Model selection for Galactic Double Neutron Star total masses

PH6130 course project
by
Shreeprasad Bhat (AI20MTECH14011)
M.Tech Artifical Intelligence
IIT Hyderabad

***Abstract*—In this project, we analyse the population of 15 known galactic Double Neutron Stars (DNSs) regarding the total masses of these systems. We fit the distribution of total masses for three models, a single Gaussian distribution, two-component Gaussian mixture model and three component Gaussian mixture model. Multiple statistical tests are performed for model selection. A pure likelihood ratio test prefers Gaussian mixture model with 3 components and than with two components, but for small datasets this test can encourage over-fitting. This can be avoided by penalizing models with higher number of free parameters. This can be achieved through several simple and well established statistical test, including including information criteria (AICc, BIC), cross-validation and Bayesian evidence ratios. We conclude that even-though a two-component and three-component mixture is consistent with the data, the model selection criteria consistently indicate that there is no robust preference for it over a single-component fit.**

Contents

1 Reference

Keitel, D., 2019. Galactic double neutron star total masses and Gaussian mixture model selection. Monthly Notices of the Royal Astronomical Society, 485(2), pp.1665-1674.

2 Introduction

The population of galactic Double Neutron Stars (DNSs) or binary neutron stars (BNSs), as the gravitational wave (GW) community prefers to call them is of high interest as a locally accessible predictor for the population of merging binaries in the wider Universe, which has recently become accessible to GW observations with LIGO and Virgo. Traditionally, a lot of work has focused on using the observed galactic sample to predict coalescence rates, though the distribution of component masses has also been studied. Here we consider the total gravitational masses M_T of 15 known DNSs. M_T is of special interest in predicting the fate of binary merger remnants and for studies of the nuclear equation of state (EoS) point out an apparent bimodality in the distribution of M_T , and with the help of Gaussian mixture models and different statistical test and caution against relying on likelihood-ratio tests alone, especially when applied to small data sets. First, for completeness, (i) visual inspection of the data set and (ii) GMM fitting and likelihood-ratio tests are briefly summarised. The additional hypothesis test methods include (iii) information criteria (AICc and BIC) that penalise under constrained parameters, (iv) a cross-validation test to understand the impact of individual DNS systems on model selection, (v) Bayesian evidence computation through nested sampling.

3 Data set and visual inspection

System	$M_T(M_\odot)$
J1411+2551	2.538(22)
J1757-1854	2.73295(9)
J0453+1559	2.734(3)
J0737-3039	2.58708(16)
J1518+4904	2.7183(7)
B1534+12	2.678428(18)
J1756-2251	2.56999(6)
J1807-2500B	2.57190(73)
J1811-1736	2.57(10)
J1829+2456	2.59(2)
J1906+0746	2.6134(3)
J1913+1102	2.875(14)
B1913+16	2.828378(7)
J1930-1852	2.59(4)
B2127+11C	2.71279(13)

Fig. 1: Dataset

Histograms of the M_T data set are shown in Fig. 2. It compares the the data by binning into 12 bins and 24 bins. Visually, $N_{bins} = 12$ makes a two-component fit appealing, while $N_{bins} = 24$ might even tempt the viewer into fitting three components. Note that the total number of data points is only 15.

4 Modelling

To probe possible structure in the M_T distribution, below we fit the data with a Gaussian distribution model, a double Gaussian distribution model and triple Gaussian distribution model, respectively.

The probability distribution function of Gaussian distribution model is given by

$$P(M_T, M_0, \sigma_0) = \frac{1}{\sqrt{2\pi}\sigma_0} exp\left(-\frac{(M_T - M_0)^2}{2\sigma_0^2}\right) \tag{4.0.1}$$

and the best fit model from maximum likelihood estimate(MLE) is shown in Fig 3. and corresponding parameters are $(M_0, \sigma_0) = (2.67, 0.10)$.

