Classification of gamma-ray bursts

Aakash

A report is submitted to

Department of Engineering Physics

Indian Institute of Technology Hyderabad

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

May 7, 2024

Declaration

I declare that this written submission represents my ideas in my own words, and where ideas or words of others have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited, or from whom proper permission has not been taken when needed.

(Signature)

Aakash

(Name)

EP21BTECH11001

(Roll No.)

Approval Sheet

This report entitled **Classification of gamma-ray bursts** by Aakash is approved for the submission of the B.Tech. project AUG-NOV, 2023.

(Dr. Shantanu Desai) Adviser
Dept. of Engineering Physics
Indian Institute of Technology Hyderabad

Acknowledgements

I express my deep gratitude to Dr. Shantanu Desai for his constant guidance and support. I have learned a lot from this project and I can't thank him enough.

Abstract

Haung et al. in 2018 have analysed the population of 15 galactic double neutron stars regarding the total mass of these systems. They suggest the existence of two subpopulations, and report a likelihood preference of two component GMM over one component GMM. But data size is very small so on the basis of likelihood ratio only we can get overfitting. So to avoid it model must be penalized with free parameters. So Re-examining different statistical tests such as AIC, BIC, cross validation , Bayesaian evidence ratios and penalized EM-test. While this re-examination also confirms preference of two component GMM over one component GMM.

1 Introduction

Galactic double neutron stars (DNSs), also known as binary neutron stars (BNSs) in the gravitational wave (GW) community, are crucial for understanding merging binaries across the Universe. Recent GW observations by LIGO and Virgo have made studying these systems easier. Traditionally, researchers have used observed Galactic DNSs to predict coalescence rates and explore component mass distributions.

A recent study by Huang et al. (2018) focused on the total gravitational masses (M_T) of 15 known DNSs, which are important for predicting merger outcomes and studying the nuclear equation of state. It was noted that there is a preference of two component GMM over one component GMM using Gaussian mixture models and likelihood ratio tests.

They only considered likelihood test as data set is very small so they also have to penalize the parametes. So it must be reevaluated using various tests, like AIC, BIC,Bayesian, Penalized EM-test.

To provide further context, the same criteria are applied to additional examples, such as simulated larger M_T datasets and a dataset of neutron star spins from Patruno, Haskell & Andersson (2017).

2 GMM MODEL SELECTION ON THE DNSMASS DISTRIBUTION

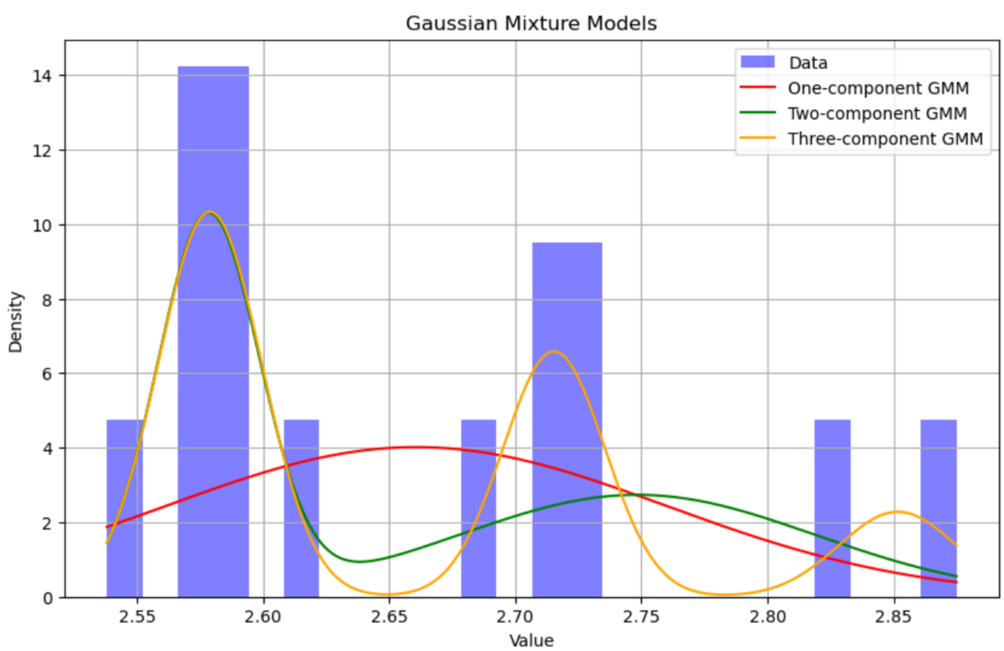

Figure 1: Gaussian Mixture Model for ncomp=1,2,3 for DNS Data

2.1 GMMs and Likelihood Ratios

For N_{data} data points x_n, the basic likelihood function for a GMM with N_{comp} means μ_k, widths σ_k, and component weights $C_k \in [0, 1]$ is the product

$$x_n L(x_n|\{\mu_k, \sigma_k, C_k\}) of$$

$$L(x_n|\{\mu_k, \sigma_k, C_k\}) = \prod_{n=1}^{N_{data}} \sum_{k=1}^{N_{comp}} C_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x_n - \mu_k)^2}{2\sigma_k^2}\right)$$

This can be amended to include measurement errors by assuming each x_n to come from a Gaussian with mean μ_n and width σ_n, then marginalizing over x_n as nuisance variables:

$$L(\mu_n, \sigma_n|\{\mu_k, \sigma_k, C_k\}) = \prod_{n=1}^{N_{data}} \int dx_n \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(x_n - \mu_n)^2}{2\sigma_n^2}\right) \sum_{k=1}^{N_{comp}} C_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x_n - \mu_k)^2}{2\sigma_k^2}\right)$$

$$= \prod_{k=1}^{N_{comp}} C_k \frac{1}{\sqrt{2\pi(\sigma_k^2 + \sigma_n^2)}} \exp\left(-\frac{(\mu_n - \mu_k)^2}{2(\sigma_k^2 + \sigma_n^2)}\right)$$

Now to fit GMMS with $N_{comp} = 1, 2, 3$ to M_T data set we can use two Python packages:

(i) `sklearn.mixture.GaussianMixture` supports basic multicomponent GMM fitting without measurement errors.

