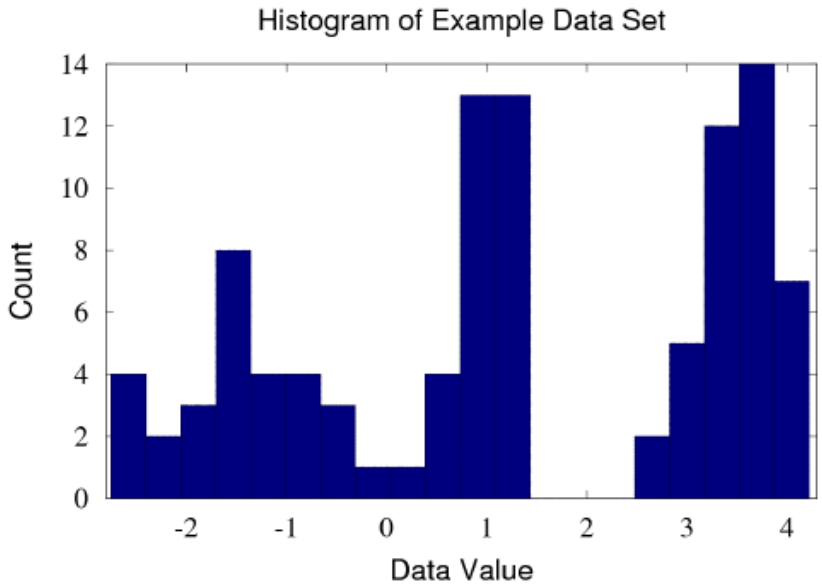
Describe some problems that plague traditional methods in statistical analysis, and show how new methods can overcome these handicaps

Look at at few examples using simulated and real data, highlighting the advantages of new methods

Questions and discussion

Univariate Mixture Modeling = Distributional Analysis

- In order to understand their data, many researchers study how their observations are distributed.
- How can we characterize the observed data set?
- How many distributions are present in the data?



Why is mixture modeling useful?

- Many processes and natural phenomena approximately follow sums of Gaussian distributions.
- Mixture modeling tries to find the underlying distributions that generate observed data.
- Use mathematical methods to calculate means and variances of distributions.
- Use statistical methodology to decide the number of distributions present in a data set.
- We can discuss these methods.

The Mixture Model Process

- The researcher first applies a distance minimization algorithm to the data set, like k-means.
- After the researcher has applied a distance minimization, the second step in the mixture modeling process is to calculate the Log-Likelihood of the mixture of Gaussian distributions.
- The goal of this step is to try to assign data points to clusters such that the Log-Likelihood is maximized, thus producing the Maximum Likelihood Estimate (MLE) of the data set.
- We can examine this process in more detail.

Traditionally use the Expectation-Maximization (EM) algorithm

- Start with mean, variance and mixture proportion values derived from k-means.
- Iteratively recalculate cluster assignments and parameter values until the change between iteration cycles is small.

$$l(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) =$$

$$\sum_{i=1}^{n} \log\left[\sum_{k=1}^{K} \pi_k (2\pi)^{-\frac{p}{2}} |\Sigma_k|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \mu_k)\right]\right]$$