For the double Gaussian distribution model we have

$$\begin{aligned} P(M_T, M_1, M_2, \sigma_1, \sigma_2, C) &= \frac{C}{\sqrt{2\pi}\sigma_1} exp\left(-\frac{(M_T - M_1)^2}{2\sigma_1^2}\right) \\ &+ \frac{1-C}{\sqrt{2\pi}\sigma_2} exp\left(-\frac{(M_T - M_2)^2}{2\sigma_2^2}\right) \end{aligned} \tag{4.0.2}$$

where $M_0(M_1, M_2)$, $\sigma_0(\sigma_1, \sigma_2)$ represent the mean and the variance of a Gaussian mass distribution, respectively, and C is defined as a weight of the first component. The best fit results from MLE are shown in Fig. 4, and the corresponding parameters are $(M_1, \sigma_1) = (2.58, 0.01), (M_2, \sigma_2) = (2.72, 0.08)$, and $C = 0.40$.

Similary, for the triple Gaussian distribution model we have

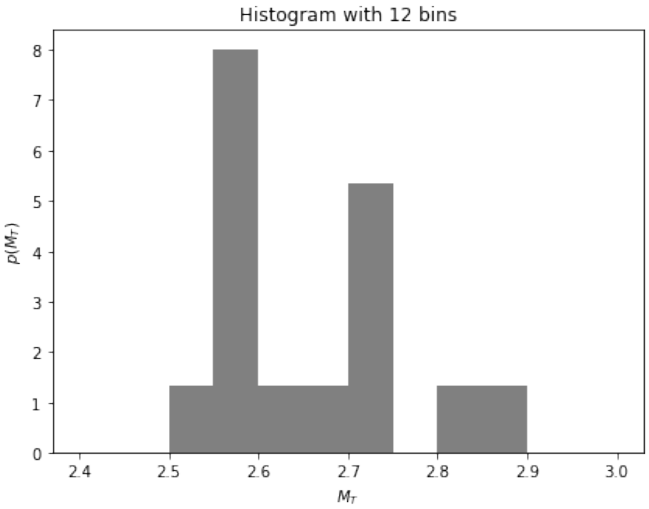
$$\begin{aligned} P(M_T, M_1, M_2, M_3, \sigma_1, \sigma_2, \sigma_3, C_1, C_2) &= \frac{C_1}{\sqrt{2\pi}\sigma_1} exp\left(-\frac{(M_T - M_1)^2}{2\sigma_1^2}\right) \\ &+ \frac{C_2}{\sqrt{2\pi}\sigma_2} exp\left(-\frac{(M_T - M_2)^2}{2\sigma_2^2}\right) \\ &+ \frac{1-C_1-C_2}{\sqrt{2\pi}\sigma_3} exp\left(-\frac{(M_T - M_3)^2}{2\sigma_3^2}\right) \end{aligned} \tag{4.0.3}$$

where $M_0(M_1, M_2, M_3)$, $\sigma_0(\sigma_1, \sigma_2, \sigma_3)$ represent the mean and the variance of a Gaussian mass distribution, respectively, and C_1, C_2 is defined as a weight of the first component. The best fit results from MLE are shown in Fig. 5.

5 Likelihood ratio test

To evaluate the significance of the presence of a structure in the M_T distribution (i.e., it consists of double Gaussian components rather than a single Gaussian component) we adopt the generalized likelihood ratio test that is widely used in hypothesis testing. Our null hypothesis is the single Gaussian distribution, and the alternative hypothesis is a double Gaussian M_T distribution. We construct the likelihood ratio first, as the mass measurement of each neutron star is independent of others. The likelihood function is simply the product of probability for each independent measured total mass, i.e.,

