Applications of Modern Statistical Methods to Analysis of Data in Physical Science

Conference Presentation

James Wicker

Ph.D. Research

Introduction

- Modern Methods in Statistics offer accurate and efficient ways of processing data
- Many researchers do not realize how processing data can affect interpretation
- Different ways of processing data can return different results
- Modern methods are firmly grounded
- Gives more confidence in interpretation

Outline of Topics

- Overview of Modern Regression Methods
- Applications of Modern Linear Regression to Spectral Analysis
- Overview of Modern Clustering Methods
- Applications of Modern Clustering Methods to Astronomy

Linear Regression

- Many phenomena show a relationship between quantities
- Simple linear regression

$$y = \beta_0 + \beta_1 x + \epsilon$$

- Multiple linear regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_k x_k + \epsilon$$

- Statistical Modeling finds best equation

Statistical Modeling

- Classical methods focus on reducing the Sum of Squared Error
- Works well for Simple Regression
- This strategy can cause problems in Multiple Regression
- Ambiguity in model selection, overfitting, ad hoc selection methods

Scored Regression

- Instead of judging competing models based on Reduction of SSE or F-test
- Assigns a score to different combinations of model variables
- Overcomes subjective thresholds of classical methods
- This method is derived from statistical theory, gives confidence in results

AIC and ICOMP

- Two examples of scoring functions used in regression
- AIC was the original scoring function

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

- ICOMP is a more modern scoring function that better models interactions of variables

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

Using scored regression

- Compute regression parameters for different combinations of models
- Use the scoring functions (AIC or ICOMP) to compute a score for these combinations
- The model combination that achieves the lowest score is the best

Regression in Spectroscopy

- Since the 1960’s, researchers have been using regression algorithms for molecular parameters
- Expand Hamiltonian in power series
- v is vibration quan. num., J is rotation quan. num. , K is z component of rotation quan. num.
- Eigenvalues of angular momentum operators
- Regressor x terms are functions of v, J and K and changes in these terms
- Response y values are the transition frequencies

Historical Stepwise Analysis

- Under the stepwise scheme, the researcher initializes the regression algorithm with a model containing terms that they believe are important
- The algorithm successively adds and deletes variables according to F-test thresholds
- The algorithm also deletes outliers until no change occurs

Information Scored Analysis

- Structure of Information scored method is analogous to the structure of stepwise process
- Algorithm starts by forcing lower-order terms and assigns scores to different combinations of higher-order terms
- The combination of variables that achieves the minimum information score is best
- Treats outliers in analogous way

Advantages of this Method

- Overcomes arbitrary F-test values for variable selection
- Overcomes ad hoc assumptions of stepwise process
- Closely connected with theory of regression
- Achieves optimization in variable selection while obeying power series requirements

Some examples of analysis

- I elected to process some historical data sets because these authors clearly stated how they analyzed the data and what model they used
- Started with low variable data
- Later analyzed data with more variables
- Theory says low variable data should agree with stepwise analysis
- More disagreement with more variables

Boyd/Kurlat: CD_3I $2\nu_4$ Low Variable

Variable	Parameter	Quantum Dependency
1	$\nu_4 + x_{44} + x_{l_4 l_4}$	$\Delta\nu_4$
2	$A_0 + 2A_e\zeta_4$	$(2K + \Delta K)\,\Delta K$
3	D_0^K	$K^4 - (K + \Delta K)^4$
4	α_4^A	$-\Delta\nu_4\,(K + \Delta K)^2$
5	α_4^B	$-\Delta\nu_4[(J + \Delta J)\,(J + \Delta J + 1) - (K + \Delta K)^2]$
6	$x_{l_4 l_4} + (1/4)A_e\zeta_4$	Δl_4^2

Variable	Kurlat et al Value	95% CI	Wicker Value	95% CI
1	2273.069	0.001	2273.070	0.001
2	3.4723	0.0002	3.4721	0.0003
3	3.81×10^{-5}	5×10^{-7}	3.77×10^{-5}	9×10^{-7}
4	0.01283	1×10^{-5}	0.01284	2×10^{-5}
5	8.7×10^{-5}	2×10^{-6}	8.7×10^{-5}	1×10^{-6}
6	8.8136	8×10^{-4}	8.8140	1×10^{-3}

Kurlat: 3rd CD_3I $\mathbf{2}\nu_4$, $\nu_4+\nu_5$, and $\nu_2+\nu_4$

Variable	Parameter	Quantum Dependency
1	$\nu_2 + ...$	$\Delta\nu_2$
2	$\nu_4 + ...$	$\Delta\nu_4$
3	$\nu_5 + ...$	$\Delta\nu_5$
4	A_0	$(2K+\Delta K)\,\Delta K$
5	$A_e\zeta_4^z$	$-2\Delta l_4\,(K+\Delta K)$
6	$A_e\zeta_5^z$	$-2\Delta l_5\,(K+\Delta K)$
7	D_0^K	$K^4-(K+\Delta K)^4$
8	α_2^A	$-\Delta\nu_2\,(K+\Delta K)^2$
9	α_4^A	$-\Delta\nu_4\,(K+\Delta K)^2$
10	α_5^A	$-\Delta\nu_5\,(K+\Delta K)^2$
11	α_2^B	$-\Delta\nu_2\left[(J+\Delta J)\,(J+\Delta J+1)-(K+\Delta K)^2\right]$
12	α_4^B	$-\Delta\nu_4\left[(J+\Delta J)\,(J+\Delta J+1)-(K+\Delta K)^2\right]$
13	α_5^B	$-\Delta\nu_5\left[(J+\Delta J)\,(J+\Delta J+1)-(K+\Delta K)^2\right]$
14	$x_{l_4l_4} + ...$	$\Delta l_4^2(J+\Delta J+1)-(K+\Delta K)^2$
15	η_5^J	$-2\Delta l_5\,(K+\Delta K)\,(J+\Delta J)\,(J+\Delta J+1)$
16	η_4^K	$\Delta l_4\,(K+\Delta K)^3$
17	η_5^K	$\Delta l_5\,(K+\Delta K)^3$
18	η_4^J	$\Delta l_4\,(K+\Delta K)\,(J+\Delta J)\,(J+\Delta J+1)$

Kurlat: 3rd CD_3I $\mathbf{2}\nu_4$, $\nu_4 + \nu_5$, **and** $\nu_2 + \nu_4$

Number	Kurlat et al Value	95% CI	Wicker Value	95% CI
1	961.799	0.009	961.793	0.006
2	2273.071	0.003	2273.0696	0.003
3	1056.202	0.007	1056.206	0.005
4	2.5797	0.0003	2.5791	0.0004
5	0.4461	0.0005	0.4464	0.0002
6	-0.816	0.0005	-0.818	0.0005
7	Not Sig		0.000037	1×10^{-6}
8	-0.0042	0.0003	-0.0040	0.0002
9	0.01287	0.00003	0.01285	0.00004
10	0.0153	0.0001	0.0154	0.0001
11	0.00178	0.00003	0.00175	0.00002
12	0.000086	0.000004	0.000086	0.000002
13	-0.00044	0.00002	-0.00042	0.00001
14	8.592	0.002	8.592	0.001
15	-4×10^{-6}	3×10^{-6}	-1×10^{-5}	2×10^{-6}
16	4×10^{-6}	2×10^{-6}	Not Sig	
17	-0.00001	0.00001	Not Sig	
18	Not Sig		Not Sig	

Modern High Resolution Data

- For higher resolution data, the regression analysis would be the first step in an iterative process that includes perturbation analysis
- We compared an initial fit of our data with the final results from Guelachvili et al. (1984)
- Original Authors sd= 9×10^{-4}
- My unweighted sd= 8×10^{-4}
- My weighted sd= 6×10^{-4}
- No hand selection, automatically select points

Guelachvili CD_3I ν_4 High Res.

Unweighted Regression Estimates of Molecular Parameters

Constant	Guelachvili et al Value	Wicker Value
ν_0	2298.54431	2298.54369
A_4	2.5825779	2.5960037
B_4	0.20139595	0.201406187
$A\zeta_4$	0.463800	0.463952

Weighted Regression Estimates of Molecular Parameters

Constant	Guelachvili et al Value	Wicker Value
ν_0	2298.54431	2298.54397
A_4	2.5825779	2.5826756
B_4	0.20139595	0.201404898
$A\zeta_4$	0.463800	0.464011

Summary of Scored Regression

- In low variable limit, modern method agrees with classical method
- As more variables are added, more disagreement appears
- My method is on firm theoretical grounds
- I think that scored regression has more general application in physics: calibration studies, intensity studies, etc.