(ii) `XDGMM` can also handle known measurement errors

Table 1: Data set of DNS total mass measurements M_T with errors δM_T, reproduced from Huang et al. (2018).

System	M_T	δM_T
J1411+2551	2.538	0.022
J17571854	2.73295	0.00009
J0453+1559	2.734	0.003
J07373039	2.58708	0.00016
J1518+4904	2.7183	0.0007
B1534+12	2.678428	0.000018
J17562251	2.56999	0.00006
J18072500B	2.57190	0.00073
J18111736	2.57	0.10
J1829+2456	2.59	0.02
J1906+0746	2.6134	0.0003
J1913+1102	2.875	0.014
B1913+16	2.828378	0.000007
J19301852	2.59	0.04
B2127+11C	2.71279	0.00013

2.2 AIC and BIC

In general, when adding additional components to a GMM the model likelihood will keep increasing. Hence, this test alone can tempt into overfitting any given data set.
So we use some AIC and BIC to find the best fit model. Considering likelihood these are also considering number of parameters to avoid overfitting. Normally we use AIC but as we have small dataset so we are

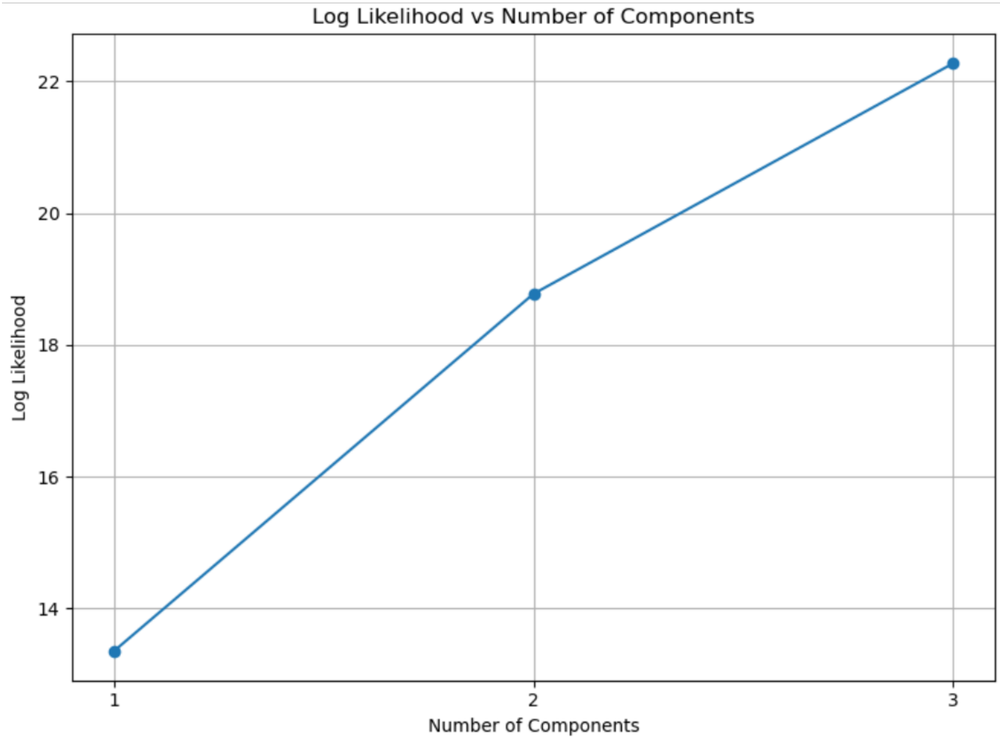

Figure 2: Log-Likelihood using XDGMM(no errors) for DNS data

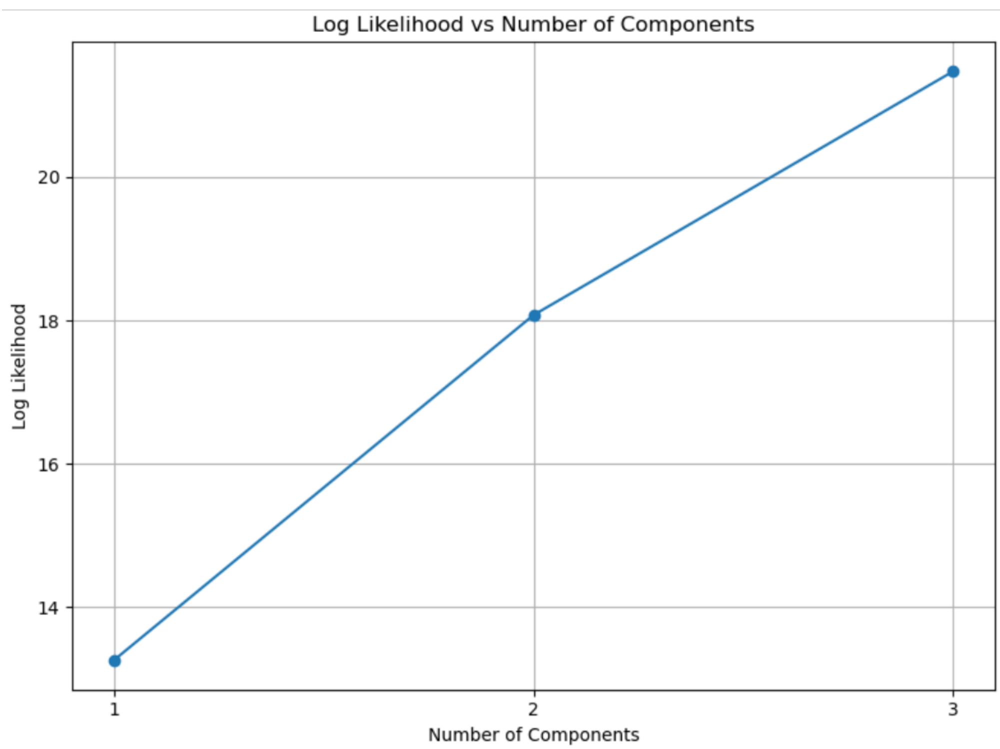

Figure 3: Log-Likelihood using XDGMM(heterosec) for DNS data

Table 2: Log Likelihood, AICc, and BIC values for different numbers of components using XDGMM(no errors) for DNS data

