PROCEEDINGS OF SCIENCE

Classifying Gravitational Waves with GMM and GLM techniques

Sourav Dutta,[a,*] **Reetanjali Moharana**[a] **and Shwetabh Biswas**[a]

[a] *Department of Physics, Indian Institute of Technology Jodhpur, Jodhpur-342030, India*

E-mail: p22ph008@iitj.ac.in

The first Gravitational Wave was discovered on September 14, 2015, by the Advanced Laser Interferometer Gravitational-wave Observatory (LIGO), and since then, we have observed 93 such events significantly. With these 93 events, we have studied the classification based on their parameters using Generalized Linear Models (GLM) and Gaussian Mixture Models (GMM). GMM is a mixture of the weighted sum of different Gaussian distributions that describe the number of classes in the data. GLM is a statistical modeling technique identifying various relationships between a response variable and predictor variables. These relationships can be normal, linear, logistic, Poisson, exponential, etc.

38th International Cosmic Ray Conference (ICRC2023)
26 July - 3 August, 2023
Nagoya, Japan

*Speaker

PoS(ICRC2023)1570

Observing Run	Starting Date	Ending Date
O1	12 September 2015	12 January 2016
O2	30 November 2016	25 August 2017
O3a	1 April 2019	30 September 2019
O3b	30 November 2019	27 March 2020

Table 1: Details of observing runs

1. Introduction

GW150914 is the first-ever gravitational wave resulting from a black hole-black hole (BH-BH) merger, which was observed on September 2015 by Laser Interferometer Gravitational-wave Observatory (LIGO). Since then, the joint search by the observatories at LIGO Livingston Observatory (LLO) in Louisiana, LIGO Hanford Observatory (LHO) in Washington, Virgo in Pisa, Italy, and KAGRA in Japan has been able to identify several GW merger events.

The observed GWs so far fall under the compact binary inspiral merger category, which is produced by an orbiting pair of massive and dense objects like a neutron star (NS), black hole (BH), and a white dwarf (WD). The most three possible classes are binary neutron star (BNS), binary black hole (BBH), and neutron star-black hole binary (NSBH). The reliable mechanism to identify the above three classes of GW events is through their masses. GW170817, the NS-NS merger that was detected on 17 August 2017, had also been identified as a short gamma-ray burst due to its electromagnetic (EM) radiation. Apart from this, there has been no GW event identified with significant EM radiation to date.

We propose here to search for the three different classes using Generalized Linear Models (GLM) and Gaussian Mixture Models (GMM) to the parameter set of the events. The Gravitational Wave Open Science Center (GWOSC) website[1][1] contains the GW events observed under three runs as listed in table 1. It is worthwhile to note that the number of events significantly increased in the third run, as compared to the first two. We used these data for our search; a detail of the data selection is mentioned in the dataset section.

2. Dataset

The data used in this work has been taken from the Gravitational Wave Open Science Center (GWOSC) website. 181 events in total have been publicly released. Out of those events, we have considered a total of 93 events that are identified as confident GW events for our analysis. The details of the number of events that have been selected from each catalog, along with the FAR threshold are given in table 2.

The events are selected by the False Alarm Rate (FAR) parameter by the GstLAL matched-filter search pipeline. The formula is given in equation 1.

$$FAR = \frac{NP(\log L^* \geq \log L|noise)}{T} \tag{1}$$

[1] https://gwosc.org/

Event release	FAR threshold	Number of events
GWTC-1-confident[2]	12.2 yr^{-1}	11
GWTC-2[3]	2.0 yr^{-1}	3
GWTC-2.1-confident[4]	2.0 day^{-1}	44
GWTC-3-confident[5]	2.0 day^{-1}	35
Total		93

Table 2: Event selection details

In the above equation, N is the number of observed candidates, T is the live time of the experiment, $\log L|noise$ is the log-likelihood ratio's noise model, and $\log L^*$ is the log-likelihood ratio of the observed candidate. Hence $P(\log L^* \geq \log L|noise)$ is the probability of observing a candidate with a log-likelihood ratio greater than or equal to $\log L$. The events that have FAR less than a specific threshold are considered as confident GW events.