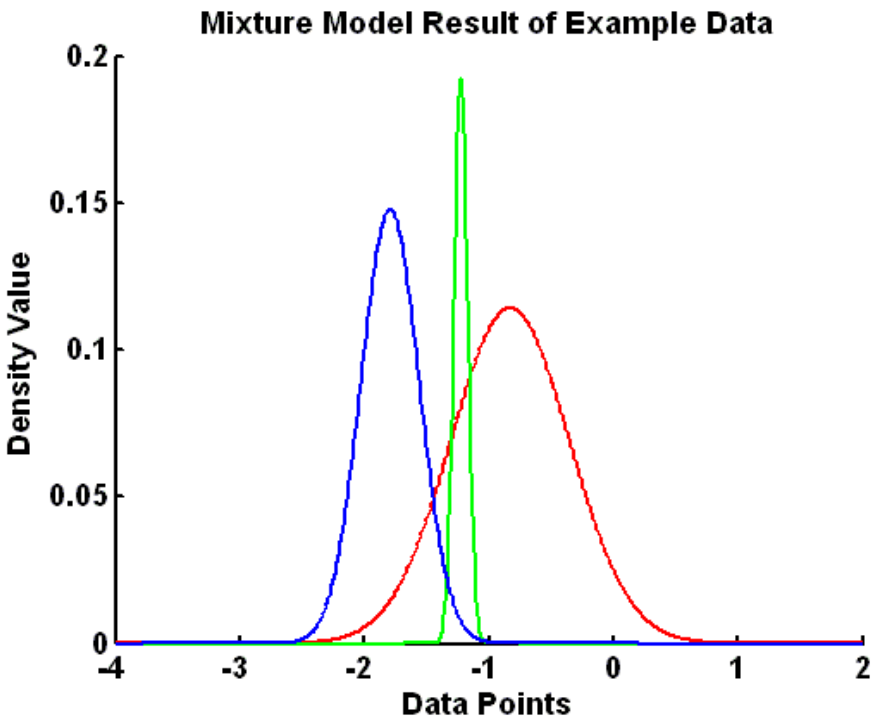
Problems with the EM algorithm approach

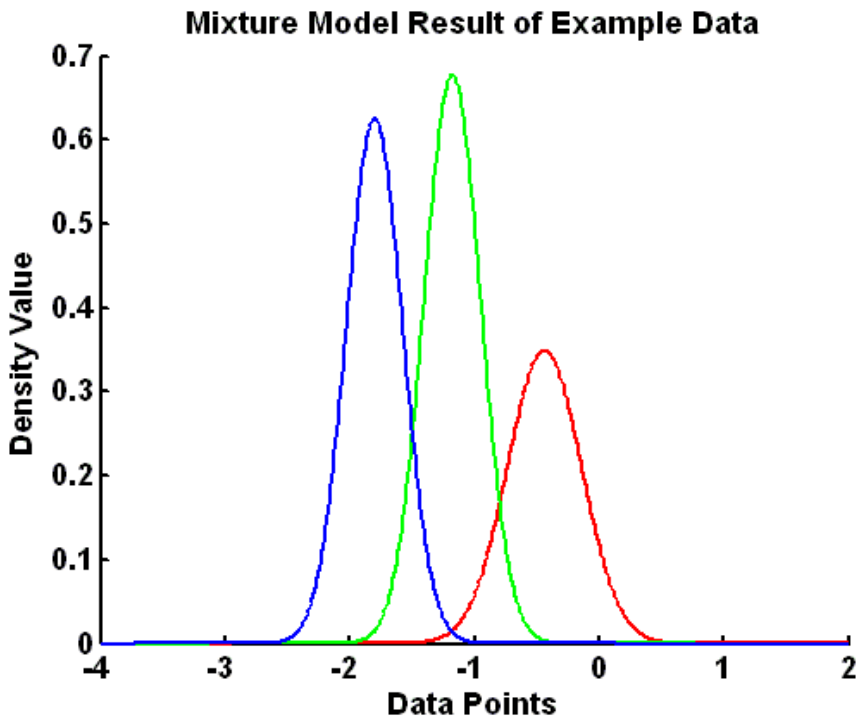
- The final outcome of the EM calculation depends on how the series sums are initialized.
- There is no guarantee that the series sum will converge for a given set of initial parameters.
- Authors admit that users should try different initializations.
- These problems are especially acute for overfitting situations.
- Due to these problems, researchers must call into question the reliability and repeatability of their cluster calculations derived from the EM algorithm.

Fundemental issues in Log-Likelihood calculations for mixture models

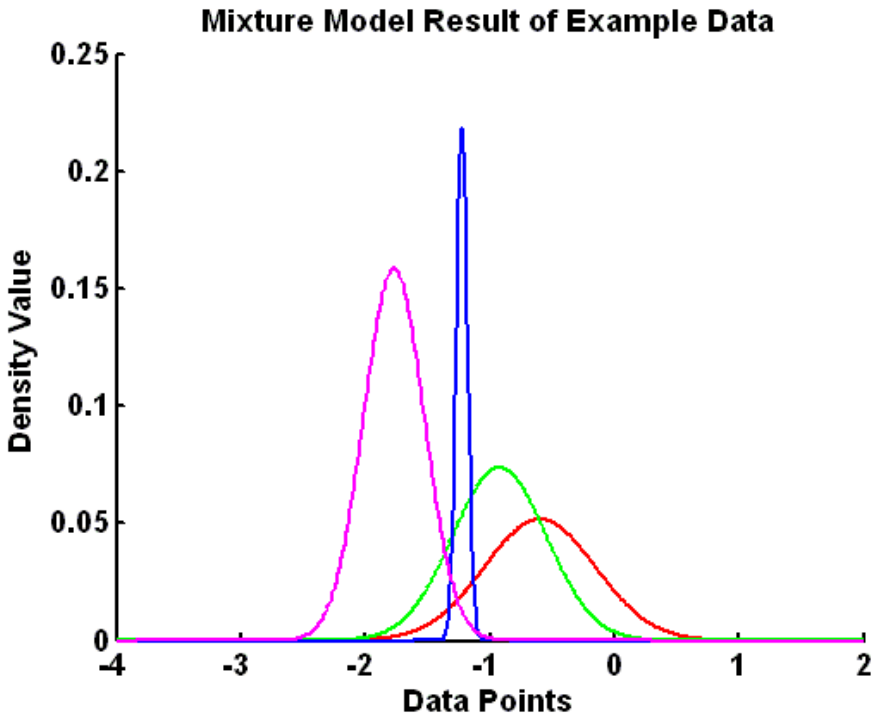
- In general, these Log-Likelihood values are infinite!
- In overfitting cases, when two means are near each other, the ratio of the variances in those clusters can become very large and thus difficult to calculate.
- Large ratios of variance values inflate the resulting Log-Likelihood sum, affecting subsequent steps in the mixture modeling process.
- How can someone compare competing model calculations with inflated Log-Likelihood values?
- These issues lead to ambiguity and inconsistent repeated calculation results.

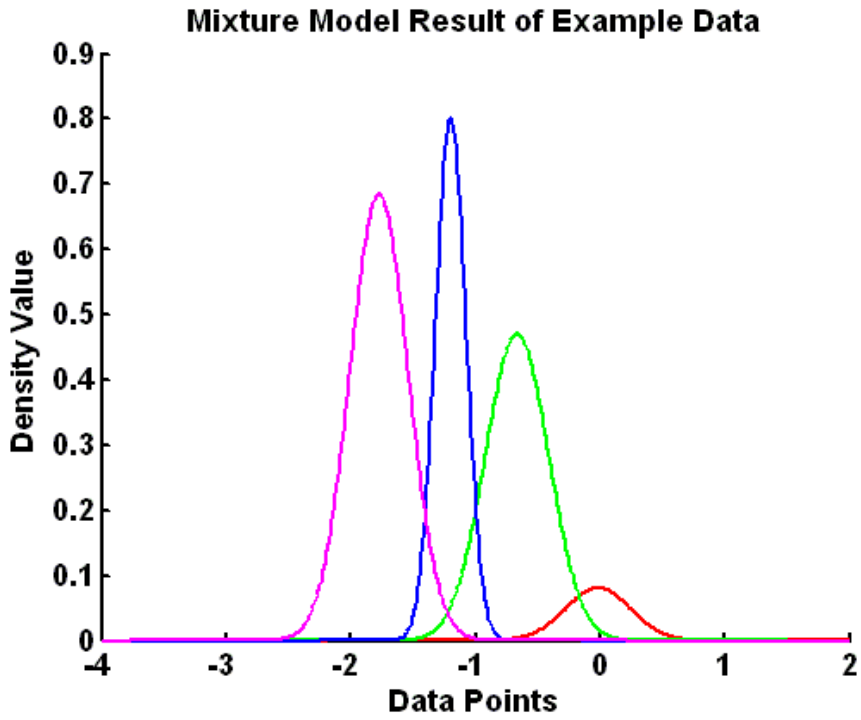
Restricting the range of variances controls the Log-Likelihood value





Another example of restricting the range of variances in the Log-Likelihood





Traditional vs. Modern Methods

In cluster analysis, traditional methods rely on "hypothesis tests" to decide the best number of clusters.

Test the hypothesis that there is one cluster vs. more than one cluster.

If more than one cluster, it is very difficult to unambiguously decide the best number of clusters in the data set due to arbitrary thresholds of acceptance and rejection of the hypothesis.