Cluster Analysis

- Tries to find structure in data
- Traditional methods use the K-means algorithm and the Expectation-Maximization (EM) algorithm
- Both of these traditional methods use initial seed values and iterative estimation
- Strong dependence on initial seed values and little optimization properties

Genetic Algorithms in Clustering

- I implemented Genetic Algorithm (GA) based methods for cluster analysis
- Does not rely on seed values
- Has proven optimization properties based on Markov Chain behavior
- My methods can more accurately identify complex data structures than K-means
- Need accurate parameter estimates in order to best use information scoring

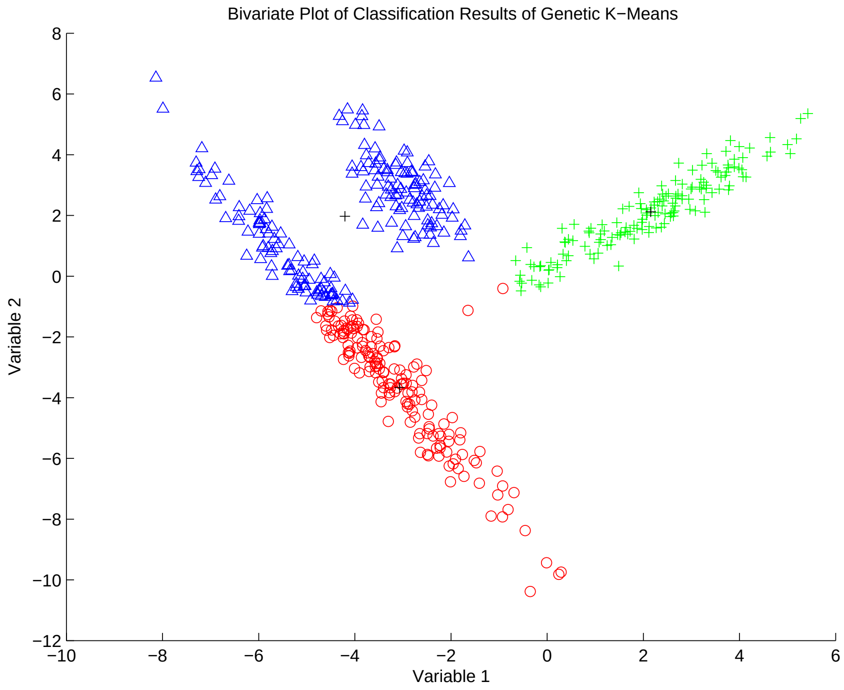
GARM

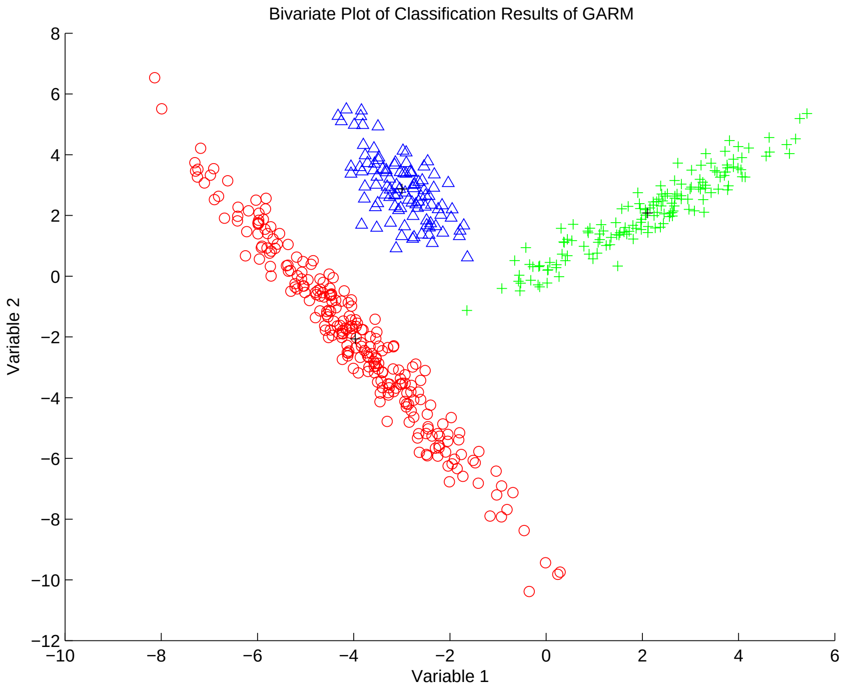
- Genetic Algorithm with Regularized Mahalanobis
- New Cluster partition method for hyperellipsoidal clustering
- Uses String-of-Group numbers representation of GA population
- New GA operations drastically reduce convergence over traditional GA methods

How does GARM work?

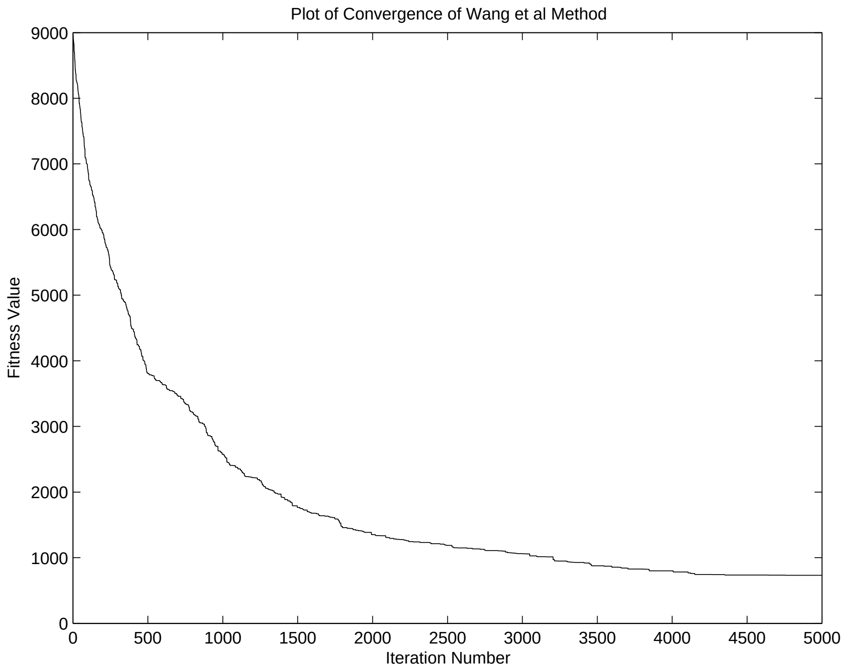
- Initialize population of cluster assignments
- Biased mutation operation – assigns data points to clusters with probability proportional to Regularized Mahalanobis Distance (RMD)
- Genetic Mahalanobis operation – assigns points with closest RMD
- Fitness function is sum of RMD
- Reproduction proportional to fitness