Number of Components	Log Likelihood	AICc	BIC
1	13.353929090443375	-21.70785818088675	-21.291757778682328
2	18.773713168546198	-20.88075967042573	-24.007175331581344
3	22.270035130493707	-4.540070260987413	-22.875668652169733

Table 3: Log Likelihood, AICc, and BIC values for different numbers of components using XDGMM(heterosc.) for DNS data

Number of Components	Log Likelihood	AICc	BIC
1	13.262121040442234	-21.36412853193342	-21.091757778682328
2	18.073713149536178	-18.35135664052321	-21.487175331581344
3	21.471046180313724	-1.791271270485112	-20.135469253621785

using AICc

$$AICc = -2\ln L + 2N_{coeffs} + \frac{2N_{coeffs}(N_{coeffs}+1)}{N_{data} - N_{coeffs} - 1}$$

Bayesian Information Criterion (BIC):

$$BIC = -2\ln L + N_{coeffs}\ln N_{data}$$

Lower the value of AICc and BIC better the model will be.

Steps to Find Log-Likelihood, AICc, BIC

Step 1: Find the log-likelihood using XDGMM or any other suitable method for your data analysis. This involves fitting your data to a Gaussian mixture model and calculating the log-likelihood of the model given the data.

Step 2: Determine the number of free parameters for each component in your Gaussian mixture model. This typically includes parameters for the mean, standard deviation, and weights of each component. The total number of free parameters for a model with n components is $2n - 1$.

Step 3: Use the formulas above to calculate the AICc and BIC:

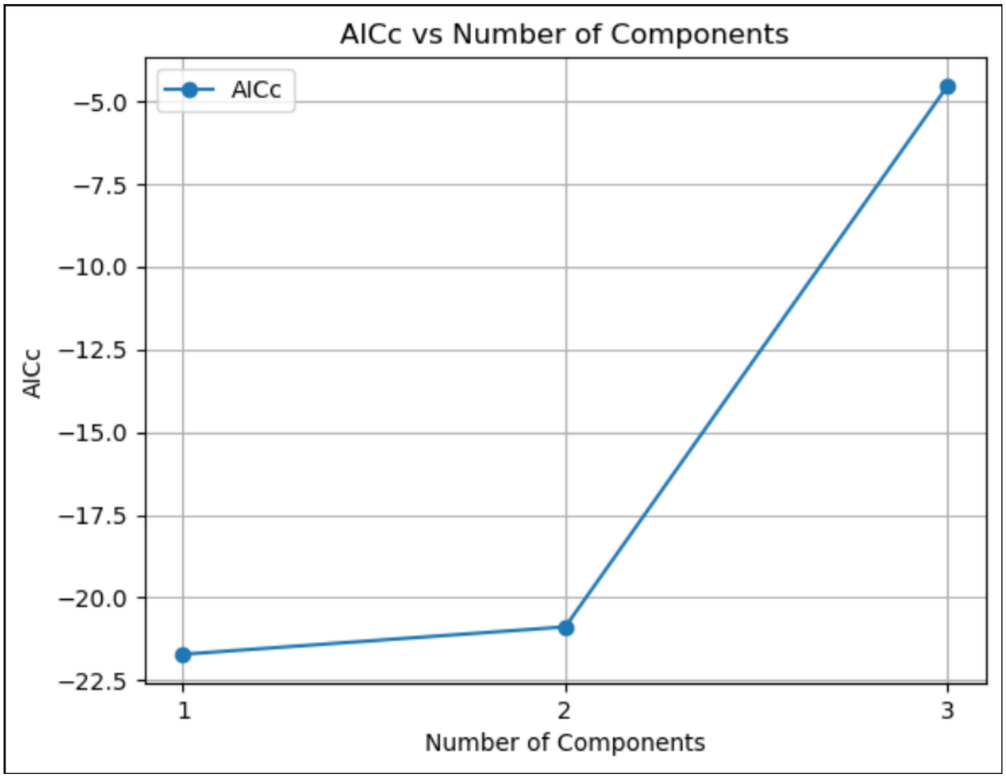

Figure 4: AICc using XDGMM(no error) for DNS data

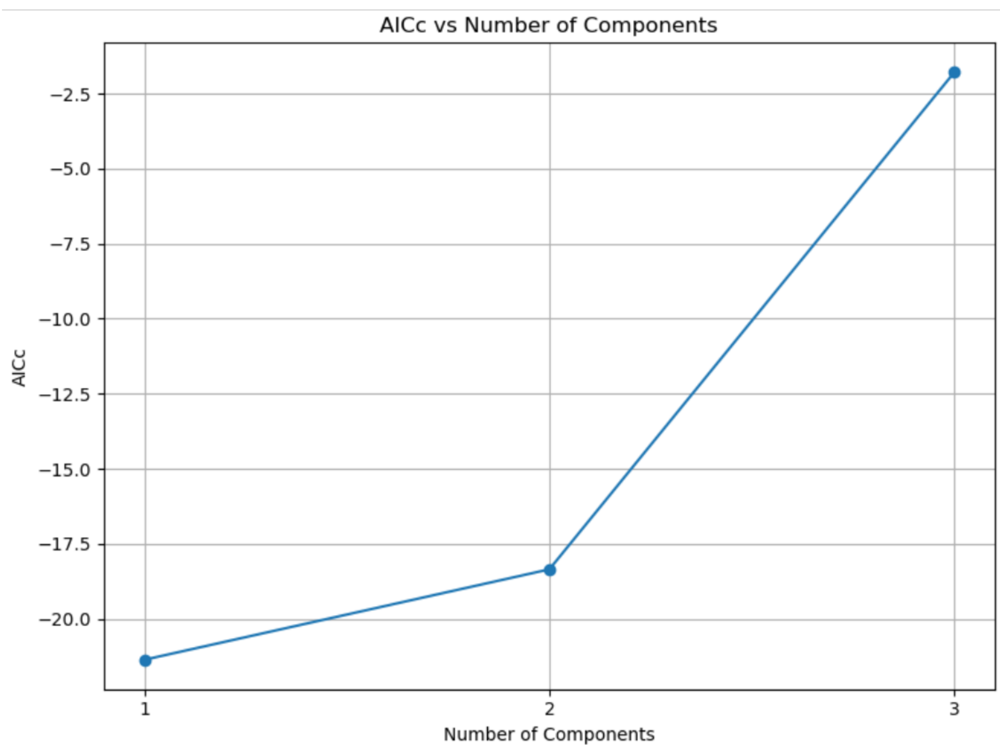

Figure 5: AICc using XDGMM (heterosec) for DNS data

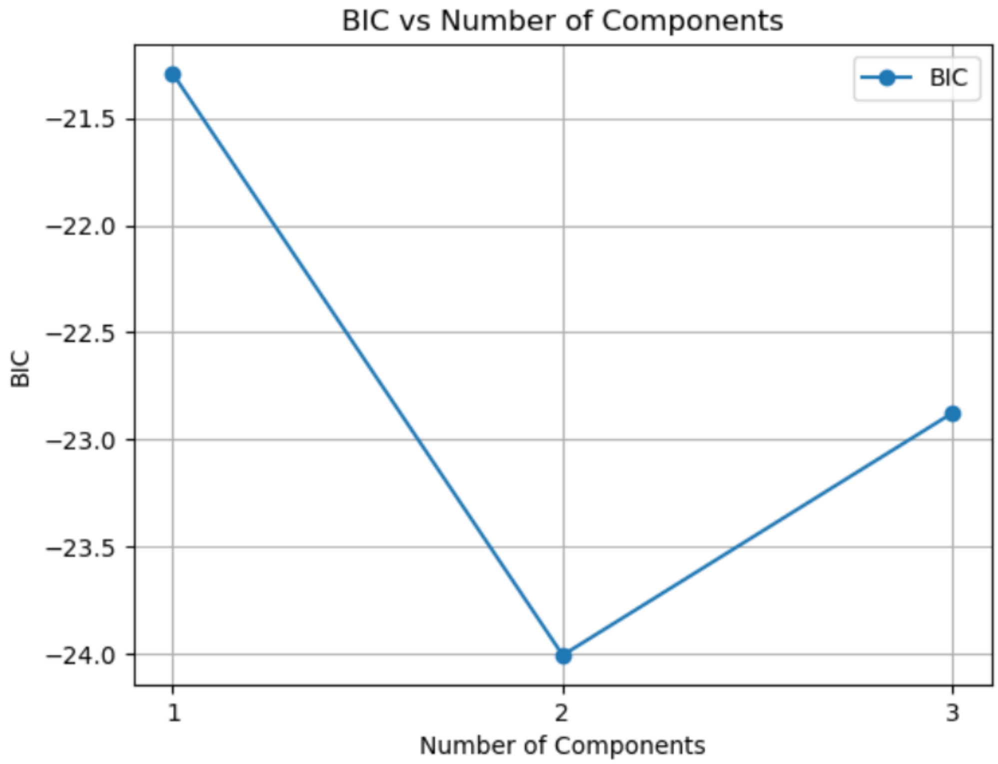

Figure 6: BIC using XDGMM(no error) for DNS data

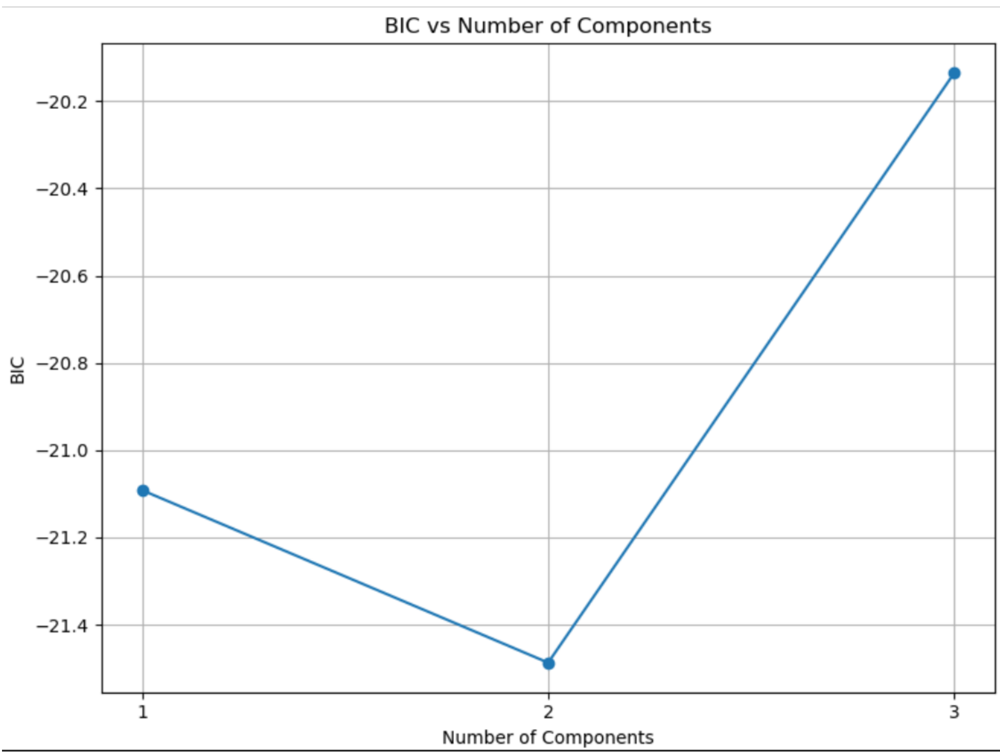

Figure 7: BIC using XDGMM(heterosec) for DNS Data

3 Compairing example: LMXB SPIN FREQUENCIES

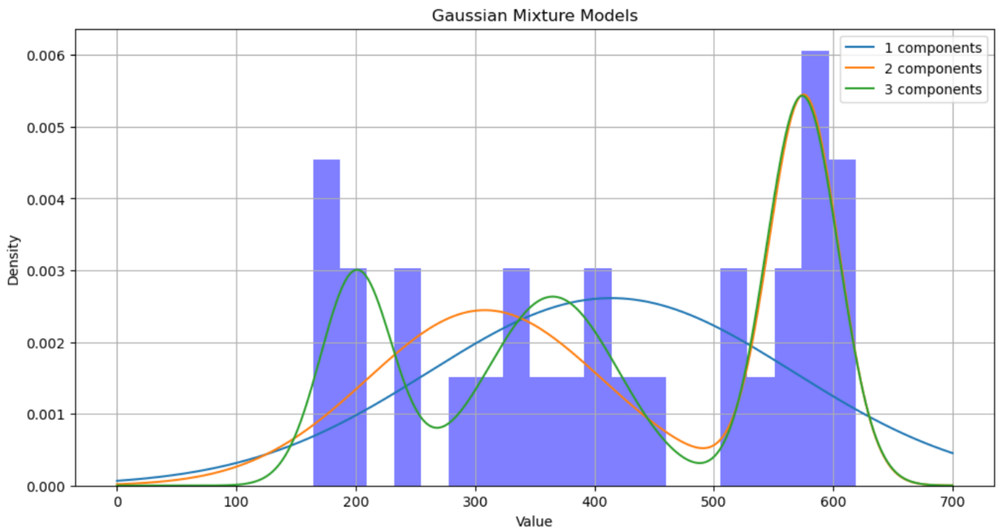

Figure 8: GMM for LMXB SPIN Frequency using XDGMM(no error)

As a comparative example, consider the same GMM analysis applied to a completely different real life data set, which shares the basic statistical properties and model selection question with the DNS study at hand: the distribution of spin frequencies fspin for a population of 29 neutron stars in LMXB systems.Ignoring measurement errors in this case let us analyze it using XDGMM(no errors).

Table 4: Comparison of Log-Likelihood, AICc, and BIC for Different Numbers of Components for LMXB SPIN Freq. data

Number of Comp	Log-Likelihood	AICc	BIC
1	-186.97760493647192	378.4167483344823	380.68980153291676
2	-178.19503304047532	368.9987617331246	373.226545230883
3	-174.3156400980943	371.8312801961886	375.5696468360804