$$L(M_i, \sigma_i | x, M) = \prod_j P(x_j | M_i, \sigma_i; M) \tag{5.0.1}$$



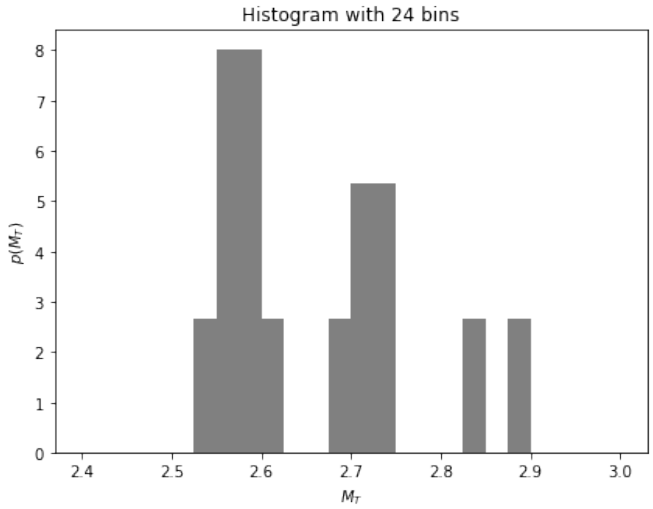


Fig. 2: Histogram of the total masses M_T

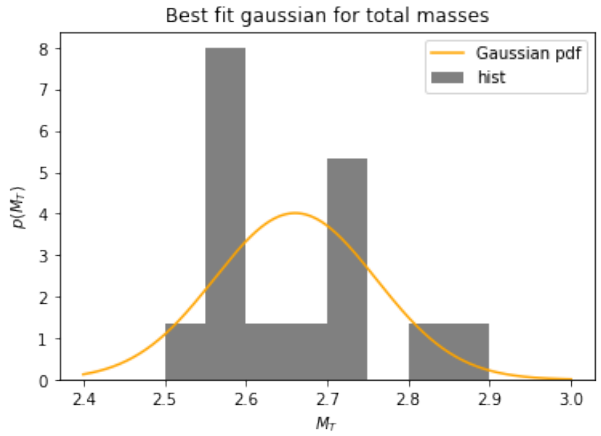


Fig. 3: Best fit Gaussian for M_T

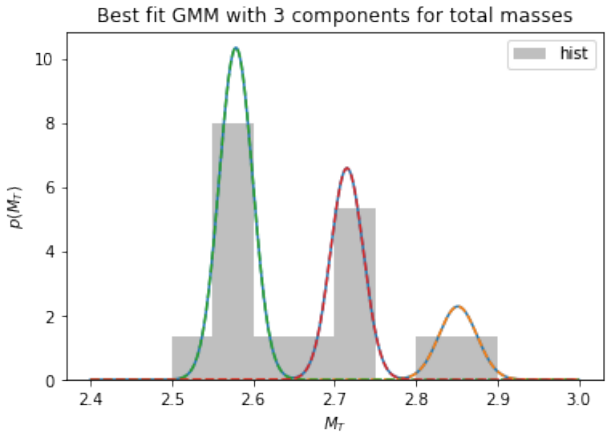


Fig. 5: Best fit GMM, $N_{comp} = 3$ for M_T

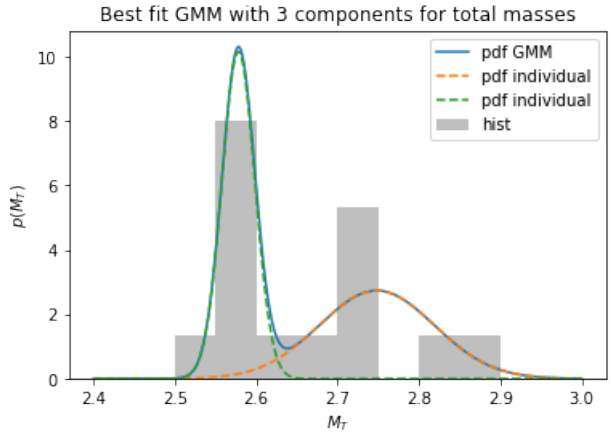


Fig. 4: Best fit GMM, $N_{comp} = 2$ for M_T

where M_i , σ_i are the fitting parameters, M represents the model we used, and x_j represents the data.