Modern Methods offer more flexibility

Since the 1970s, some researchers in statistical analysis have proposed “information scoring functions” as a way to judge competing models.

Overcome problem of setting arbitrary thresholds inherent in hypothesis-based statistical analysis.

The model that has the minimum information score is regarded as the best to describe the system under study.

The Information Theory approach has many advantages

- Information scoring overcomes the inherent subjectivity of hypothesis testing.
- It computes scores for different statistical models, and the best model is the minimum score to describe the phenomenon under study.
- Information scores include Baysian Information Criterion (BIC), Akaike Information Criterion (AIC) and ICOMP
- Some researchers in the field of information theory regard ICOMP (Bozdogan 1993) as the best because it is derived from considering the distributions present.

Information Scoring functions attempt to find a better balance

- Instead of relying on the ever-increasing sum of squared error term, AIC and ICOMP have two terms which counteract each other.
- The first term is Log-Likelihood, which accounts for error.
- The second term tries to protect from overfitting.

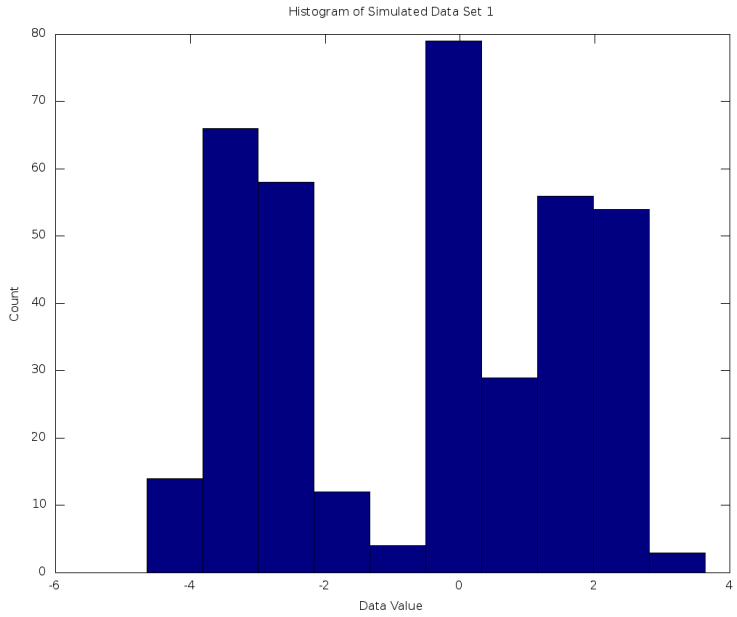
$$AIC = -2logL(\Theta_k) + 2m(k)$$

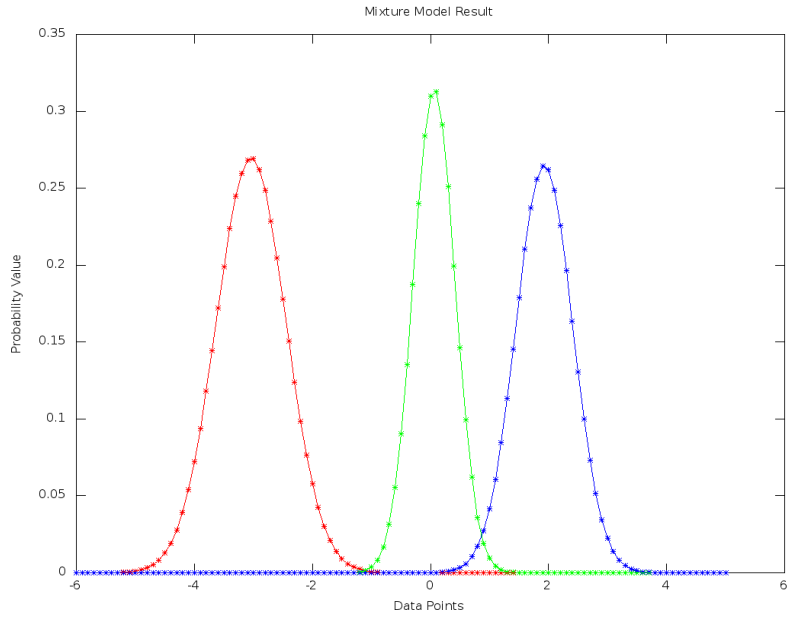
$$ICOMP = -2logL(\Theta_k) + 2C_1(\Sigma_{Model})$$