3. Analysis Methods

We used the GMM method to classify the GW events following[6]. We then performed the known maximum likelihood classification method, the GLM [7], to search for different classes of GWs. We studied GLM on the scattering distribution of the mass parameters, M_C vs χ_{eff}, η vs χ_{eff}, M_C vs η, where M_C, η and χ_{eff} are calculated as:

$$\begin{aligned} M_C &= (m_1 m_2)^{3/5}(m_1+m_2)^{-1/5} \\ \eta &= m_1 m_2 (m_1+m_2)^2 \\ \chi_{eff} &= \frac{\chi_1 \cos\alpha_1 + \frac{m_2}{m_1}\chi_2 \cos\alpha_2}{1+\frac{m_2}{m_1}} \end{aligned} \tag{2}$$

where m, χ, and α are the mass, spin magnitude, and tilt angle of the respective components ($m_1 \geq m_2$).

3.1 Gaussian Mixture Models (GMM)

We use the machine learning python package scikit-learn[2] to use the GMM method. This method helps to avoid the classic binning method. GMM is a mixture of the weighted sum of different Gaussian distributions that describe the number of classes in the data. For k Gaussian components $(C_i, i = 1, 2...., k)$ the i^{th} Gaussian distribution for x number of data is,

$$N(x|\mu_i, \Sigma_i) = \frac{1}{2\pi}\frac{1}{\sqrt{\Sigma_i}} exp\{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\} \tag{3}$$

with μ_i, and Σ_i as mean and covariance, respectively. Hence, the probability distribution function for complete data $X = x_j (j = 1, 2. . . ., N)$

[2] https://scikit-learn.org/stable/

Value of Δ_i	Remarks on model
Δ_i=0	Best model
$0 < \Delta_i < 2$	Model is also supported
$2 < \Delta_i < 6$	Positive evidence against the model
$\Delta_i > 6$	Strong evidence against the model
$\Delta_i > 10$	Very strong evidence against the model

Table 3: Model selection criteria

$$P(X|\omega, \mu, \Sigma) = \sum_{j=1}^{N} \left(\sum_{i=1}^{k} \omega_i \; N(x_j|\mu_i, \Sigma) \right), \tag{4}$$

where w_i is the weight of i^{th} Gaussian distribution.

The GMM method uses an iterative algorithm alternating between **E**xpectation and **M**aximization steps (E and M steps) to estimate the parameters of i^{th} number of Gaussian distributions as explained in[6].

3.2 Selection Method

The Bayesian information criterion (BIC), also known as the Schwarz[8] criterion is used to choose a model from a finite set of Gaussian distribution models. The model with the smallest value of BIC (BIC_{min}) is the preferred model. BIC is given by:

$$BIC = p \ln N - 2 \ln P_{max} \tag{5}$$

where P_{max} is the maximum likelihood (ML) achieved by the models, p is the number of parameters of the model, and N is the sample size.

To compare the two or more models, BIC difference (Δ_i) is calculated

$$\Delta_i = BIC_i - BIC_{min} \tag{6}$$

The remarks on model selection based on Δ_i value is shown in table 3

3.3 Generalized Linear Models (GLM)

GLMs can model a wide range of relationships between the response and predictor variables, including linear, logistic, Poisson, and exponential relationships. Generalized Linear Models (GLM) are powerful, decisive tools that offer flexible and computationally attractive models for large datasets. Model choice is made important due to this resulting flexibility and complexity. GLM has been used in a wide range of areas, from longitudinal data analysis to classification of Gamma Ray Bursts[9].

3.4 GLM analysis

The "flexmix" model, available in R[7], has been used for GLM analysis. The "flexmix" model is a general framework for finite mixtures of regression models, implementing the EM algorithm.