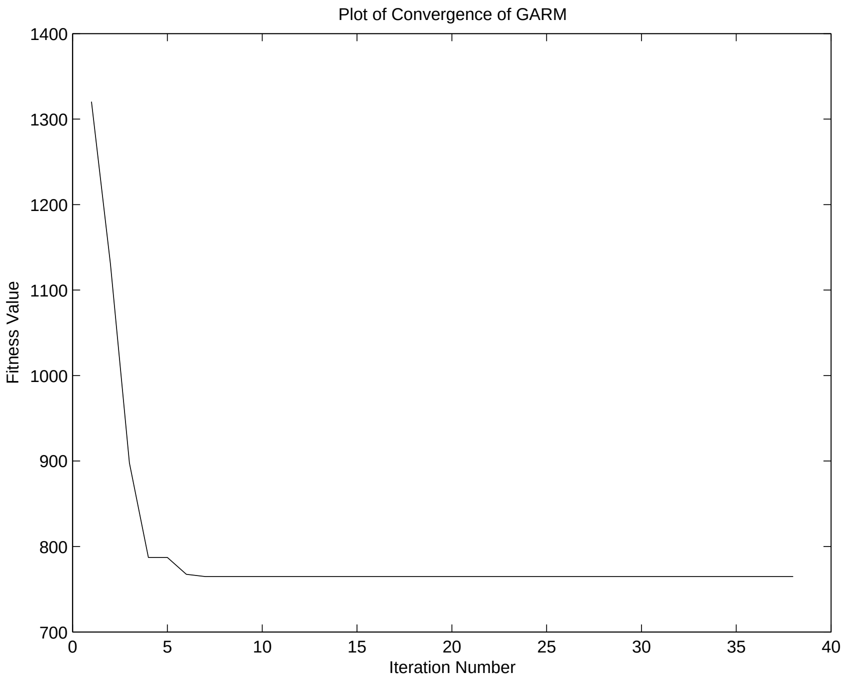
Example: 80.2% vs. 100%



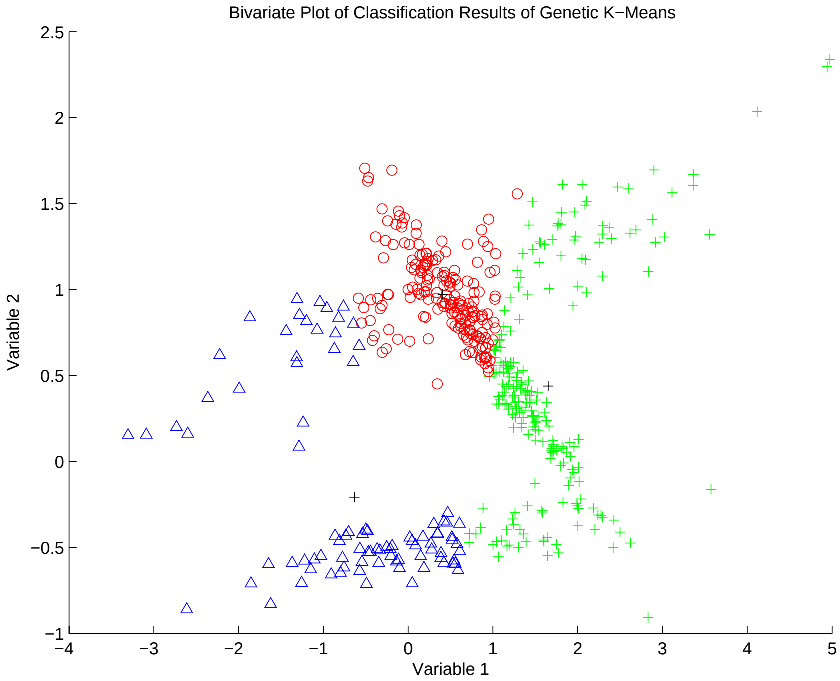


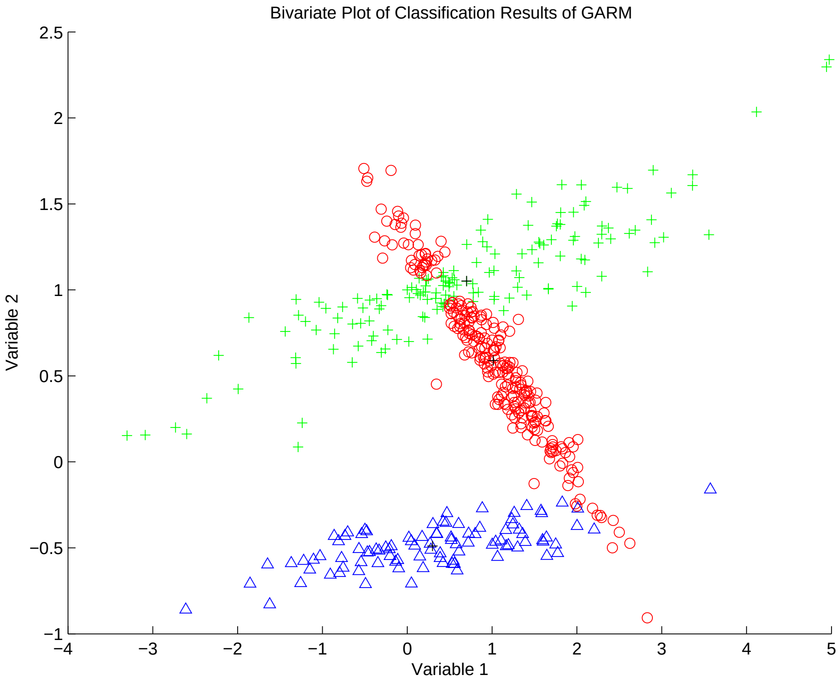
Convergence: 1000's vs. 10's



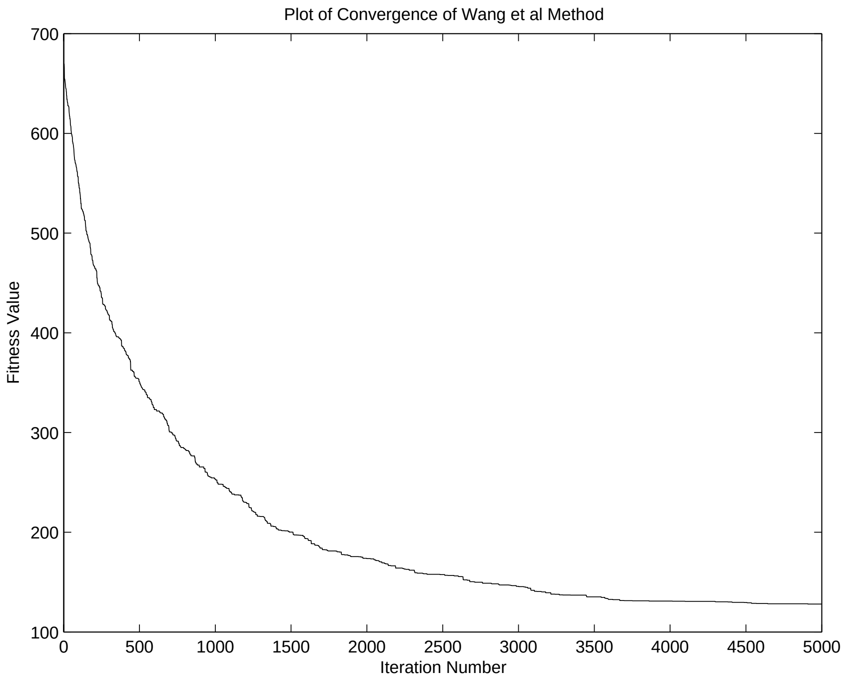


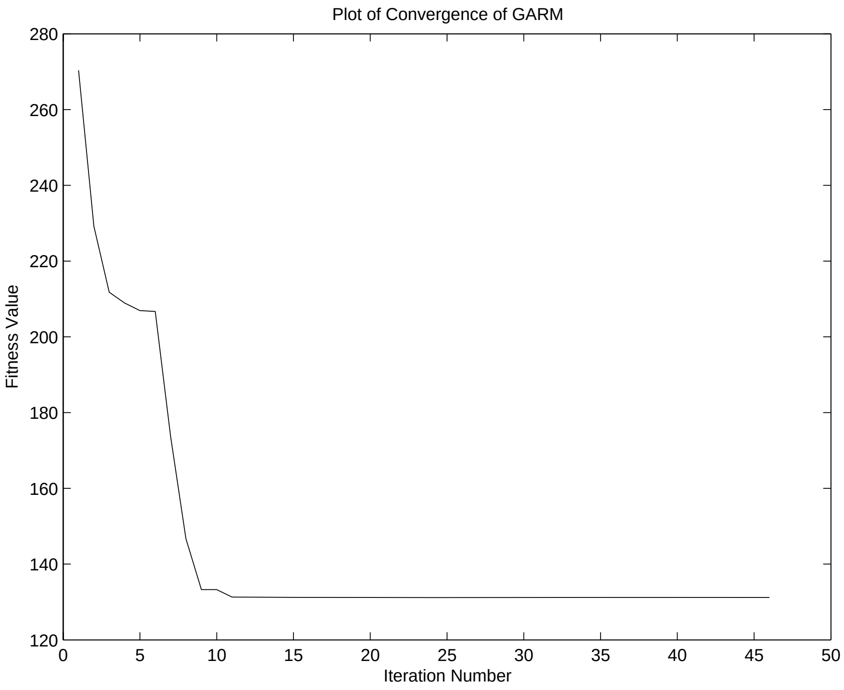
Example: 48.8% vs. 93.8%





More Fast Convergence





Mixture Modeling

- Mixture Modeling classifies data according to probabilities arising from defined distributions
- Most common is normal mixture models

$$g_k(\mathbf{x}; \mu_k, \Sigma_k) = (2\pi)^{-\frac{p}{2}} |\Sigma_k|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu_k)^T \Sigma_k^{-1} (\mathbf{x} - \mu_k)\right]$$

- Maximize Posterior Probabilities of group membership

$$f(\mathbf{x}; \boldsymbol{\pi}, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k g_k(\mathbf{x}; \mu_k, \Sigma_k)$$

Traditionally use EM algorithm

- Traditional EM algorithm starts cluster assignments with K-means initialization
- Iteratively recomputes log-likelihood of mixture model and cluster assignments

$$l(\pi, \mu, \Sigma) =$$

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

- Continues this process until the change in the log-likelihood value is small

Traditional EM algorithm

- Authors admit this is a local maximizer
- Final results depend strongly on starting values
- Parameter space of cluster values generally highly nonlinear with many local maxima
- In data with complex covariance structure, it may be difficult for traditional EM algorithm to find global maximum