My work – Combining all three steps

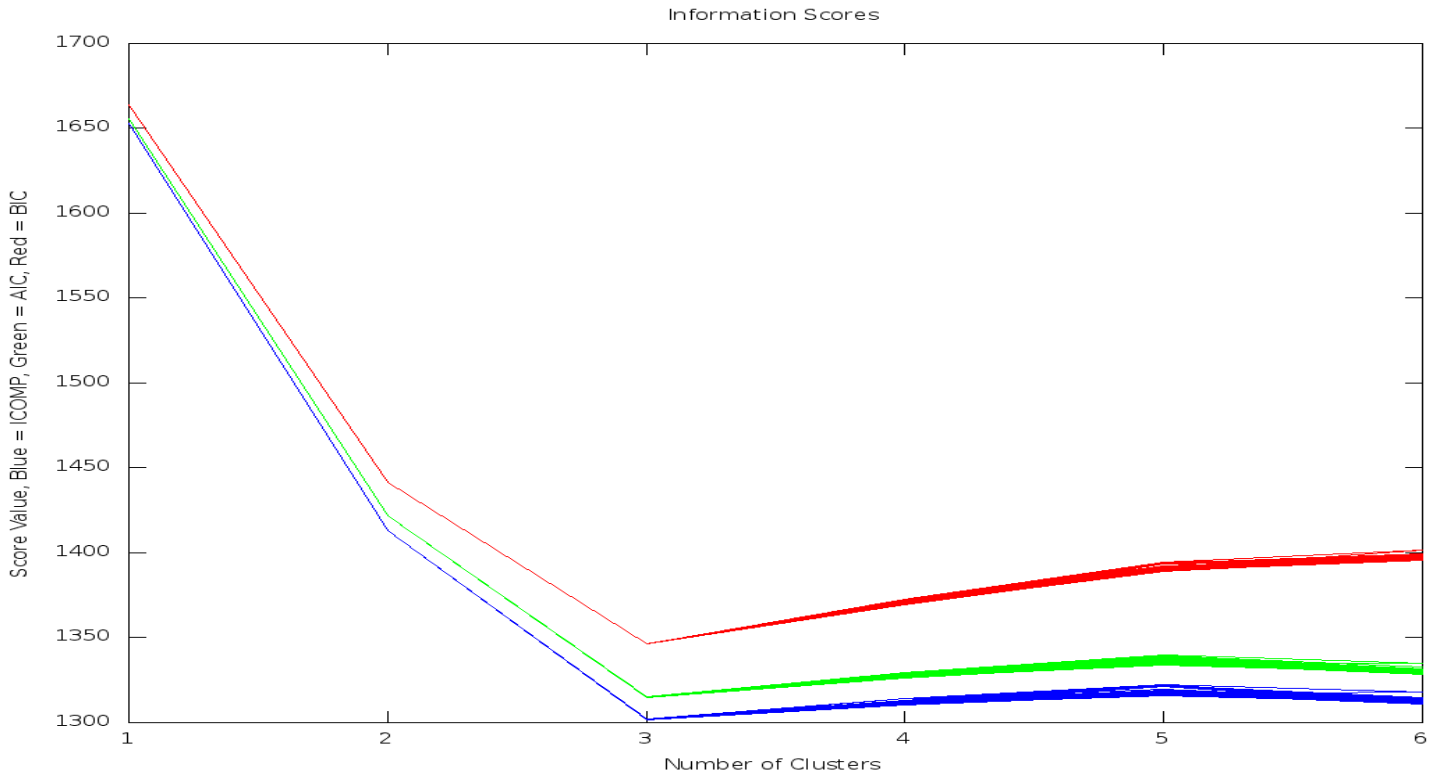
- I have been developing a new method to calculate Mixture Models with information scores.
- This uses a Genetic Algorithm to calculate the Log-Likelihood and hence MLE values for a given number of component distributions.
- My new method incorperates the variance ratio restriction which leads to finite solutions for maximizing Log-Likelihood values.
- We can see the output of some tests on simulated data.

Simulated data with three clusters

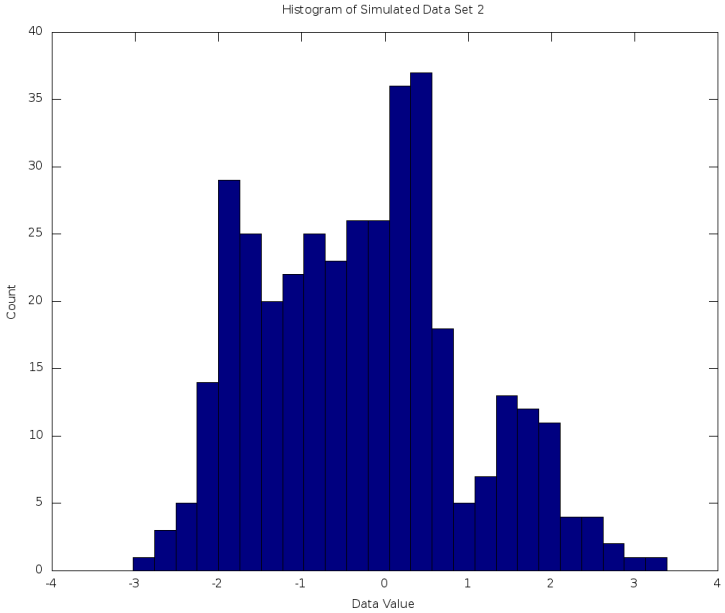


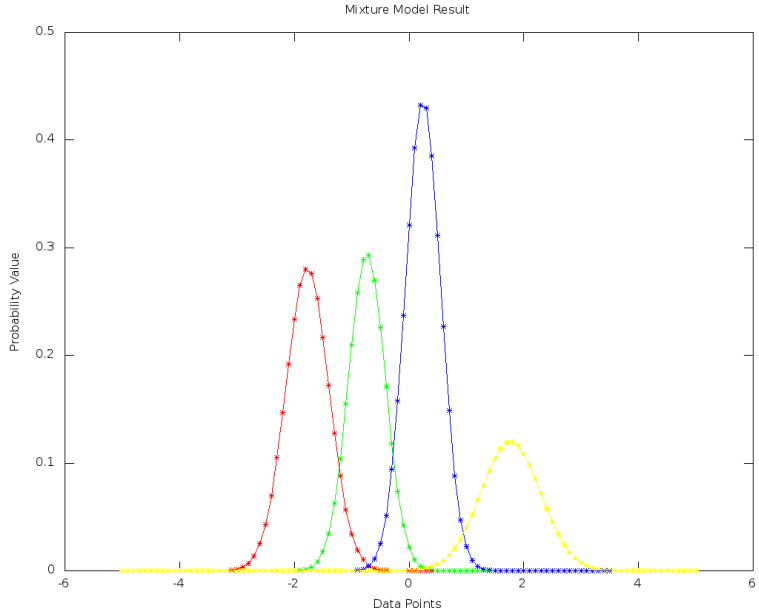


Information Scoring Results of 100 trials of Simulated Data set with three clusters