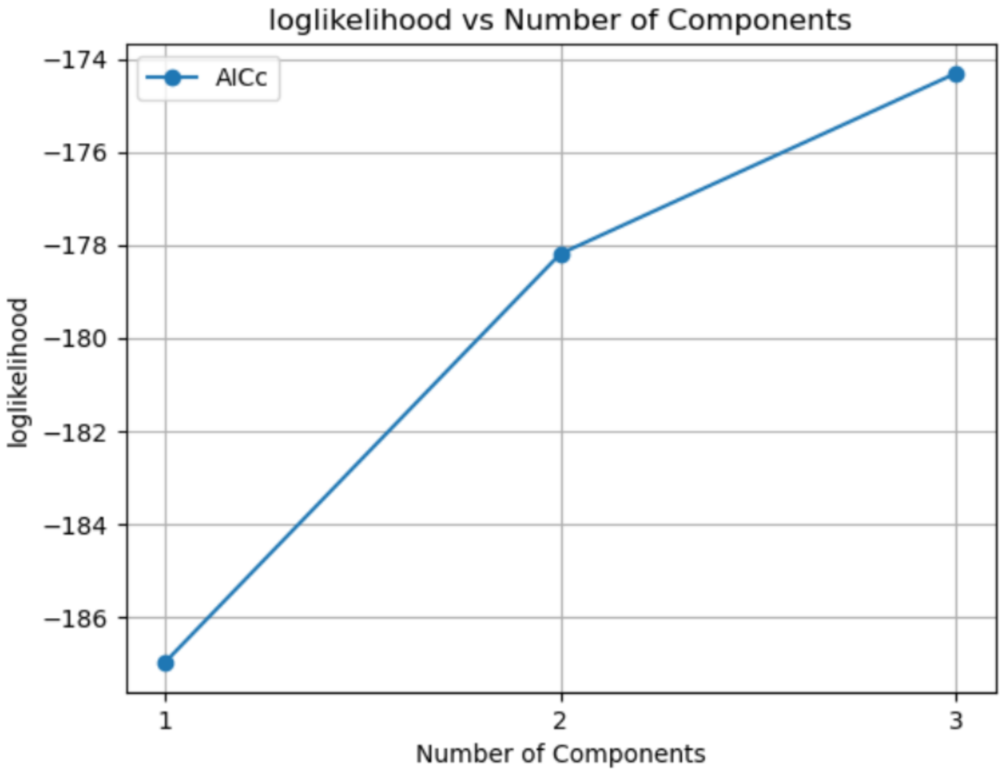

Figure 9: log-likelihood using XDGMM(no error) for LMXB SPIN Freq. Data

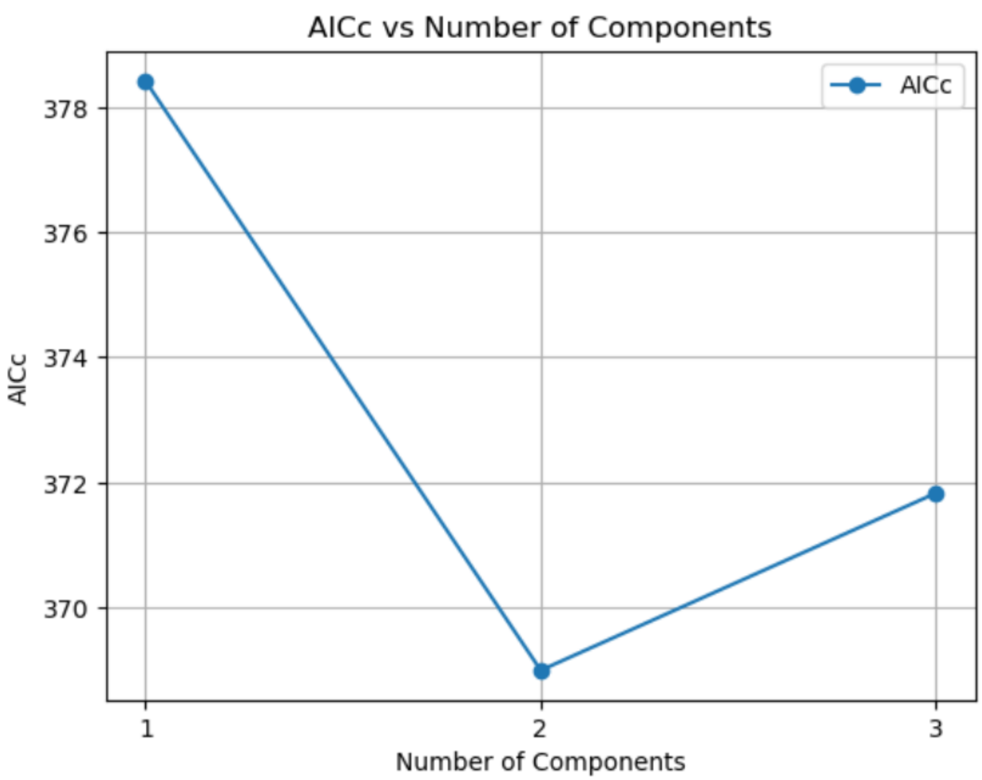

Figure 10: AICc using XDGMM(no error) for LMXB SPIN Freq. Data

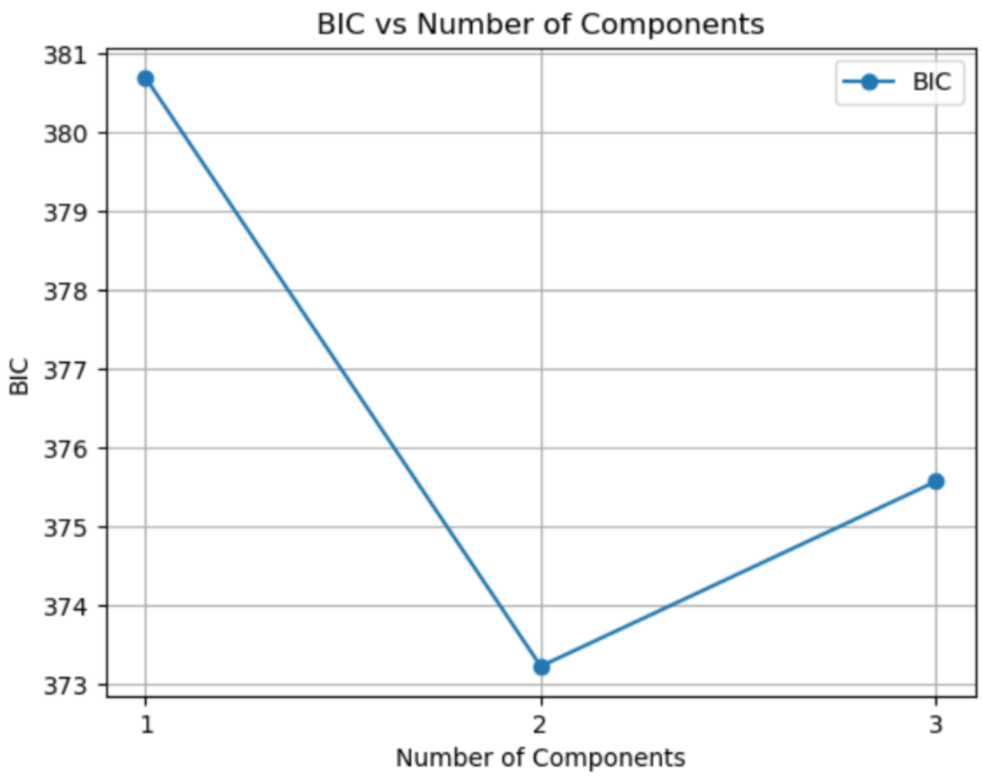

Figure 11: BIC using XDGMM(no error) for LMXB SPIN Freq. Data

From the above result we can see that two component gaussian mixture model is preferred over one component gaussian mixture model

Prior Distribution

In Bayesian statistics, the prior distribution represents our initial belief or knowledge about the parameter of interest before observing any data. It encapsulates any relevant information available before the data collection process. The choice of prior can significantly influence the posterior distribution and subsequent inference. Priors can be informative or uninformative, and they are denoted as $P(\theta)$, where θ is the parameter of interest.