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

$$L(x_n|\mu_k, \sigma_k, C_k) = \sum_{k=1}^{N_{comp}} \frac{C_k}{\sqrt{2\pi}\sigma_k} exp\left(-\frac{(x_n - \mu_k)^2}{2\sigma_k^2}\right) \tag{5.0.2}$$

And the likelihood ratio of single component Gaussian mixture to the double component Gaussian mixture is

$$L_1/L_2 = 0.0044 \tag{5.0.3}$$

And the likelihood ratio of double component Gaussian mixture to the triple component Gaussian mixture is

$$L_2/L_3 = 0.03 \tag{5.0.4}$$

6 Information criteria: AICc 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. A more robust way of model selection is provided by information criteria which introduce a penalty term for higher numbers N_{coeffs} of coefficients. The Akaike Information Criterion (AIC) is given in its modified form as

$$AICc = -2\ln L + 2N_{coeffs} + \frac{2N_{coeffs}\left(N_{coeffs+1}\right)}{N_{data} - N_{coeffs} - 1} \tag{6.0.1}$$

Here the second term is the original Akaike penalty for complex models, and the third term is a correction to produce more reliable rankings when N data is small. (The AICc converges to the original AIC for large N data) A popular alternative is the Bayesian Information Citerion (BIC):

$$BIC = -2\ln L + N_{coeffs}\ln N_{data} \tag{6.0.2}$$

The AICc and BIC computed for different Gaussian mixture model are

Likelihood ratio test	AICc	BIC
GMM, $N_{comp} = 1$	-21.71	-21.29
GMM, $N_{comp} = 2$	-20.88	-24.00
GMM, $N_{comp} = 3$	-4.54	-22.88

Despite its name, it is in general not equivalent to a full Bayesian evidence comparison between two models. Lower values of either criterion indicate a preferred model with a better balance between goodness-of-fit and parsimony. The strength of preference is given purely by the differences between models: any overall additive constant can be ignored. There is no universal agreement on how large a difference constitutes clear preference between models, though values between 3 and 5 are usually quoted. Note also that these criteria are formally motivated by asymptotic considerations which cannot be invoked for the small-N data problem under consideration here. Information criteria are generally expected to converge on a consistent answer when the data are indeed informative about the model selection question. Hence, it appears that for the M_T distribution of Galactic DNS systems, the data set is simply not yet large (and/or precise) enough to conclusively answer the question.

7 Cross-validation

Another independent check for overfitting is cross-validation(CV). The basic idea is to check the intra-sample variance of a data set by re-evaluating fits on subsets of the data. For each iteration, a figure of merit (e.g. log-likelihood) is computed on the left-out data points, and in the end averaged over iterations. (In other words, for each iteration, the left-out data are a 'test' set for a model 'trained' on the remaining data.) Overly complex models are expected to get over-fit to the training subsets and then provide inferior prediction performance on the test subsets. The conceptually simplest version is leave-one-out CV, where all possible subsets of $N_{data} - 1$ data points are exhaustively evaluated. Numerical cross-validation scores turn out not to be useful for this small data set, as the variance is too large to make any robust statements. However, an illustrative analysis in the spirit of leave-one-out CV is easily done by fitting GMMs for all 15 subsets of 14 data points each. This also helps identify systems that have a large effect on the fit. The individual fitted distributions for each iteration are compared in Fig. 3. When ignoring measurement errors, individual systems in the M_T . 2.65M range have a large influence on the two-component fits, with the lower-mass peak sometimes even shifting to within the visually apparent 'gap'.

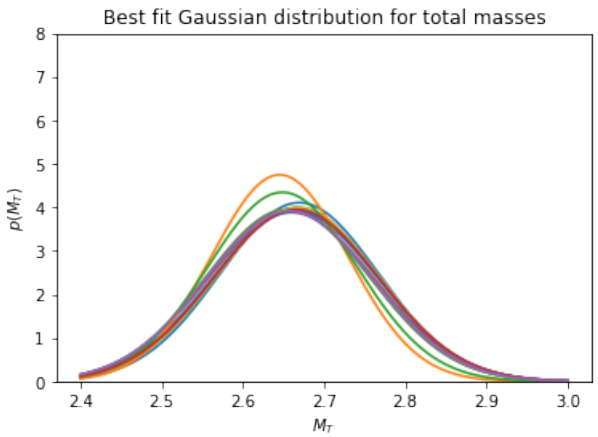


Fig. 6: Cross validation of Gaussian fit M_T

8 Bayesian evidence

Since the DNS data set is so small ($N_{data} = 15$), it is computationally cheap to obtain Bayesian posterior estimates and evidences for model selection.