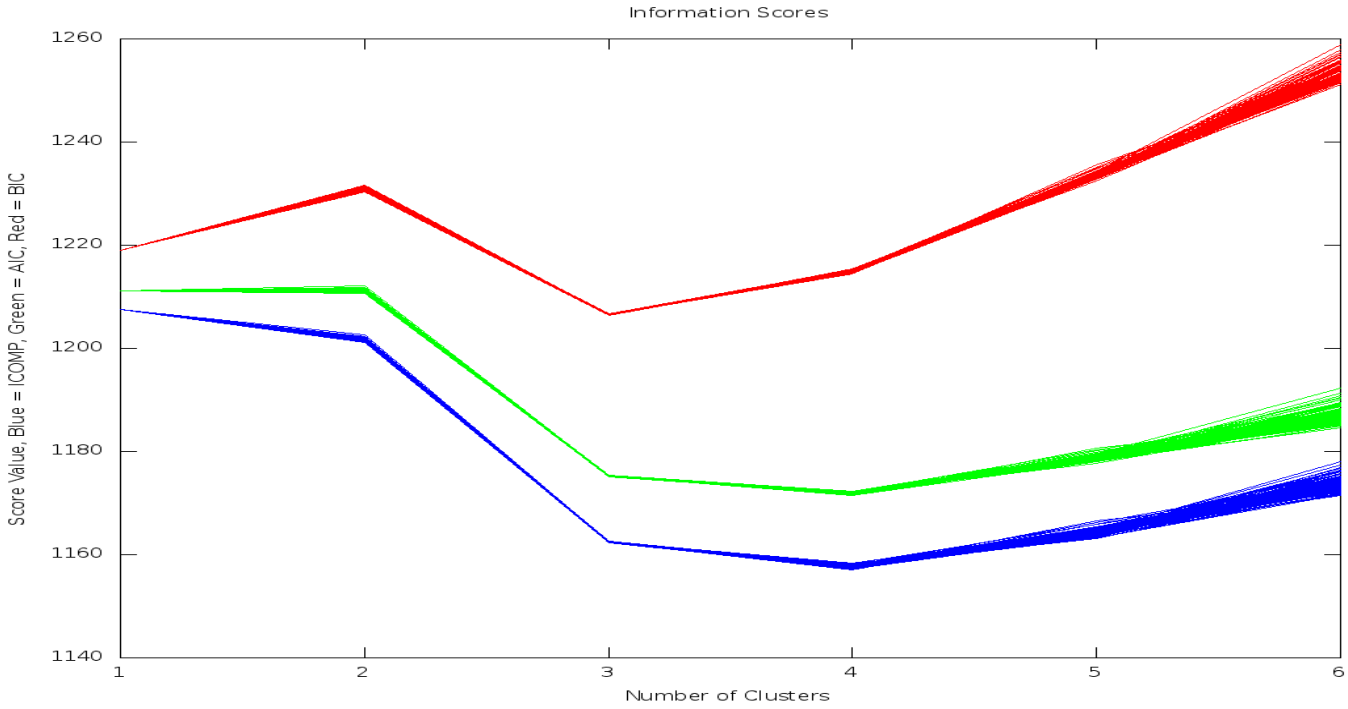


Simulated data with four clusters

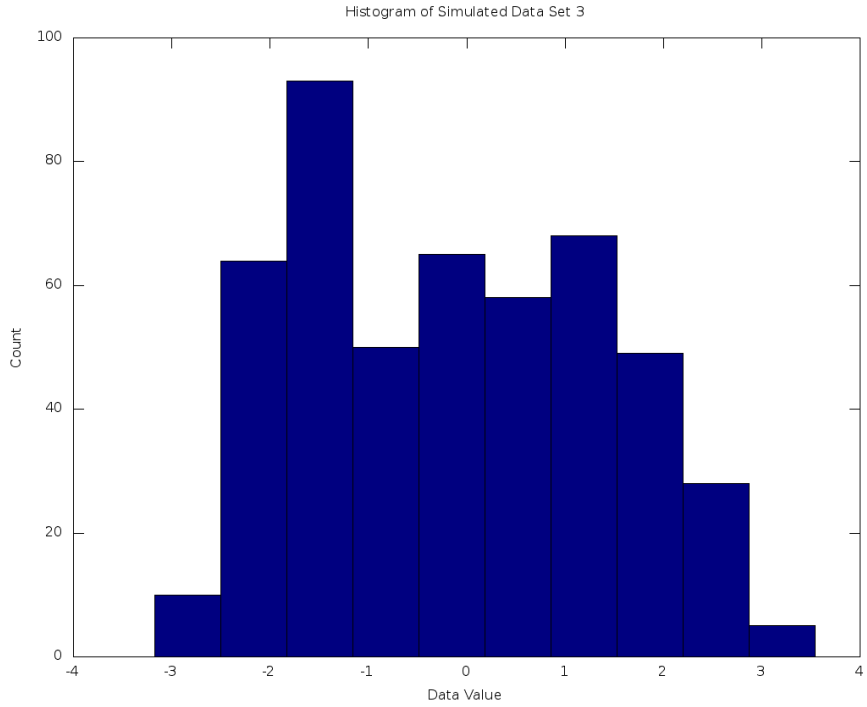




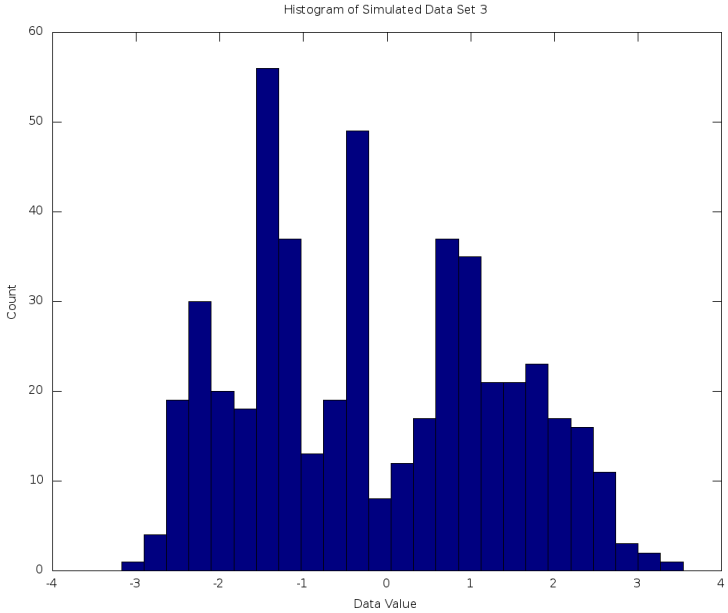
Information Scoring Results of 100 trials of Simulated Data set with four components

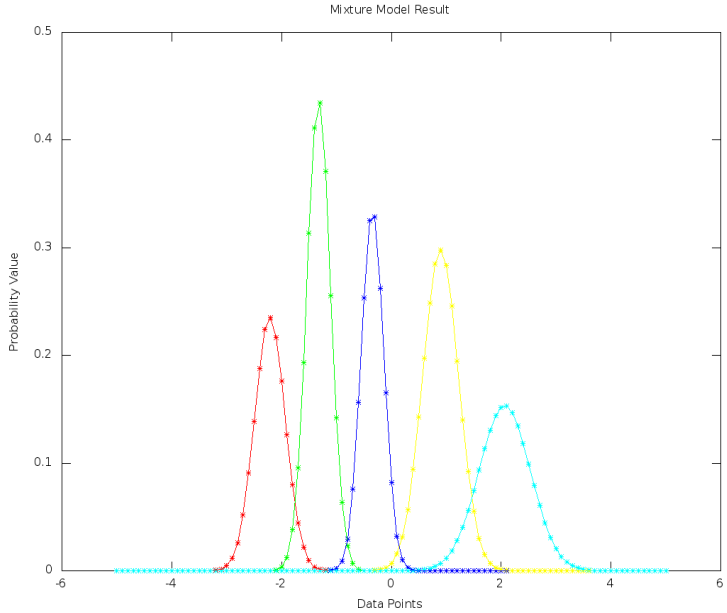


How Many Clusters Do You Think There Are Here?