Parameter Space	Components for minimum BIC	
	Real	Simulated
χ_{eff}	2	1
m_1	3	2
m_2	3	3
M_C	2	3
η	4	2
M_C-χ_{eff}	2	1
η-χ_{eff}	2	1
M_C-η	3	2

Table 4: Variation of BIC_{min} for various parameters

The E-step and all data handling are provided, while the user can supply the M-step to define new models easily. The default generalized linear models (glm) method has been used for clustering the following parameter space, which consists of a Normal distribution for modeling the dependent variable Y, with an identity link function, such that:

$$E(Y) = \mu = kX; Y \sim N(\mu, \sigma^2)$$
$$f(Y, \mu) = \frac{1}{2\pi\sigma^2} exp[-\frac{1}{2\sigma^2}(Y - \mu)^2]$$

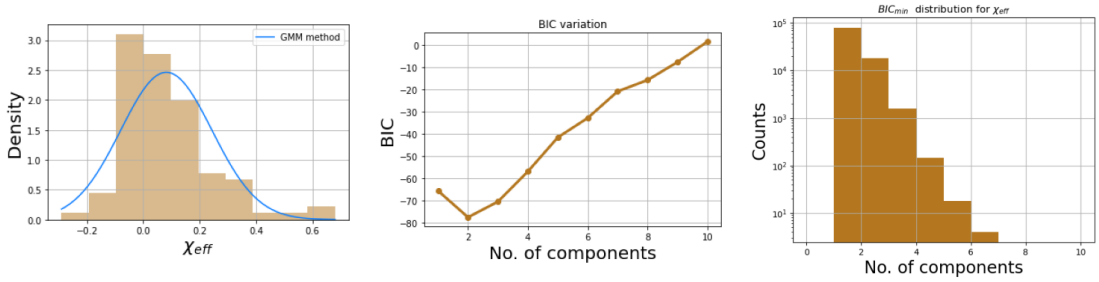
where $f(Y, \mu)$ denotes the probability distribution function of Y, with X being the independent variable, and k is the coefficient to be estimated. The link function is the identity function.

4. Results and Discussions

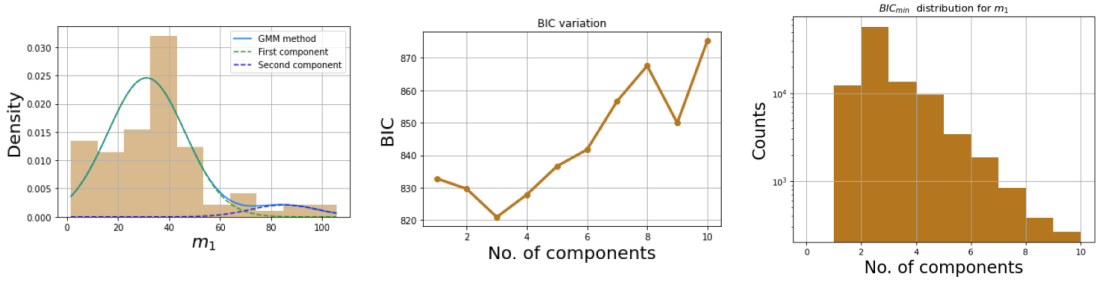
The number of components/classes in the GW events are identified according to the BIC value. table 4 summarizes the number of classes for which we got the minimum BIC value for the GMM and GLM analysis with 10^{-6} as the convergence threshold of the log-likelihood. The results of our analysis are shown in fig. 1 and 2. The left-hand-side plots show the distribution of the actual data. The central plots show the BIC variation on the actual data points. A summary is shown in table 4.

The data set contains huge error bars due to the uncertainty in noise and signal. To understand the effect of these error bars on classification, we have generated 10^5 values for each real data point using random Gaussian distribution, with the observed value as the mean and the error as the standard deviation. The current system time (in ms.), from epoch (1970), is used as seed during the random points generation. The right-hand side plots show the distribution of BIC_{min} values after 10^5 iterations on the simulated dataset.