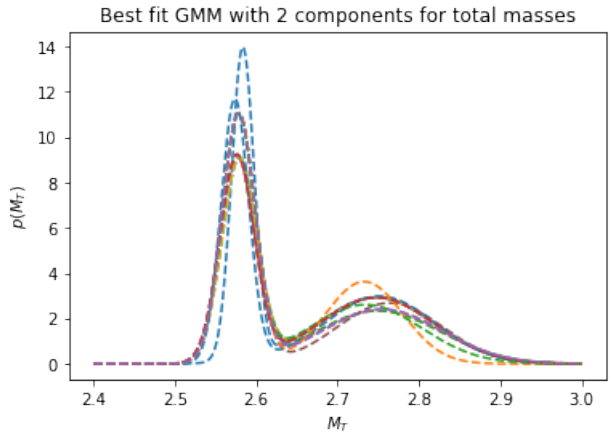


Fig. 7: Cross validation of GMM, $N_{comp} = 2$

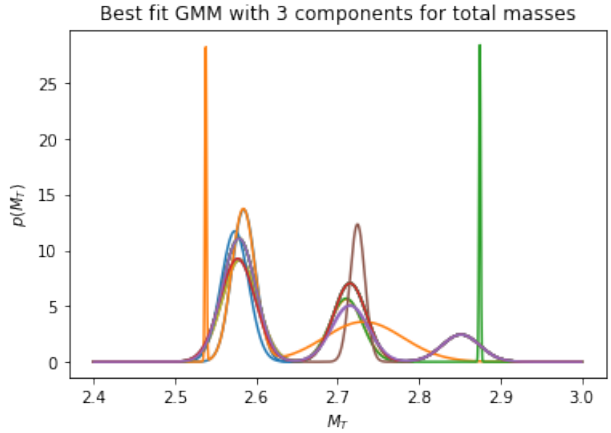


Fig. 8: Cross validation of GMM, $N_{comp} = 3$

Starting from some prior knowledge I, a prior distribution $P(\theta|H,I)$ for the parameters θ of a model $H(\theta)$, and the GMM likelihood $P(x|\theta,H,I) = L(x_n|\mu_k,\sigma_k,C_k)$, the posterior distribution for θ under that model follows from Bayes theorem:

$$P(\theta|x,H,I) = \frac{P(\theta|H,I)P(x|\theta,H,I)}{P(x|H,I)} \qquad (8.0.1)$$

The Bayesian evidence for a model H is defined as its likelihood marginalised over its whole prior support,

$$Z_H = P(x|H,I) = \int d\theta P(\theta|H,I)P(x|\theta,H,I) \quad (8.0.2)$$

Note that this is still dependent on the model H, whereas the total evidence $P(x|I)$ would be a model-independent normalisation factor. Evidence ratios, also called Bayes factors, are a convenient quantity for model selection, as priors need to be defined only over the parameter space of each model, but not between models. To evaluate Z_H for GMMs of different N comp , we can use CPNest, a python implementation of the nested sampling algorithm, with the likelihood function and $N_{live} = 1024$ sampler live points. The outcome of Bayesian inference in general depends on the choice of priors $P_(\theta|H,I)$; the following results are obtained from weakly informative priors. Overall, the CPNest posterior estimates and evidence ratios appear stable under reasonable prior changes. CPNest results are also included in Table.

Model	log Z
GMM, $N_{comp} = 1$	12.62 ± 0.012
GMM, $N_{comp} = 2$	-8.67 ± 0.05

The $N_{comp} = 1, 2$ posteriors are also illustrated in Fig 9. and Fig 10. In addition, Fig shows the median reconstructed GMM distribution functions and their 90% intervals. No CPNest likelihood point estimates are included in Table 1 since these might be misleading without context: Near the posterior median, $log_{10}L$ is generally close to the previous fit results, while higher values can be found in some overall less favoured parts of parameter space. The main quantity of interest for model comparison, the model evidence Z_H , is not derived from a point estimate, but it takes into account the whole sampled volume. At $Z_2/Z_1 \leq 1.4, Z_2Z_3 \approx 1.1$, the evidence ratios are indecisive, meaning that the increased prior volume of GMMs with higher N_{comp} just about makes up for the higher likelihoods achieved, and no clear preference for either model can be found.