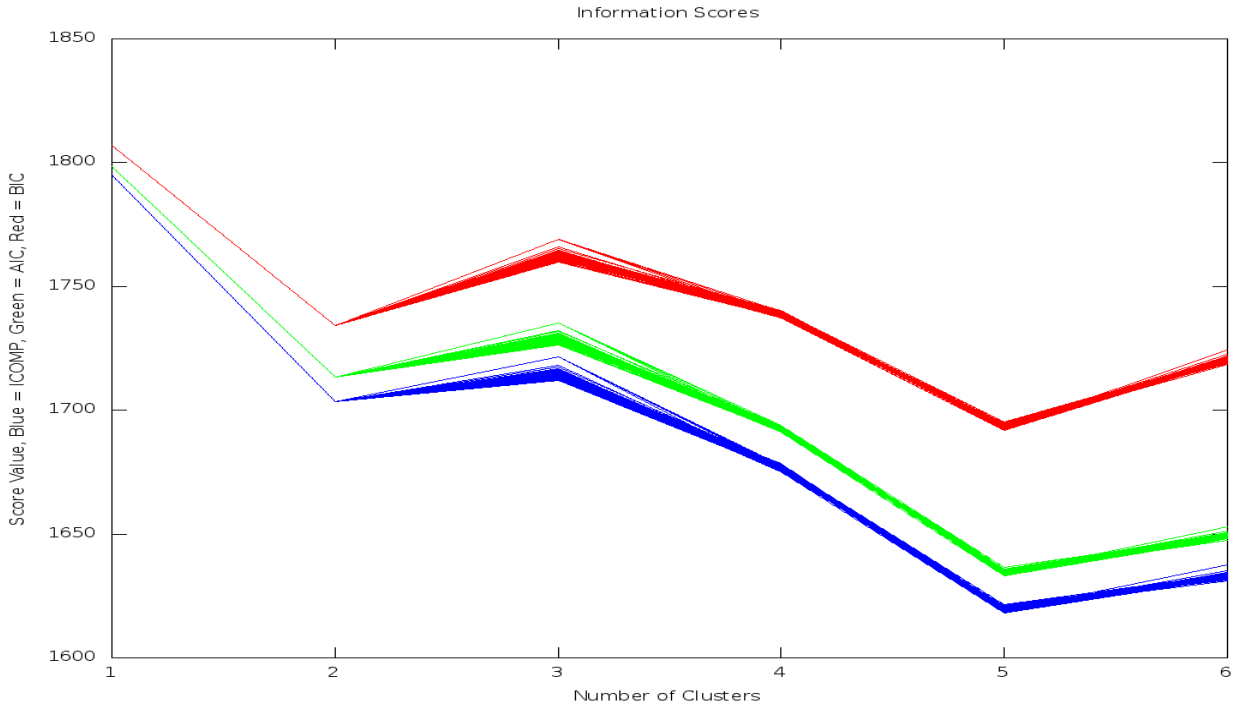


Simulated data with five clusters





Information Scoring Results of 100 trials of Simulated Data set with five components



Applying the new analysis method to Astronomy data

- Astronomers have many kinds of data that need to be classified, including galactic metallicity distributions and kinematic data that generally follow Gaussian distributions.
- Astronomy data represent good test cases for new methods in data analysis for a few reasons.
- The data are representative of many other types of data, including the anomolies and difficulties.
- Astronomy data are free of cost and legal restrictions.

Distribution analysis in Astronomy

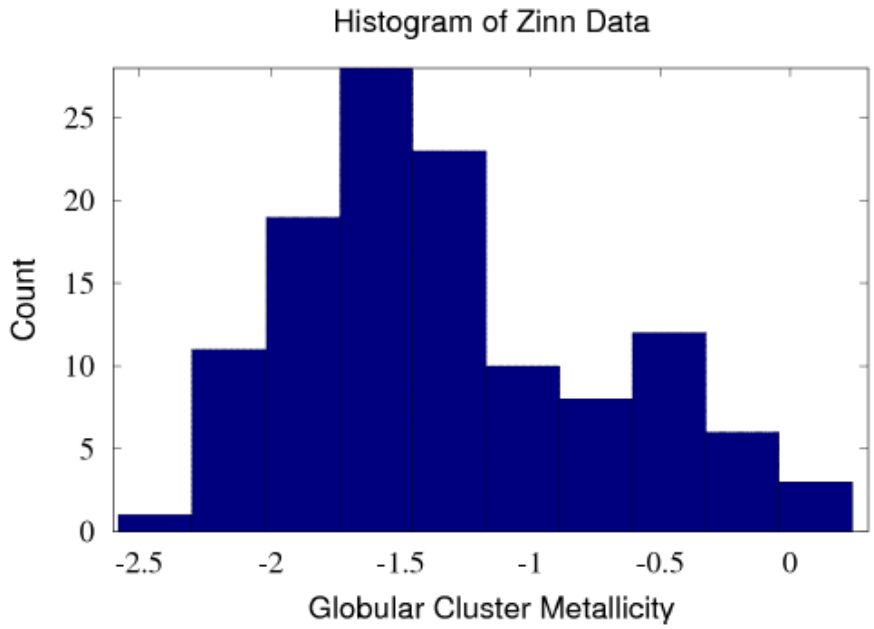
- The KMM algorithm was introduced in the late 1980s and popularized in Astronomy by Ashman, Bird and Zeph (1994).
- KMM = Kaye's Mixture Model
- This algorithm uses series-based iterative sums to calculate Log-Likelihoods.
- Basically the same as the EM algorithm.
- Although this method was introduced in the 1980s, it is still being used in astronomical research.
- I found it referenced in recent astronomy articles.