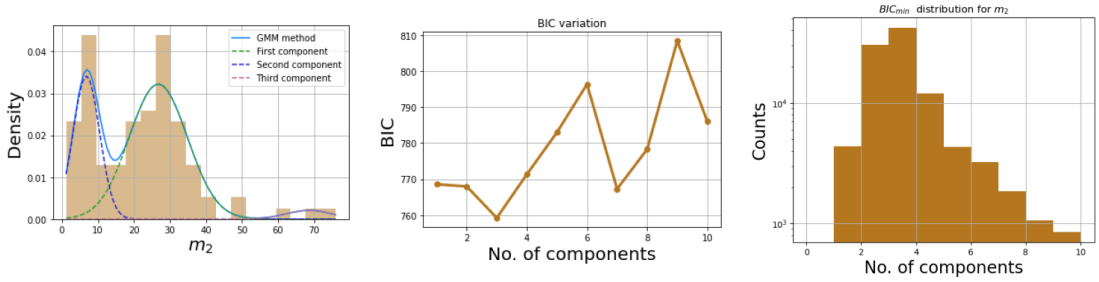
From the table 4, the GWs are classified into three classes based on the m_2 and M_C parameter spaces. While from the other parameter spaces like m_1, η & $M_C - \eta$, the GWs can be classified into two classes.