9 Conclusions

The distribution of total masses M_T of Galactic DNS systems shows can be fit into unimodal, bimodal or trimodal gaussian distribution. A pure likelihood ratio test prefers those two components over one, with estimating the significance of this preference as 2σ. We have not considered measurement errors into account for any model fit. But considered more robust model selection criteria: Neither the frequentist information criteria (AICc and BIC) which amend the likelihood ratio test with a penalty for the higher number of free parameters in multi-component GMMs, nor a Bayesian evidence ratio test find any robust preference for more than one component. The various GMM fitting methods employed here still all agree with that a

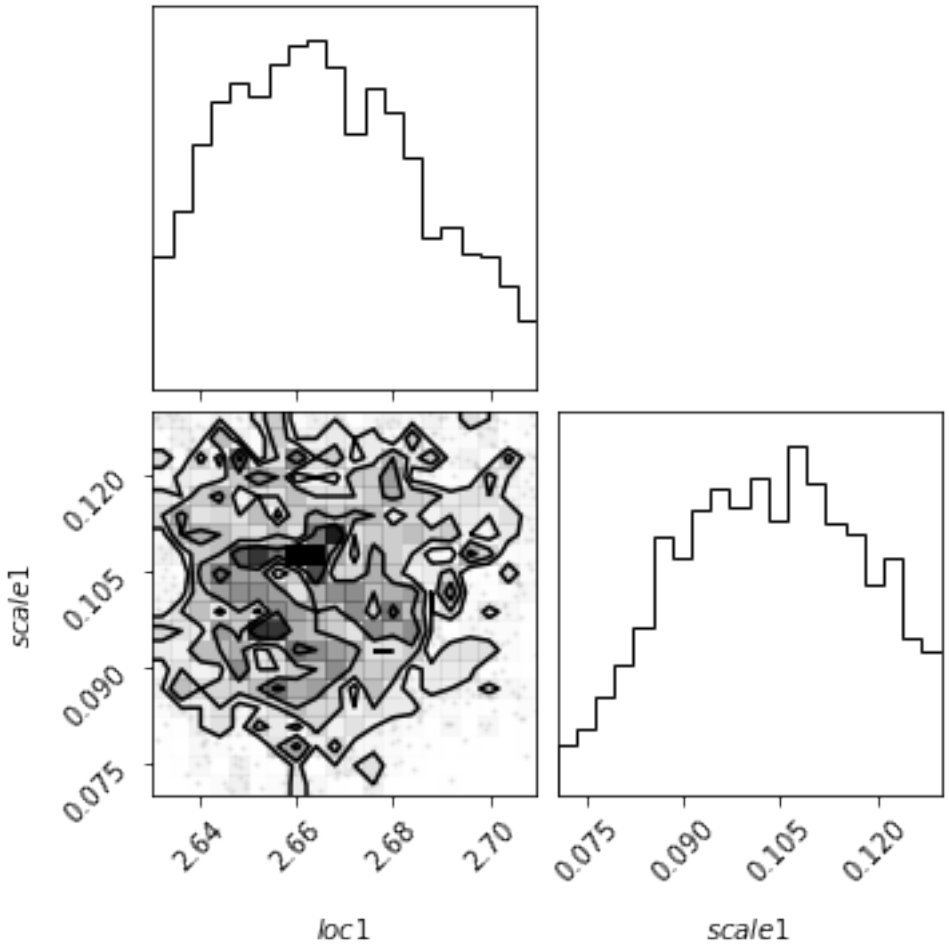


Fig. 9: Corner plots for Gaussian fit

two component GMM certainly provides a good fit to the data; the scenario is not ruled out either, could have interesting consequences for stellar evolution models and GW astronomy. But it appears that the present set of known DNSs is simply too small, and some systems masses are not constrained well enough, to robustly decide between one or two components.