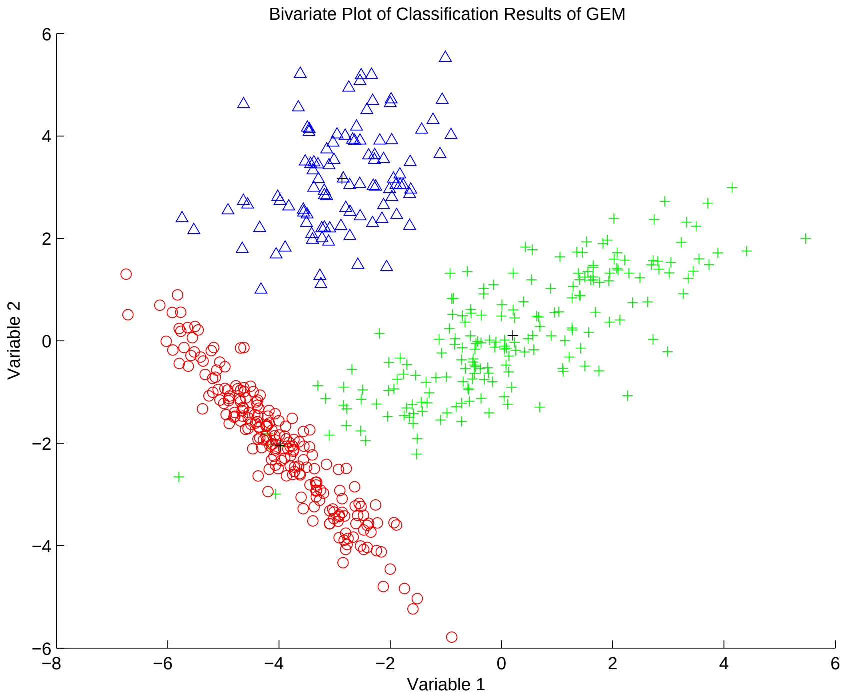
Genetic Expectation Maximization

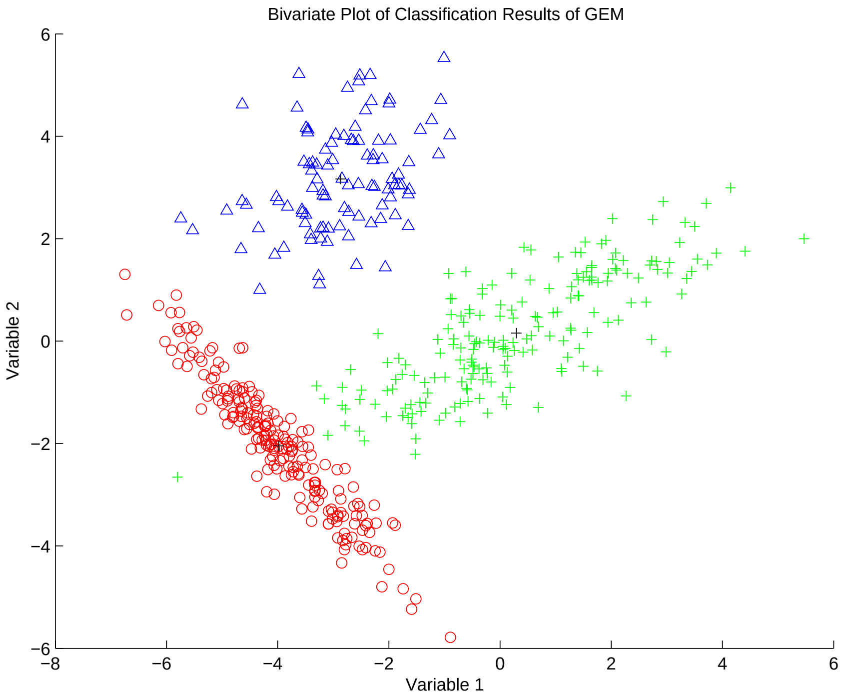
- I introduced the GEM algorithm as a new way to do mixture model cluster analysis
- Analogous string-of-numbers population and GA operations as GARM
- Biased mutation assigns values according to posterior probabilities
- Posterior Probability Operation
- Fitness is the log-likelihood of mixture
- Reproduction proportional to fitness

Advantages of GEM

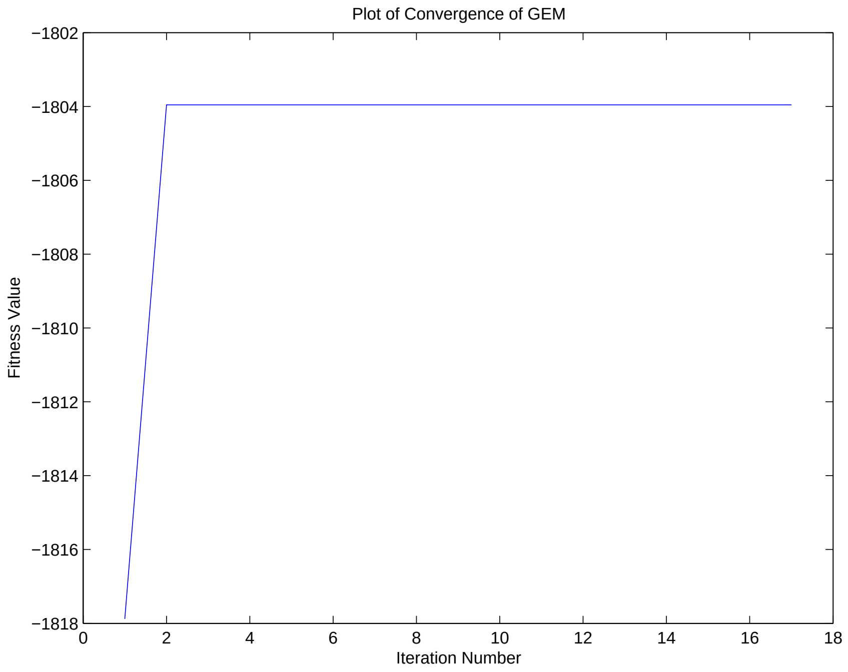
- Does not show strong dependence on initial values
- Relatively fast convergence: traditional algorithm can take 1000's of iterations
- Optimization based on Markov Chain properties
- Better able to accurately model complex covariance structure

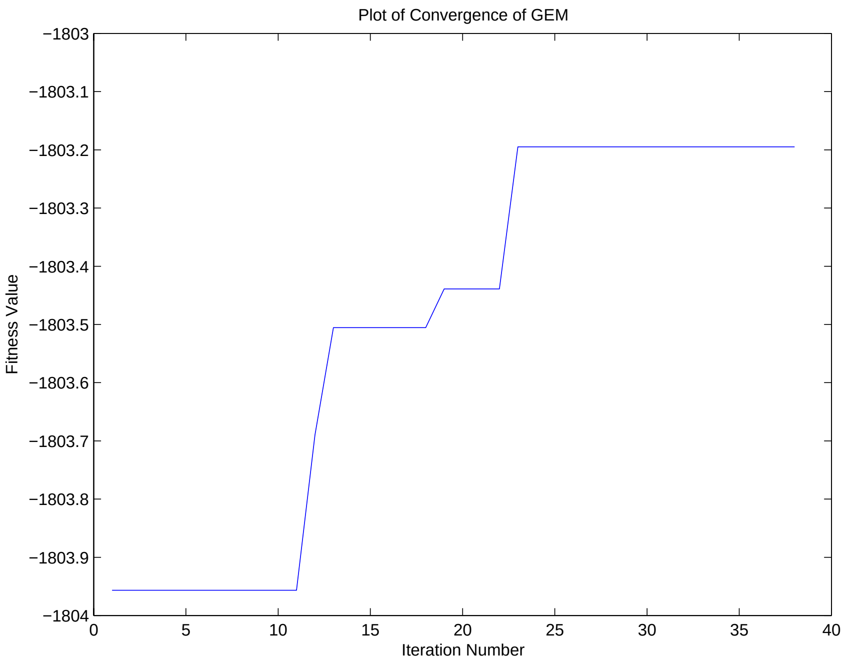
Different initializations:
GARM: 97.8% and GKM: 98.2%



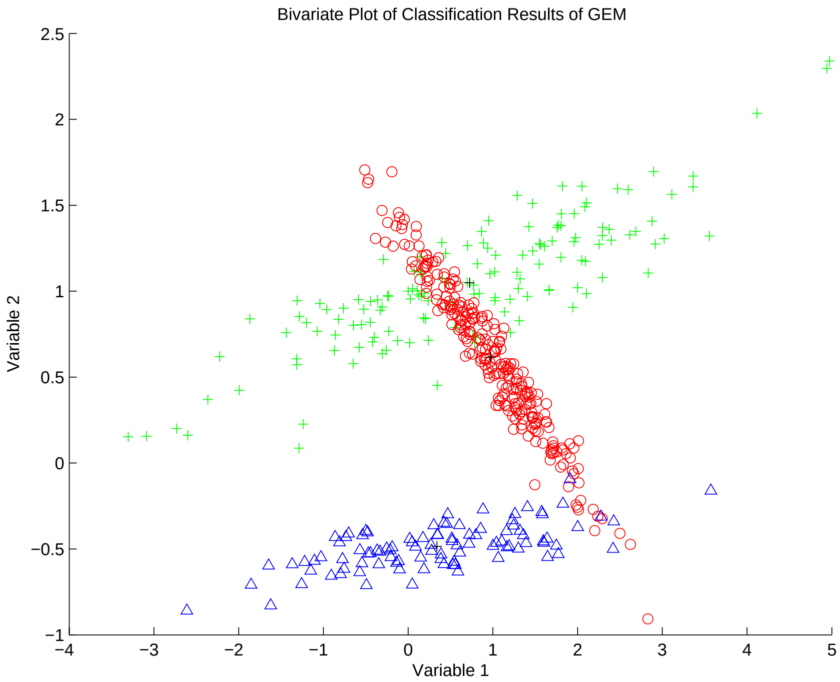


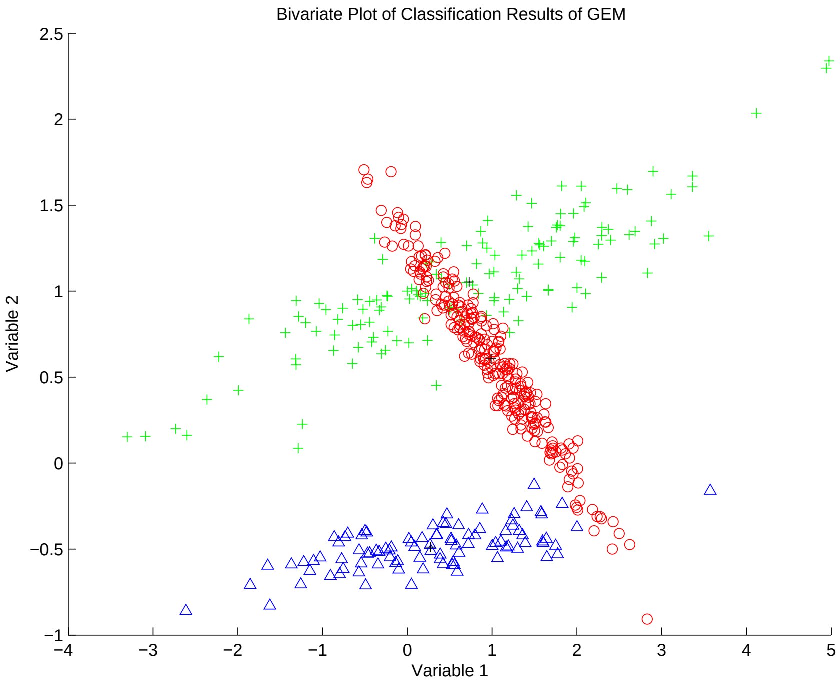
Convergence of Log-Likelihoods with GARM and GKM initializations