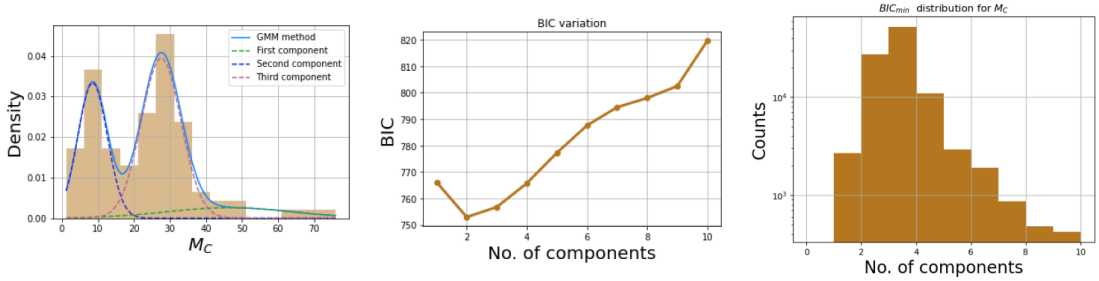
(a) Analysis of χ_{eff} values



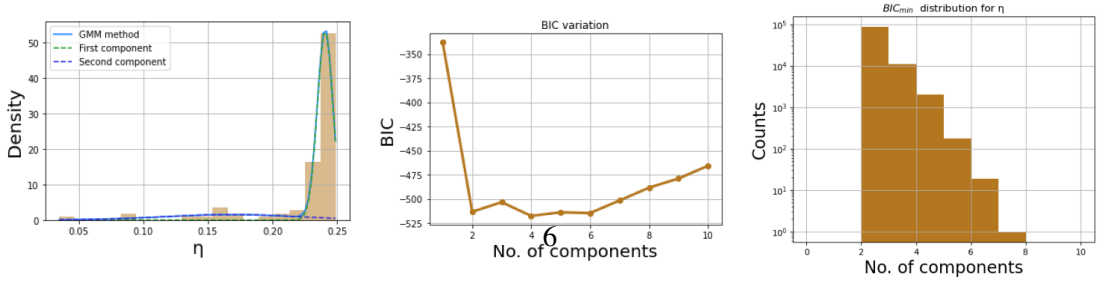
(b) Analysis of m_1 values



(c) Analysis of m_2 values

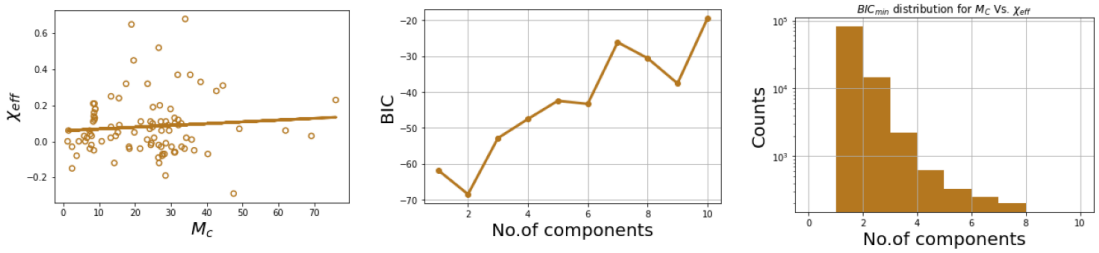


(d) Analysis of M_C values

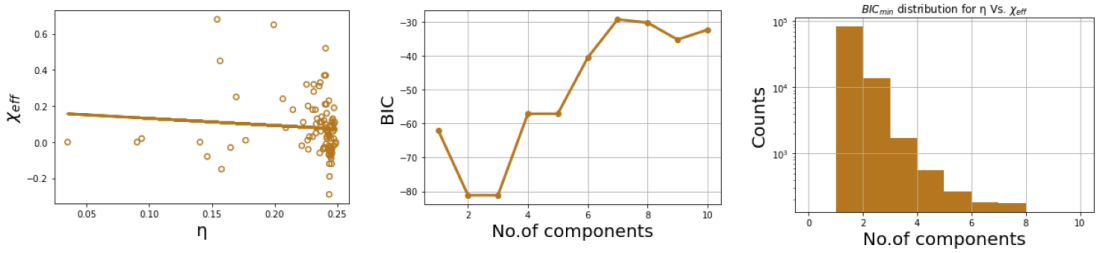


(e) Analysis of η values

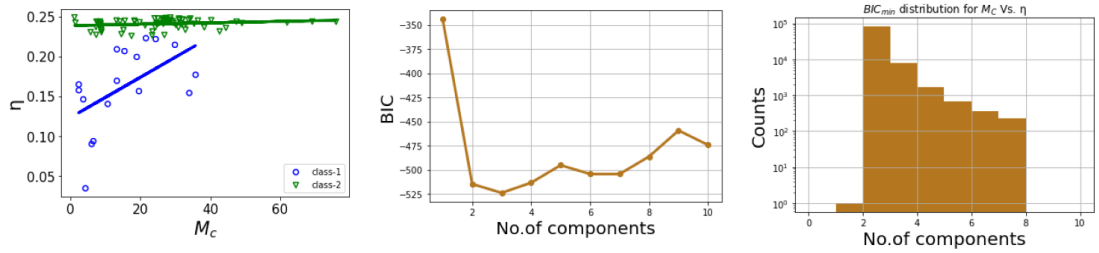
Figure 1: Result of GMM analysis.



(a) Analysis of M_C Vs. χ_{eff}



(b) Analysis of η Vs. χ_{eff}



(c) Analysis M_C Vs. η

Figure 2: Result of GLM analysis. The left-hand-side plots show the scatter plot of the actual data points. The central plots show the BIC variation on the actual data points. The right-hand side plots show the distribution of BIC_{min} values after 10^5 iterations on the simulated dataset.

References

[1] Abbott, R., et al.*Open data from the third observing run of LIGO, Virgo, KAGRA and GEO.*, *arXiv preprint* [gr-qc/2302.03676].

[2] Abbott, B. P., et al. *GWTC-1: a gravitational-wave transient catalog of compact binary mergers observed by LIGO and Virgo during the first and second observing runs.*, *Physical Review X 9.3 (2019): 031040* [astro-ph.HE/1811.12907v3].

[3] Abbott, R., et al. *GWTC-2: compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run. Physical Review X 11.2 (2021): 021053* [gr-qc/2010.14527v3].

[4] Abbott, R., et al. *GWTC-2.1: Deep extended catalog of compact binary coalescences observed by LIGO and Virgo during the first half of the third observing run. arXiv preprint* [gr-qc/2108.01045].

[5] Abbott, R., et al. *GWTC-3: compact binary coalescences observed by LIGO and Virgo during the second part of the third observing run. arXiv preprint* [gr-qc/2111.03606].

[6] Zhang, Zhi-Bin, et al. *Classifying gamma-ray bursts with Gaussian Mixture Model. Monthly Notices of the Royal Astronomical Society 462.3 (2016): 3243-3254* [astro-ph.HE/1603.03680v3].

[7] Gruen B, Leisch F (2023). *flexmix: Flexible Mixture Modeling. R package version 2.3-19, https://CRAN.R-project.org/package=flexmix*.

[8] Schwarz, Gideon. *Estimating the dimension of a model." The annals of statistics (1978): 461-464*.

[9] Dutta, S., Moharana, R., Kumar, M..*Generalized Linear Models of T_{90}-T_{50} relation to classify GRBs. arXiv preprint* [astro-ph.HE/2305.03947v2].