(a) Priors for ncomp=1

Parameter	Value
Coefficient	1
Mean (μ)	500
Standard Deviation (σ)	100

(b) Priors for ncomp=2

Parameter	Value	
	Value 1	Value 2
Coefficient	0.5	0.5
Mean (μ)	300	600
Standard Deviation (σ)	100	30

Table 6: Priors for ncomp=3

Parameter	Value		
	Value 1	Value 2	Value 3
Coefficient	0.5	0.5	0.5
Mean (μ)	200	300	500
Standard Deviation (σ)	100	100	100

Expectation-Maximization (EM) Algorithm

The EM algorithm is an iterative method used to estimate parameters in statistical models where some variables are unobserved (hidden or latent). It iteratively alternates between performing an "expectation" step (E-step), where the expected values of the unobserved variables are computed given the current parameter estimates, and a "maximization" step (M-step), where the parameters are updated to maximize the likelihood function.

Number of Components	Log Likelihood
1	-186.98
2	-178.20
3	-174.32

Table 7: Number of Components vs Log Likelihood for LMXB SPIN Freq. data

Posterior Distribution

After observing the data, we want to update our beliefs about the parameter using Bayes' theorem. The posterior distribution represents our updated belief about the parameter after considering the observed data. It is calculated by combining the prior distribution and the likelihood function through Bayes' theorem. Mathematically, the posterior distribution is given by $P(\theta|x)$, where x is the observed data.

(a) Posterior for $n_{comp} = 1$

Parameter	Value
Coefficient	1
Mean (μ)	414.03
Standard Deviation (σ)	152.71

(b) Posterior for $n_{comp} = 2$

Parameter	Value	
	Value 1	Value 2
Coefficient	0.6029	0.39708607
Mean (μ)	307.72	575.45322653
Standard Deviation (σ)	98.23	29.39170582641993

Table 9: Posterior for ncomp=3

Parameter	Value		
	Value 1	Value 2	Value 3
Coefficient	0.16224989	0.43123332	0.40651679
Mean (μ)	184.88018036	348.88381074	574.60693939
Standard Deviation (σ)	13.245383111574958	69.74699050084328	29.896924435430158