Why did I analyze older data?

- I wanted to compare the results of the new data processing method with previously published results.
- Previous authors published numerical values from the statistical tests that they applied to their data.
- I can compare the numerical values of their tests to judge how similar or different my results are from the previous authors' results.
- This can also test the reliability of the new analysis method by comparing the results with physical properties of the systems.

Examples from Astronomy

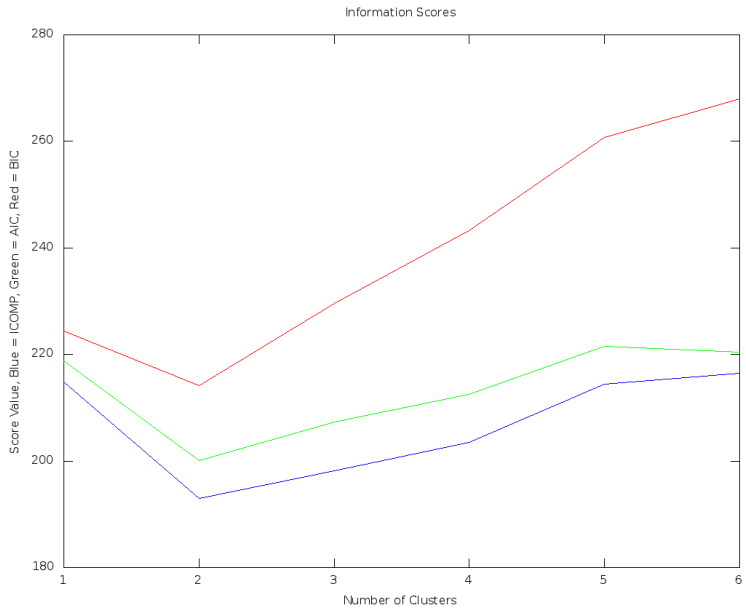
- Globular clusters are star groupings that orbit the nucleus of our galaxy, the Milky Way.
- Zinn (1985) studied the chemical composition of 121 globular clusters.
- Zinn thought there are two populations of globular clusters based on the histogram.



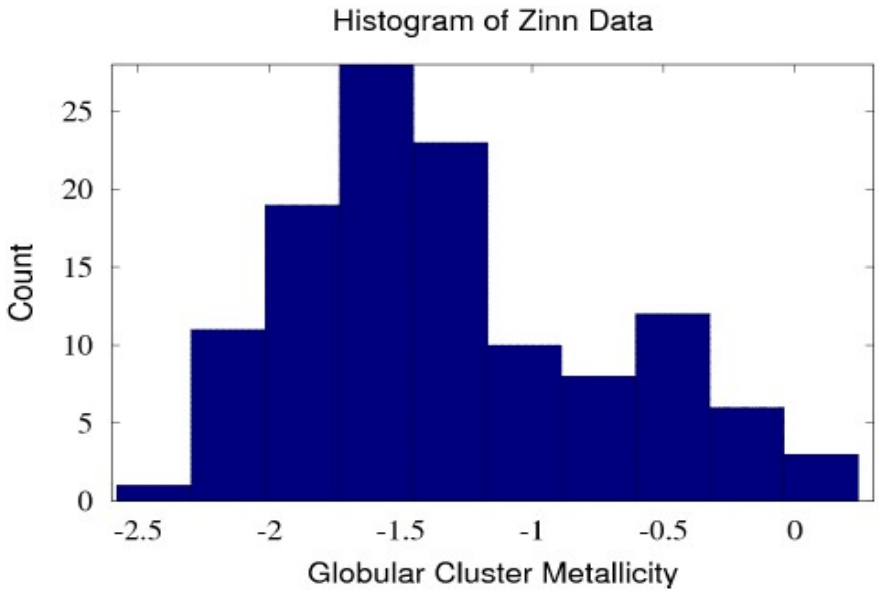
Information scores agree with Zinn's result

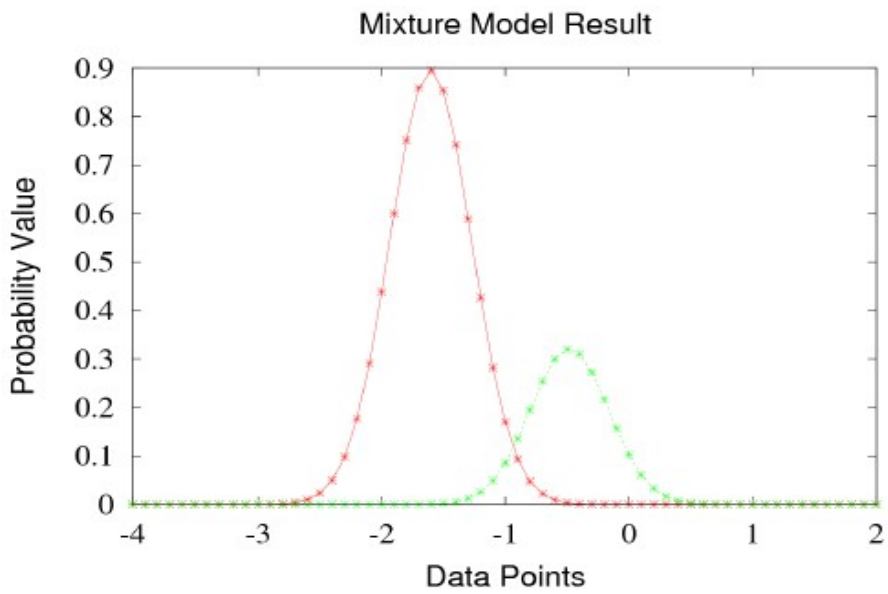
Information scoring found that this data set is best modeled by a two component mixture model.

Zinn did not use a statistical test, but these results confirm his conclusion.