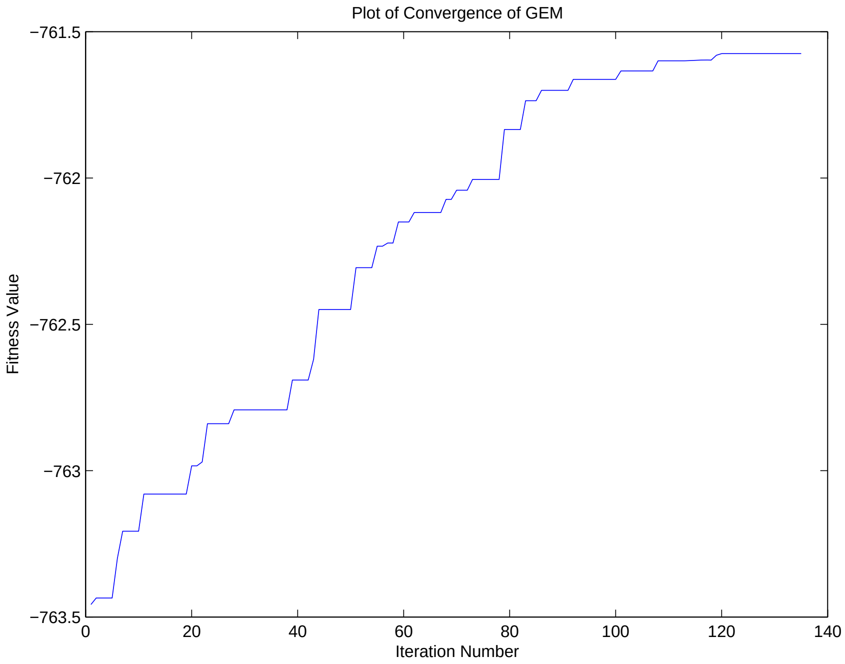


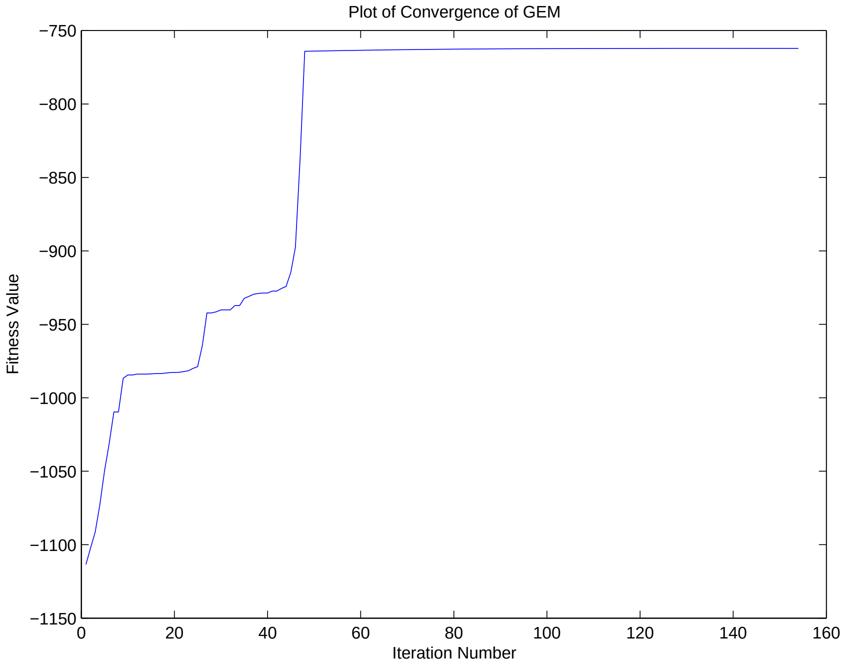
Different initializations: GARM: 92.4% and GKM: 92.4%





Convergence of Log-Likelihoods with GARM and GKM initializations

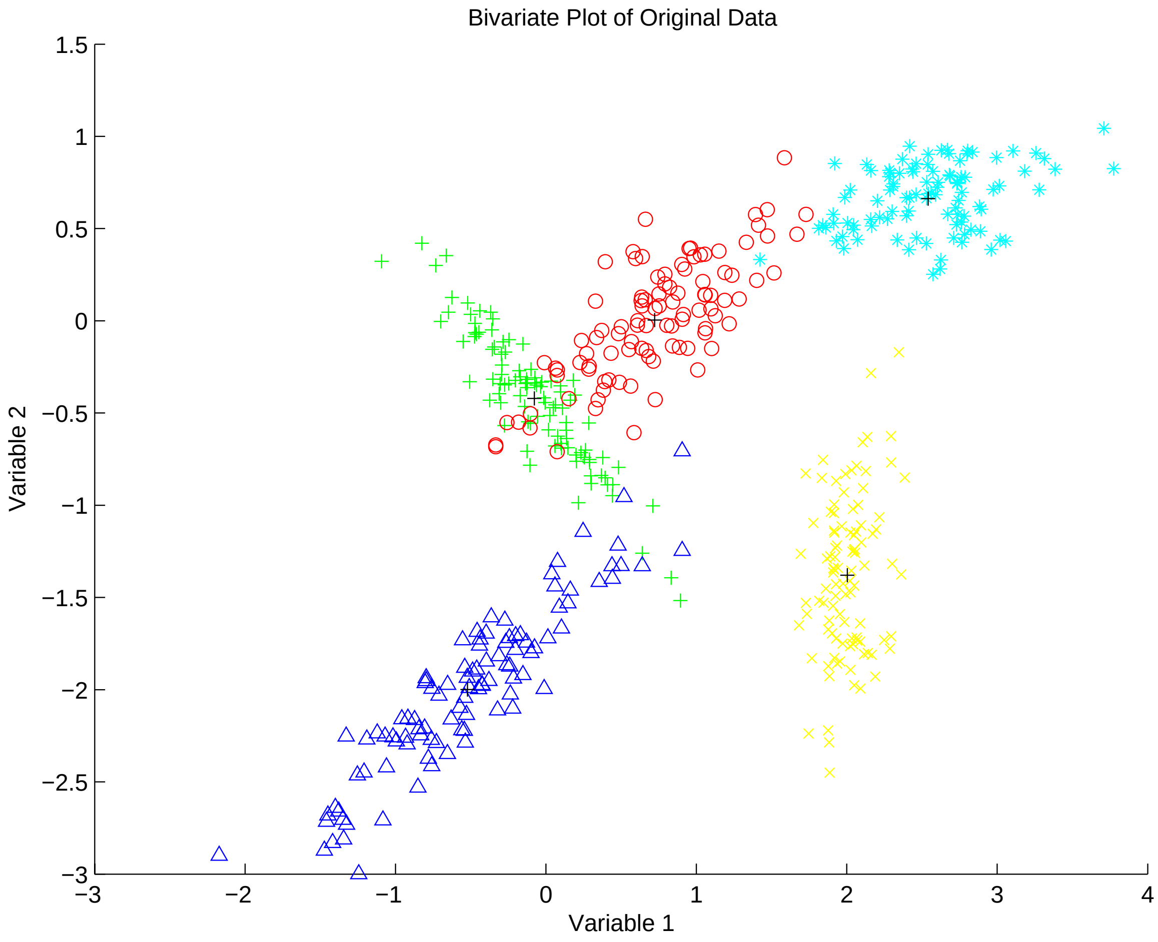




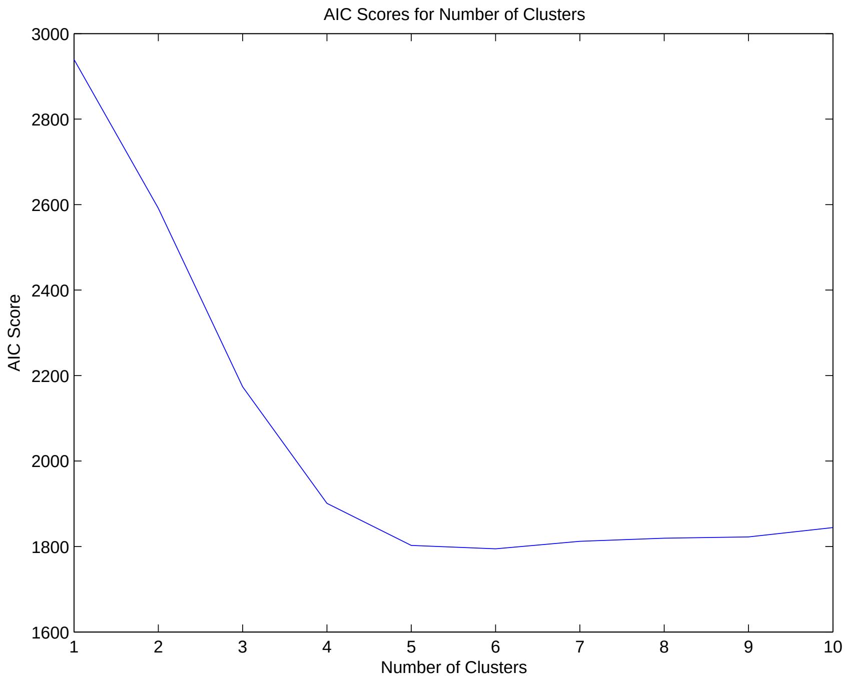
Use GEM in Information Scored Mixture Model analysis