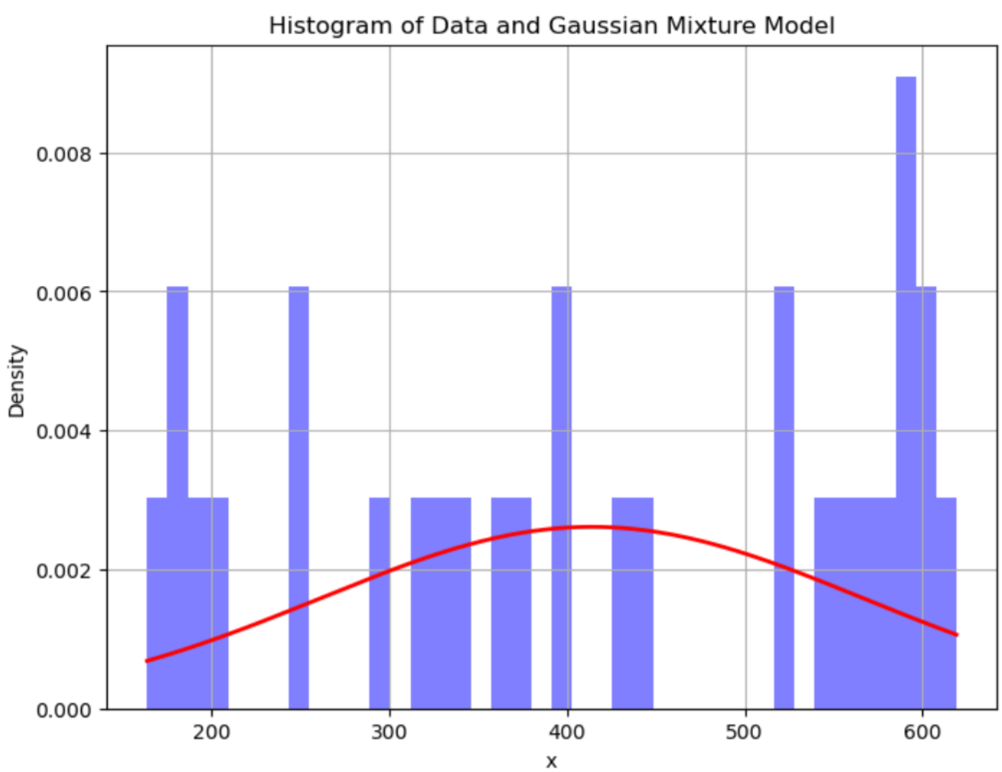

Figure 12: GMM for ncomp=1 for LMXB Spin Freq. Data

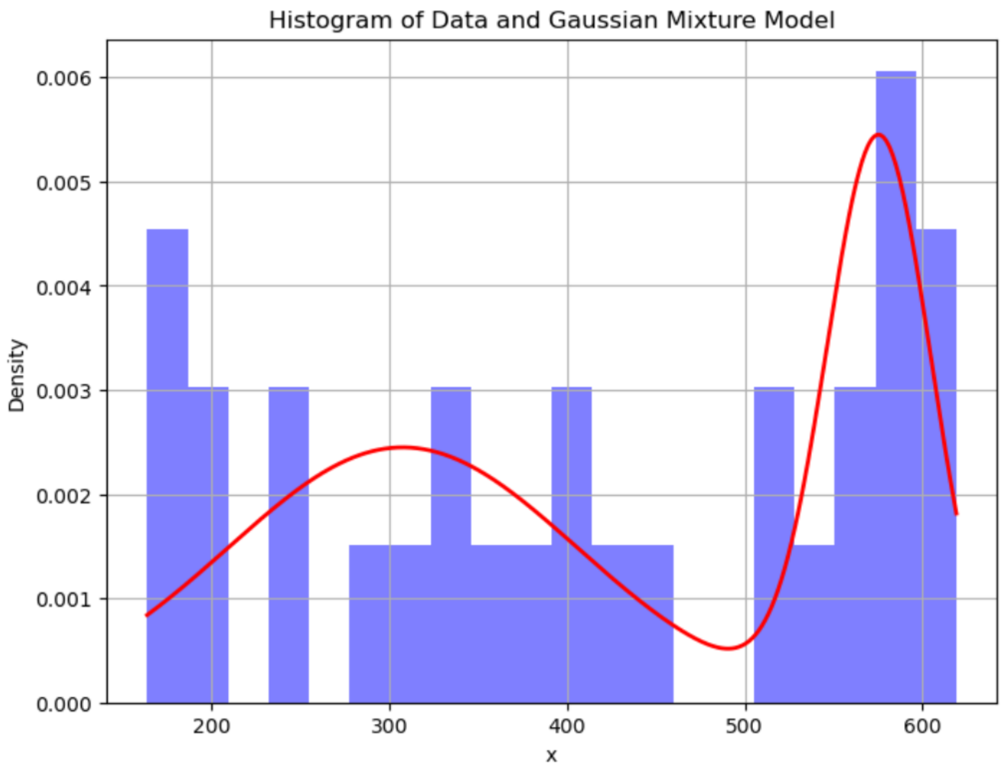

Figure 13: GMM for ncomp=2 for LMXB Spin Freq. Data

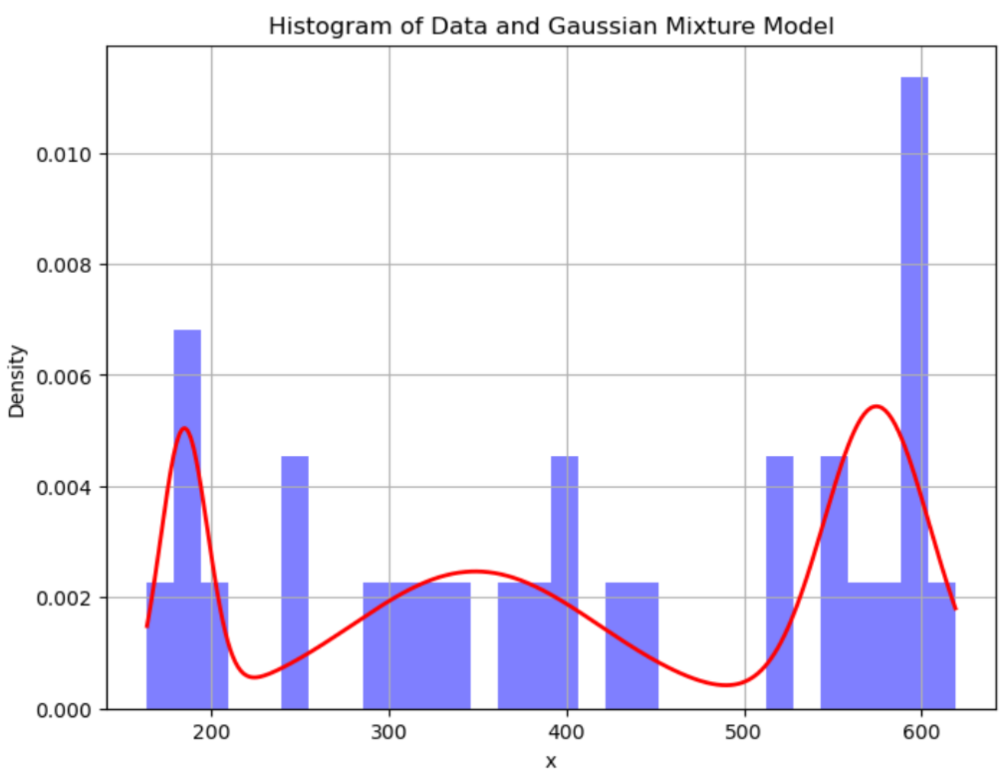

Figure 14: GMM for ncomp=3 for LMXB Spin Freq. Data

Table 10: Data set of LMXB spin frequencies f_{spin}

Source	f_{spin} (Hz)
4U 1728−34	363
KS 1731−260	524
IGR J17191−2821	294
4U 1702−429	329
SAX J1750.8−2900	601
GRS 1741.9−2853	589
EXO 0748−676	552
4U 1608−52	619
4U 1636−536	581
MXB 1659−298	567
Aql X−1	550
IGR J00291+5934	599
PSR J1023+0038	592
XSS J12270−4859	593
SAX J1808.4−3658	401
XTE J1751−305	435
XTE J0929−314	185
XTE J807−294	190
XTE J1814−338	314
HETE J1900.1−2455	377
Swift J1756.9−258	182
SAX J1748.9−2021	442
NGC 6440 X−2	206
IGR J17511−3057	245
Swift J1749.4−2807	518
IGR J17498−2921	401
IGR J18245−245	254
MAXI J0911−655	340
IGR J17602−6143	164

Conclusions

After Re-estimation by considering penalising tests (AICc, BIC) and Log-Likelihood two component GMM is better fit over one component GMM. On applying EM and Bayesian on LMXB SPIN Freq. data also give preference to two component GMM over one component GMM. But the data set is very small so we cannot clearly say that two component will get prefernce over one component or not we need more data to comment on it. Hope in future we will get more data and will re-analyze it and tell that two component GMM is preferred over one component GMM or not.

References

1. Galaxy-Double Neutron Star.
2. Duke EM.
3. GFG.
4. ASTROML.