Two component mixture model of Zinn's data



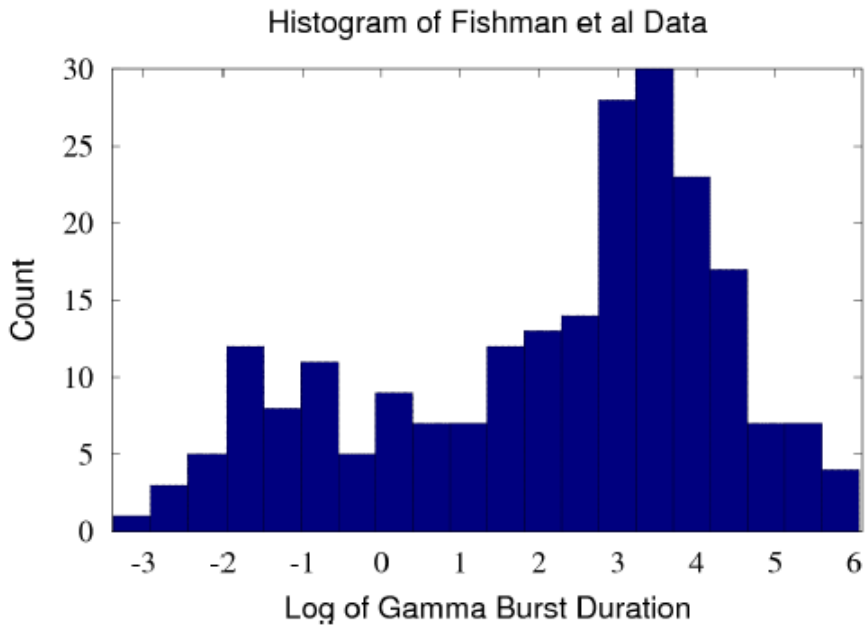


I rechecked an example from Ashman, Bird and Zeph (ABZ)

- The ABZ article applied KMM to several data sets.
- The article gives numerical values for the statistical tests.
- We can check how well information scoring agrees with the KMM result.
- Comment on why I agree or disagree with the previous author.

Gamma Ray Bursts

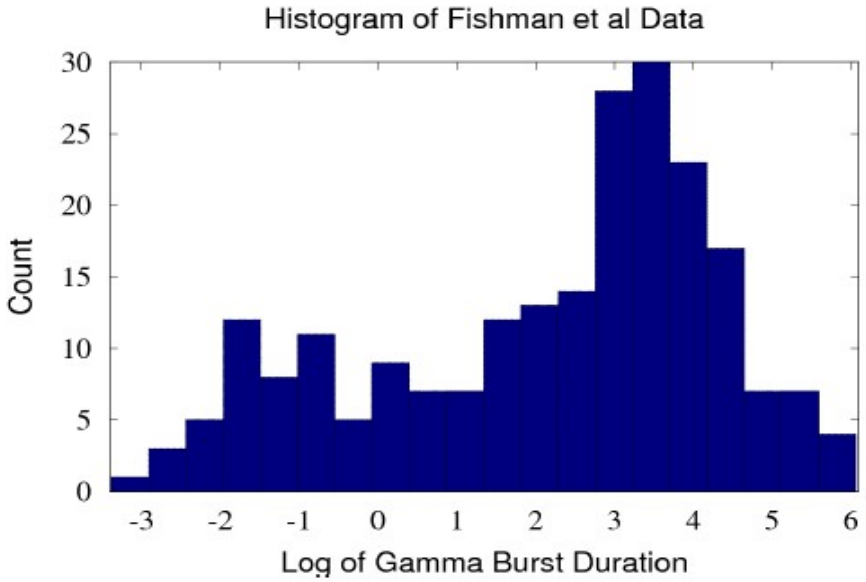
- Huge explosions that occur very far from our solar system.
- ABZ took the data from Fishman et al. (1994)
- There are 222 data points representing the Log of duration of Gamma Ray bursts.

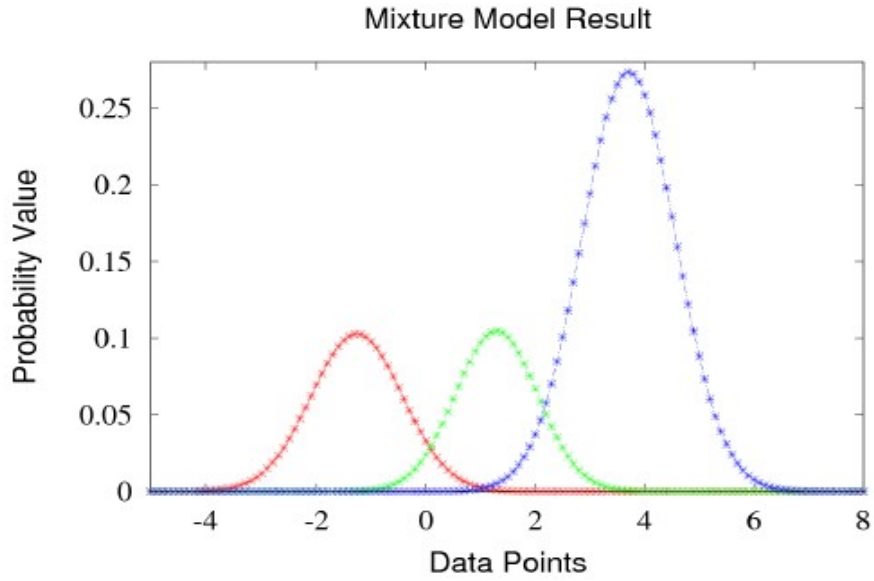


Results of Gamma Ray Burst data

- The traditional hypothesis-based test rejected the single component hypothesis with $P < 0.001$.
- Does not tell how many clusters are present in the dataset other than “more than one”
- My analysis found that this can be best modeled by three clusters.
- These results can help astronomers better understand the physical processes occuring in Gamma-Ray Bursts.

Three component mixture model of Fishman et al.'s data





Future Plans

The example data sets are univariate cases. In the future, I want to also develop multivariate cluster analysis with information scoring that can process more complex data sets where the variables could also show correlations.

In addition, if the underlying probability distributions that generate the data are not Gaussian, there are still related statistical methods we can use to analyze them.

Summary

Modern methods of cluster analysis based on information scoring overcome handicaps inherent in traditional methods that rely on hypothesis tests.

The methods I am developing can more accurately model the underlying probability distribution function that generated a dataset.

Still in testing the algorithm, but I'm optimistic about future prospects for the line of research.

Questions? Comments?

If you're interested in this work, I am happy to talk more about it with you.

Email: jewicker@gmail.com

Linked in: James Wicker in Beijing

If you are interested in astronomy, look at the journal where I work: http://www.raa-journal.org