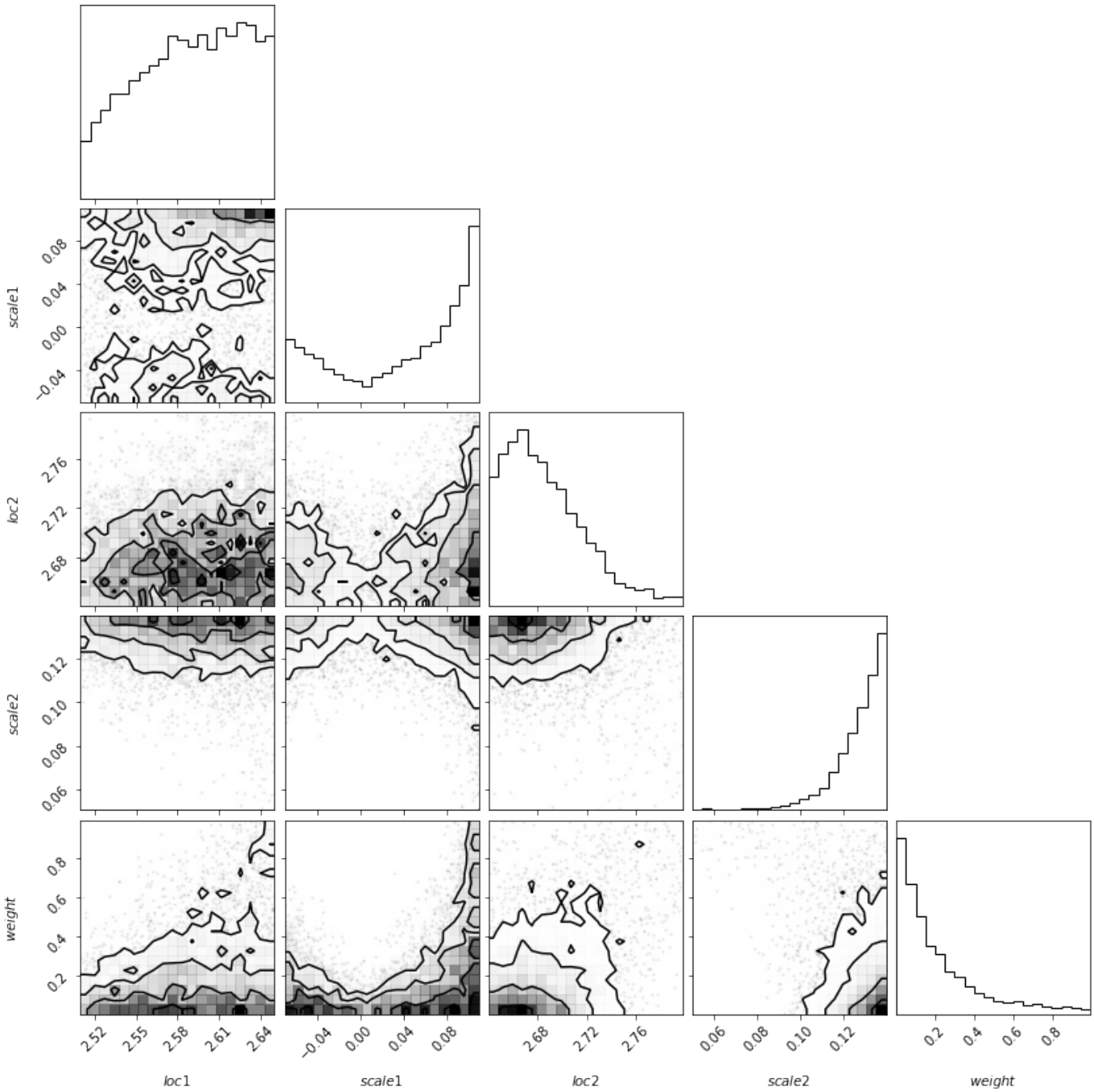


Fig. 10: Corner plots for GMM, $N_{comp} = 2$