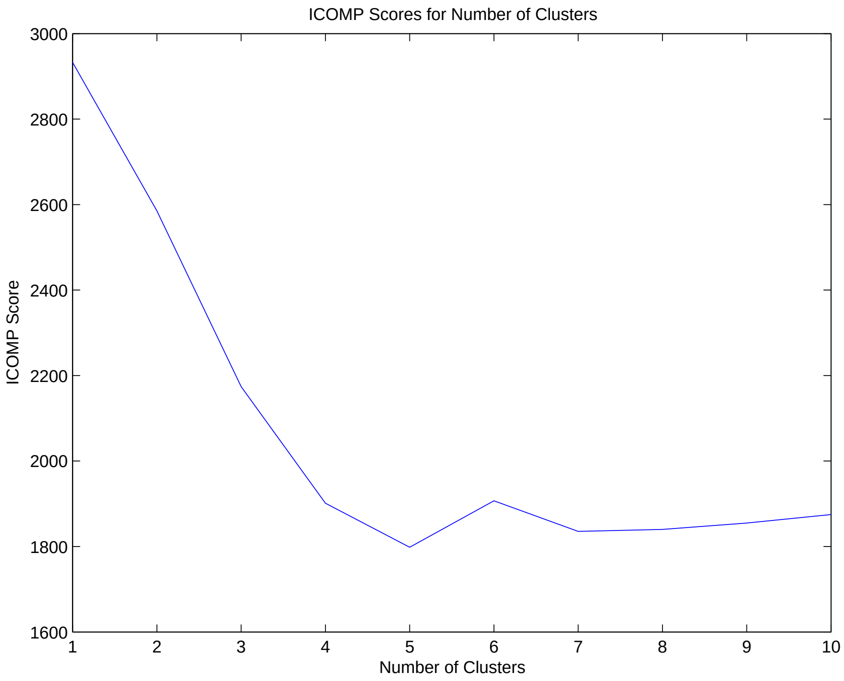
- We can derive information scoring functions AIC and ICOMP in mixture modeling situation
- Depends on number of parameters
- Need accurate estimations of means and covariance parameters to use scoring
- Use GEM to calculate mixture components
- Assign information scores to components
- Components with minimum score is best

Test Data: 5 components

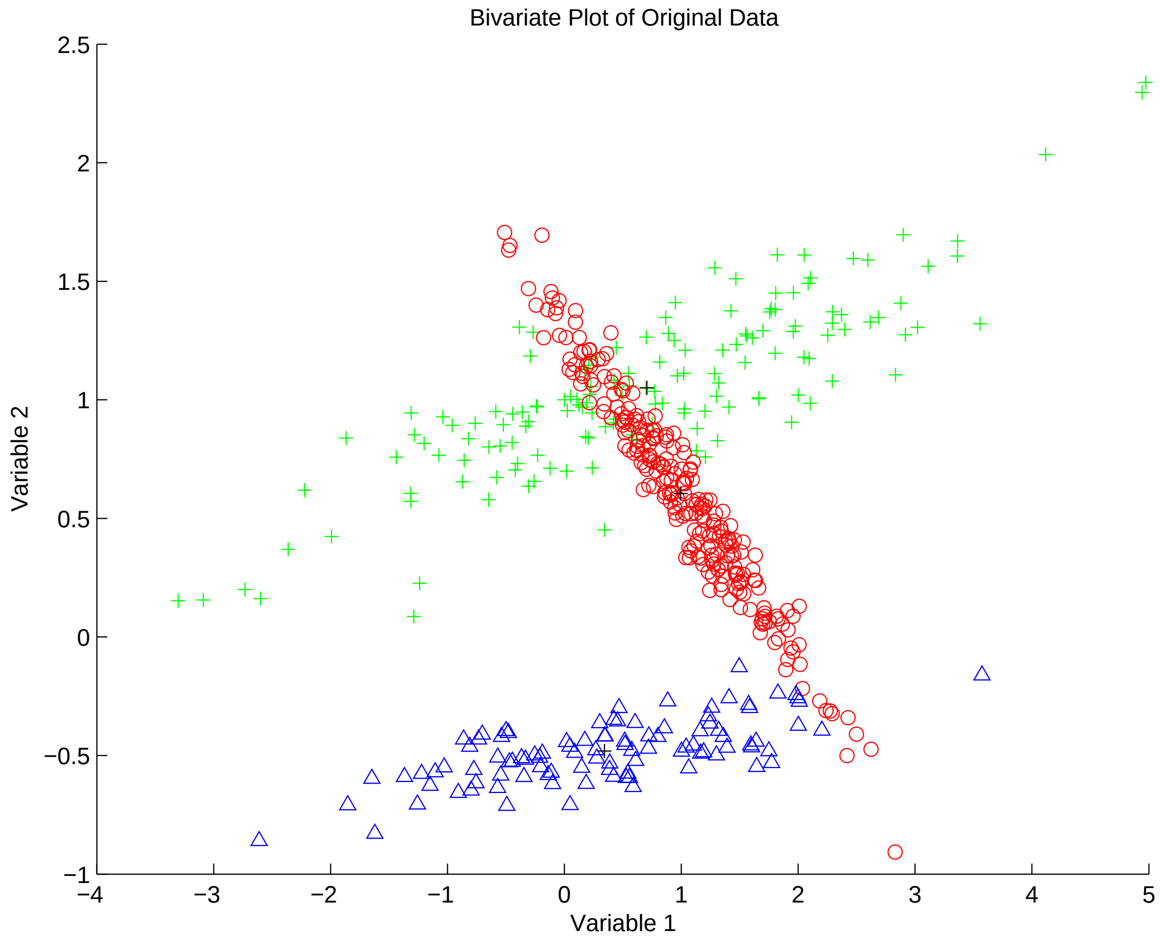


AIC and ICOMP scores

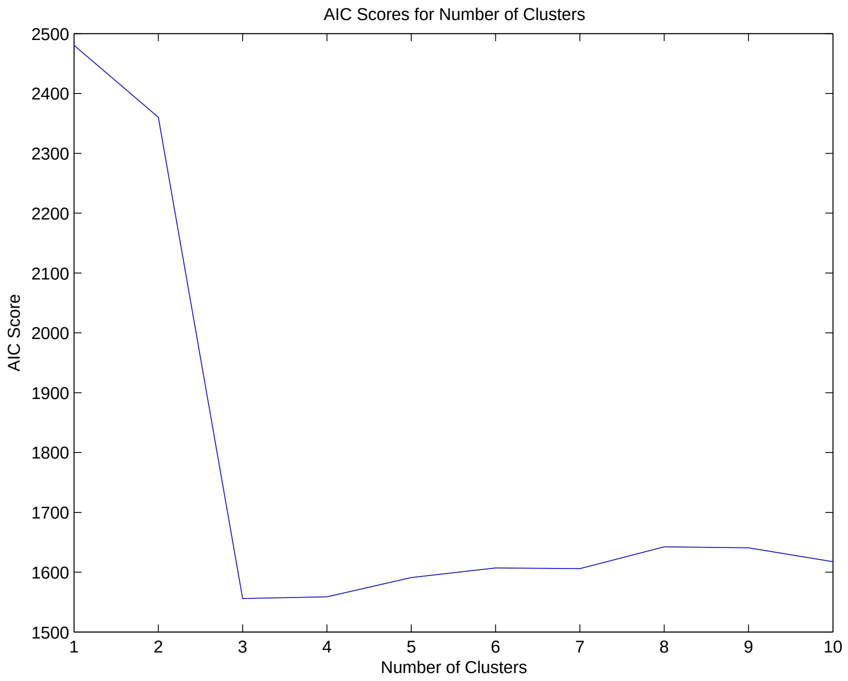


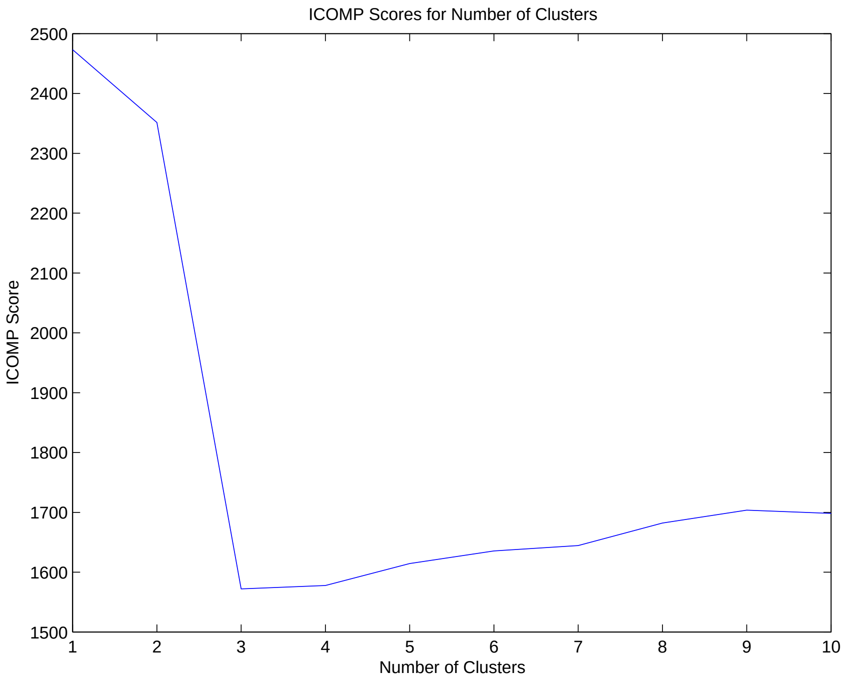


Test Data: 3 components



AIC and ICOMP scores





GA Cluster Analysis Summary

- GARM type operations improve convergence and accuracy of analysis
- GEM finds global maximum in log-likelihood value with little dependence on initial conditions
- GEM returns accurate estimates of cluster means and covariances
- Uses accurate parameter estimates of GEM in information scored cluster analysis

Mixture Models in Astronomy

- I used GEM with information scoring to analyze some astronomical data
- Astronomy data continually grows
- Need automated ways of classifying increasingly multivariate data
- Paper from 2004 states that currently over 100 Tb of data warehoused in astronomy
- Human Genome ~ 1 Gb
- Library of Congress ~ 20 Tb

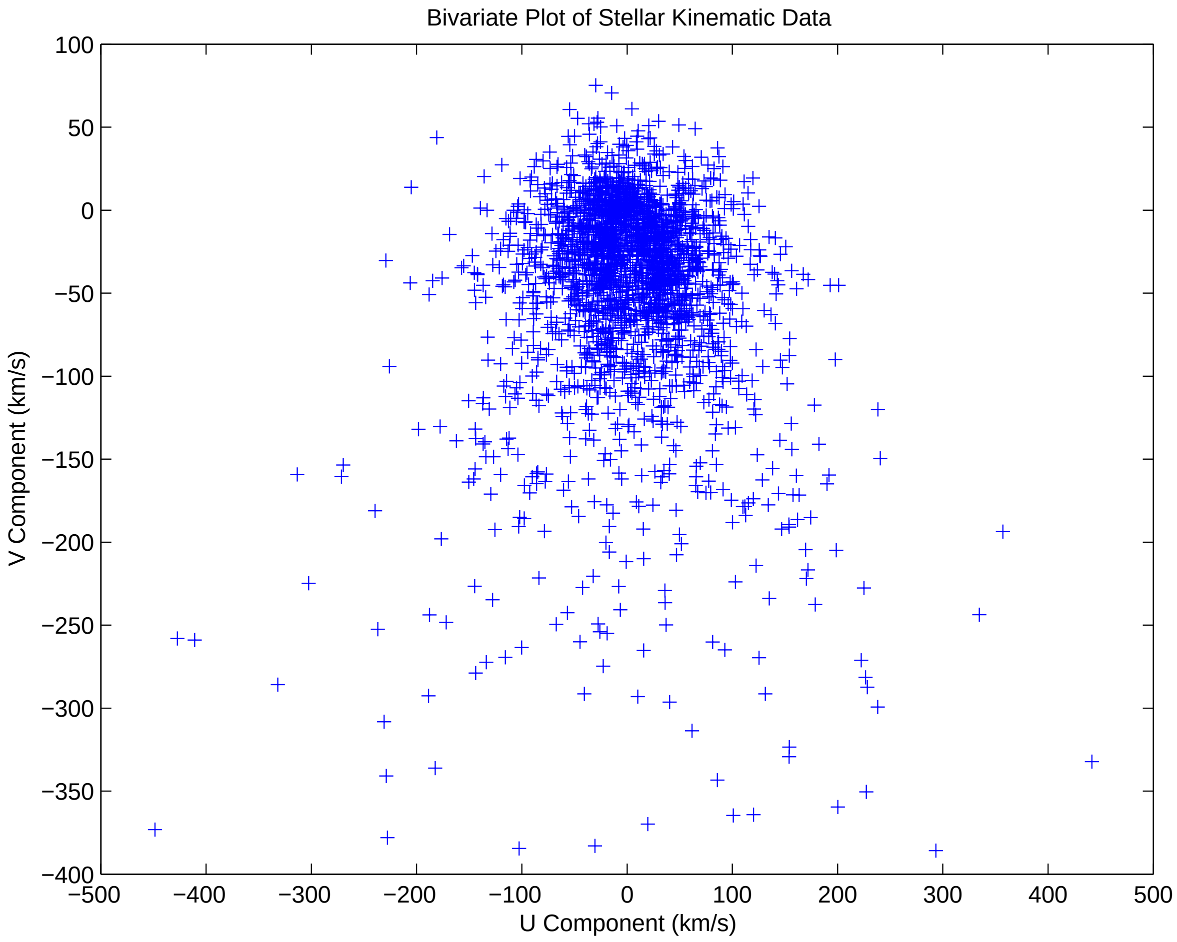
Stellar Kinematic Data

- Soubiran (1993) studied the proper motion of stars in our Galaxy
- Data compiled from photographic plates of 7 square degrees near globular cluster M3
- Plates taken over 40 year time span
- Proper Motion: V component towards galactic pole
- U component in rotational component of galactic motion

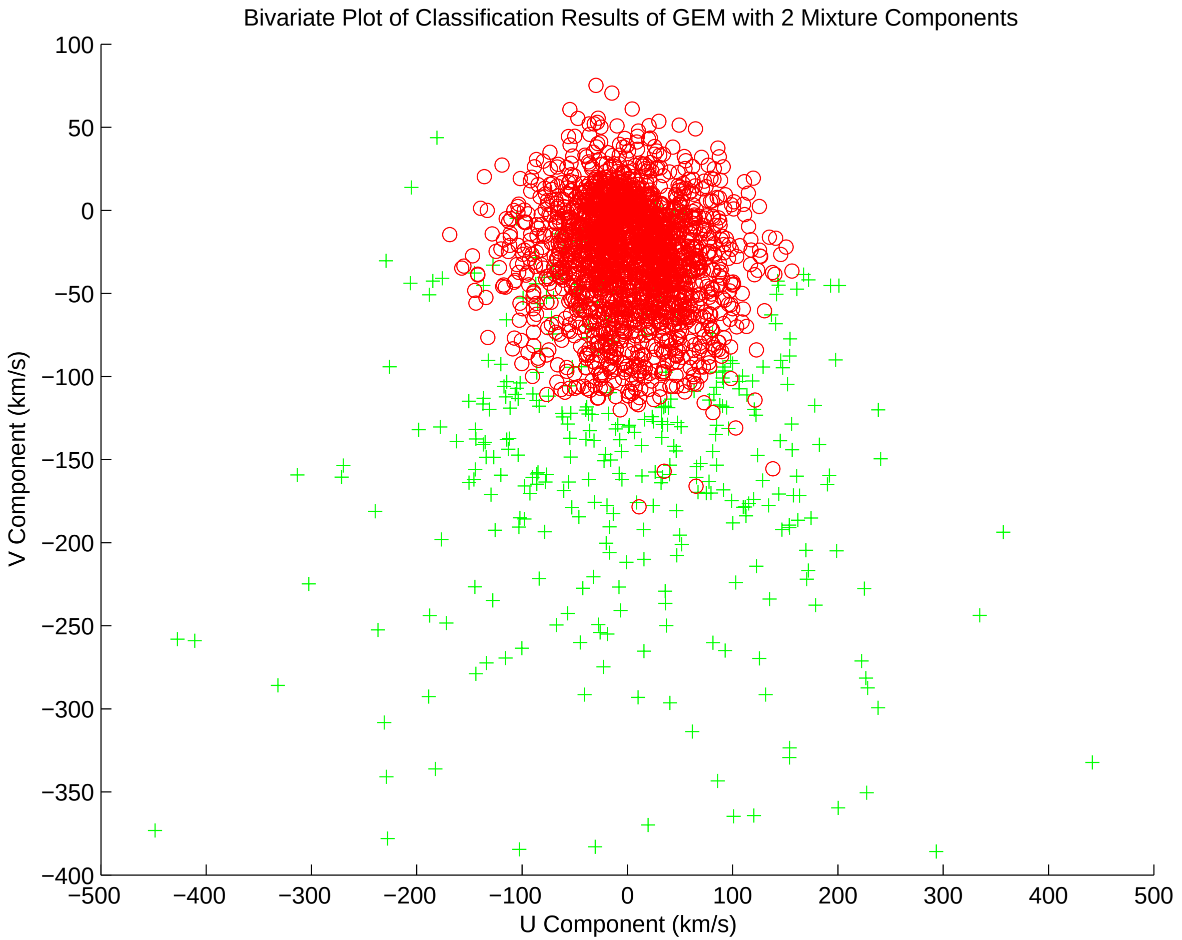
How many Stellar Populations?

- The historical paradigm is that galaxy has two populations of stars
- Disk and Halo
- Differ in ages, motions, metalicities
- Since 1990’s, evidence of three populations
- Thin disk, Thick disk, Halo
- How can we judge which model is best?

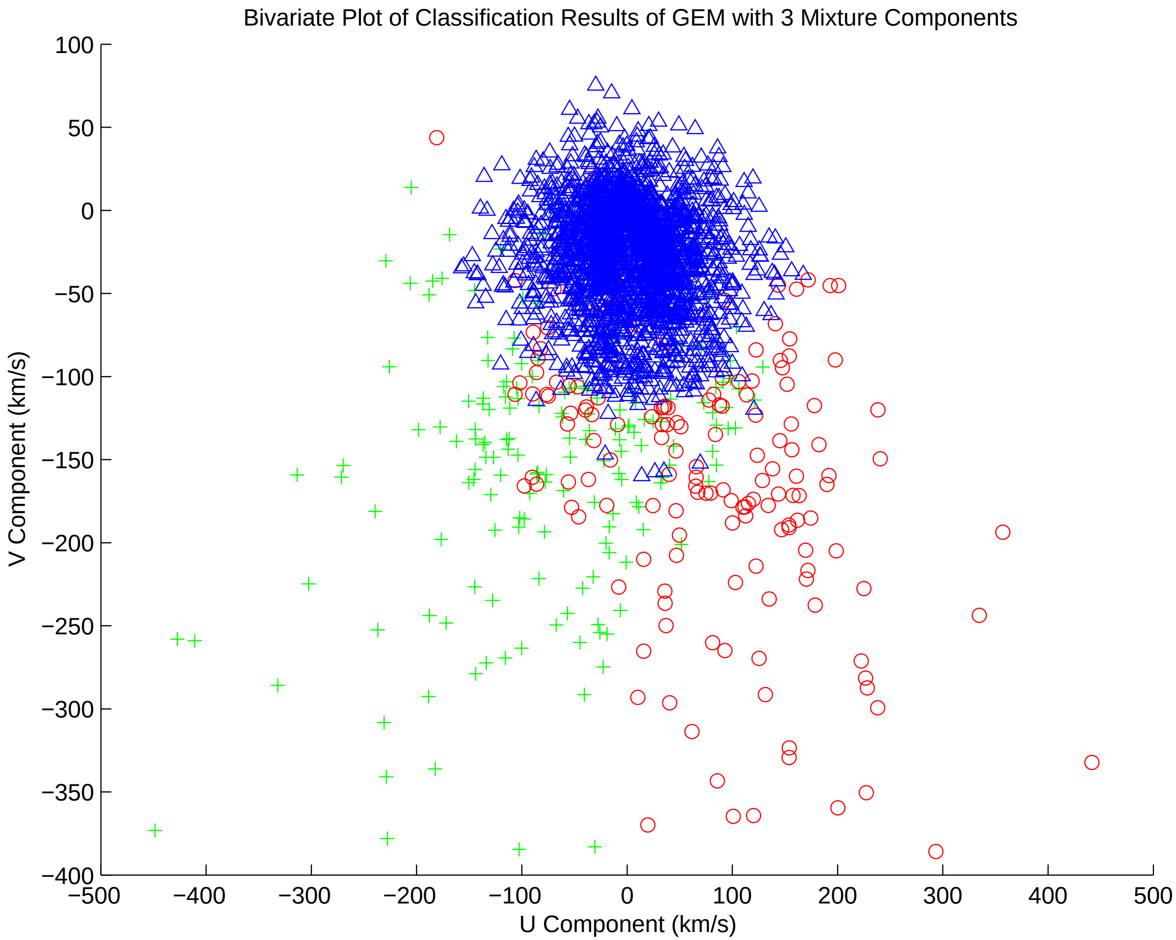
Plot of data set
No Obvious Structure



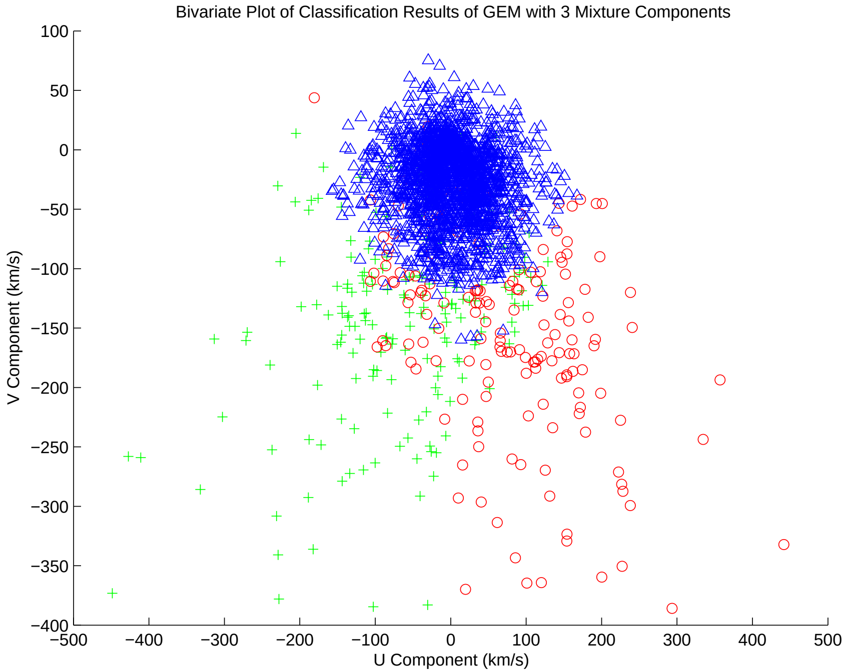
GEM Result with 2 components
AIC = 51024.4 , ICOMP = 51044.2



GEM Result with 3 components
AIC = 51007.6 , ICOMP = 51025.5



Scores indicate 3 components



- Minimum ICOMP and AIC scores indicate that 3 components is preferred over 2
- Agrees with Bensmail et al. (1997) using Bayes factors
- Further evidence to support 3 stellar populations hypothesis

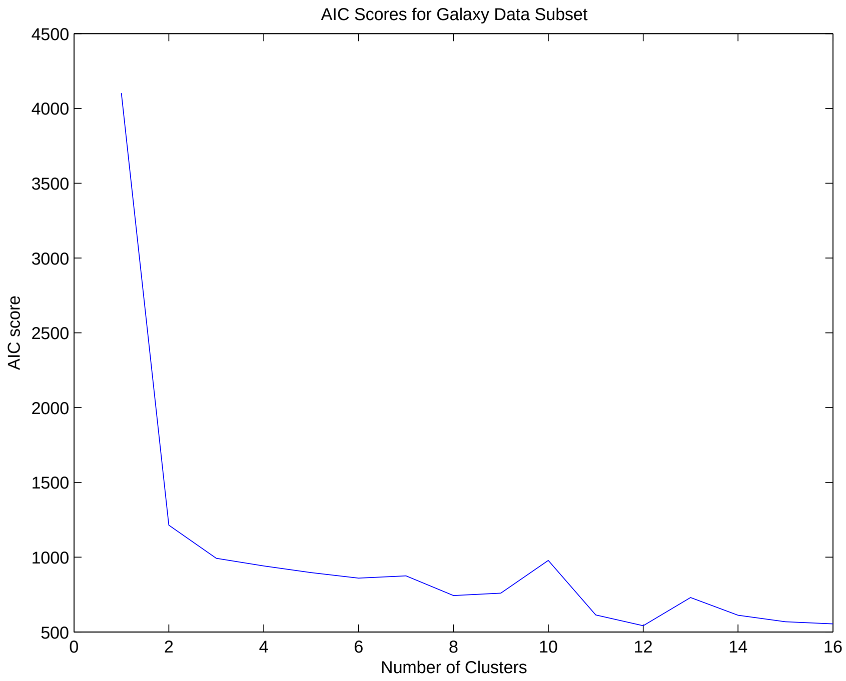
Data that tests classification

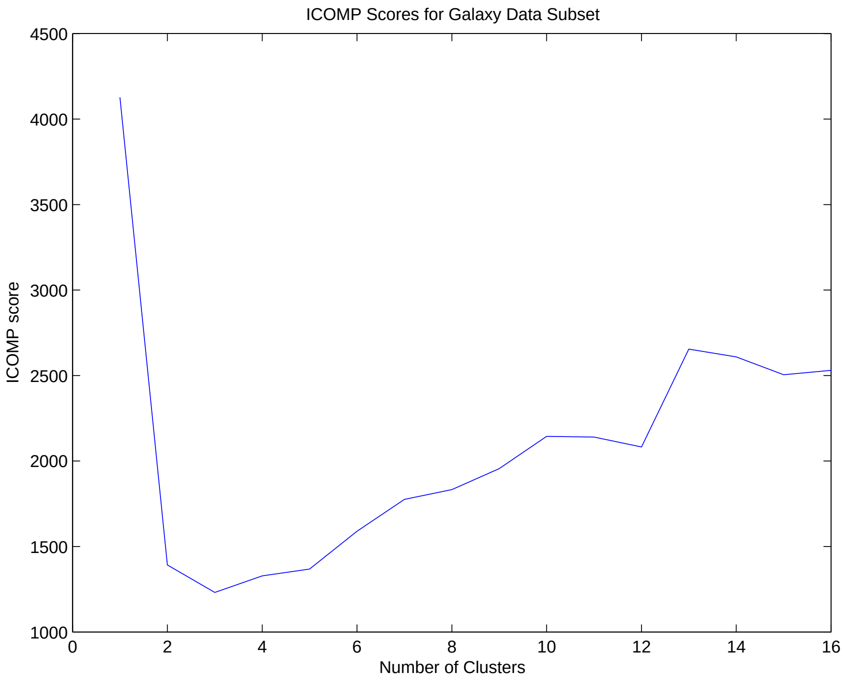
- Zhang and Zhao (2003, 2004) compiled data that can test classification algorithms
- Compiled data from USNO, 2MASS Infrared, and Rosat X-ray RASS catalogs
- Data are 10 dimensional, with parameters describing the intensities in different bands
- Analogous to Optical Color Index (B – V) except that covers visible, IR, and X-Ray

Classification Methods

- Zhang and Zhao used artificial intelligence algorithms
- Combined PCA preprocessing with Backprop NN, Kohonen NN, SVM, LVQ
- Trained NN’s on half of data, tested classification on other half
- I applied scored GEM mixture modeling
- No need for training
- Can identify covariance structure of data

Example:
Galaxy Subset: 173 points
min AIC = 12 , min ICOMP = 3





Classification in Astronomy Spectral Telescope LAMOST in China

Conclusion

- This work represents the first time that information scoring methods have been applied to physics and astronomy data
- I think that information scored regression can have wider application in physics
- GA based log-likelihood analysis can be extended to mixture of kernels (already did calculations) and nonlinear clustering

Questions? Comments?

- This work is in my Ph.D. dissertation
- Online at University of Tennessee library website
- Currently drafting publications
- Contact me: jwicker@utk.edu or jewicker@gmail.com
- I am looking for opportunities to collaborate and apply this research