Galactic Double Neutron Star total masses and Gaussian mixture model selection

David Keitel

ABSTRACT

Huang et al. (2018) have analysed the population of known galactic Double Neutron Stars (DNSs) regarding the total masses of these systems. They suggest the existence of two sub-populations, and report likelihood-based preference for a two-component Gaussian mixture model over a single Gaussian distribution. This note offers a cautionary perspective on model selection for this data set: A pure likelihood ratio test can encourage overfitting, which can be avoided by penalising models with a higher number of free parameters. Re-examining the DNS total mass data set within the class of Gaussian mixture models, this is achieved through several simple and well-established statistical tests, including information criteria (AICc, BIC), cross-validation and Bayesian evidence ratios. While this re-analysis confirms the basic finding that a two-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. Additional DNS discoveries will be needed to settle the question of sub-populations.

1 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 (Abbott et al. 2017a). Traditionally, a lot of work has focused on using the observed galactic sample to predict coalescence rates (see Abbott et al. 2018b, and references therein), though the distribution of component masses has also been studied (Schwab et al. 2010; Zhang et al. 2011; Özel et al. 2012; Kiziltan et al. 2013).

In a recent paper, Huang et al. (2018) (in the following: Huang+) have considered 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) (Baiotti & Rezzolla 2017; Margalit & Metzger 2017; Ma et al. 2018; Abbott et al. 2017b, 2018a). Huang+ point out an apparent bimodality in the distribution of $M_T$, and with the help of Gaussian mixture models (GMMs) and a likelihood ratio test, they arrived at a 2$\sigma$ preference for two components over one.

In this note, I suggest additional statistical tests not originally considered by Huang+, and caution against relying on likelihood-ratio tests alone, especially when applied to small data sets. Hence, let us re-evaluate the suggested preference for a two-component GMM fit to the observed DNS $M_T$ distribution with a series of simple tests: (i) visual inspection of the data set (Sec. 2.1), (ii) GMM fitting and likelihood-ratio tests for up to three components (Sec. 2.2), (iii) information criteria (AICc and BIC) that penalise underconstrained parameters (Sec. 2.3), (iv) a cross-validation test to understand the impact of individual DNS systems on model selection (Sec. 2.4), and (v) Bayesian evidence computation through nested sampling (Sec. 2.5).

To provide more context for the model selection results, the same criteria are also applied on additional examples: simulated larger $M_T$ data sets (appendix A) and a physically different, but statistically not dissimilar data set of NS spins from Patruno et al. (2017) (appendix C).

2 GMM MODEL SELECTION ON THE DNS MASS DISTRIBUTION

2.1 Data set and visual inspection

This analysis reuses the $M_T$ values, with measurement errors, as collected in Table I of Huang+. (For references to the original measurements, please see that table.) Individual component masses (pulsars and companions) are not considered here; this could be a fruitful topic for further study.
2.2 GMMs and likelihood ratios

For $N_{\text{data}}$ data points $x_n$, the basic likelihood function for a GMM with $N_{\text{comp}}$ means $\mu_k$, widths $\sigma_k$ and component weights $C_k \in [0,1]$ is the product $\prod x_n L(x_n | \{\mu_k, \sigma_k, C_k\})$ of

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

(1)

This can be amended to include measurement errors by assuming each $x_n$ to come from a Gaussian with mean $\mu_n$ and width $\sigma_n$, then marginalising over $x_n$ as nuisance variables:

$$L(\mu_n, \sigma_n | \{\mu_k, \sigma_k, C_k\}) = \int_0^\infty \frac{dx_n}{\sqrt{2\pi} \sigma_n} \exp\left(-\frac{(x_n - \mu_n)^2}{2\sigma_n^2}\right) \sum_{k=1}^{N_{\text{comp}}} \frac{C_k}{\sqrt{2\pi} \sigma_k} \exp\left(-\frac{(x_n - \mu_k)^2}{2\sigma_k^2}\right),$$

and again taking the product over data points $(\mu_n, \sigma_n)$.

To fit GMMs with $N_{\text{comp}} = 1, 2, 3$ components to the $M_T$ data set, we can use two independent python packages:

(i) sklearn.mixture.GaussianMixture (sklearnGMM for short, Pedregosa et al. 2011) supports basic multi-component GMM fitting without measurement errors.

(ii) XDGMM (Xtreme Deconvolution GMM, Hollenen et al. 2017b,a) can also handle known measurement errors; it serves as a wrapper for the astroML (VanderPlas et al. 2014) implementation of a method by Bovy et al. (2011).

Fit results are collected in Table 1 and compared with those from Huang+. The main results of interest are those from XDGMM under consideration of measurement errors (heteroscedastic case). These agree well with Huang+ for one- and two-component GMMs, with the small differences consistent with different fitting implementations. The likelihood ratio of $L_1/L_2 \approx 0.14$ is also similar to their reported 0.011, though going to three components provides another factor of $L_2/L_3 \approx 0.03$ which is less strong, but already illustrates the danger of model selection by likelihood ratios alone.

As a sanity check, XDGMM with errors set to zero produces the same best-fitting GMM parameters and likelihoods as sklearnGMM. Using uniform errors (homoscedastic case) is equivalent to no errors, besides reducing the estimated $\sigma_T$, as expected from Eq. 2. Parameter estimates are consistent in all three cases, though likelihoods are different and the full error treatment is important in assessing statistical robustness, as we will see through the following series of tests.

2.3 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 $N_{\text{coeffs}}$ of coefficients. An astronomy-focused review and pedagogical introduction to such information criteria is provided by Liddle (2007). See also Burnham & Anderson (2002) for a more in-depth exposition.

The Akaikes Information Criterion (AIC), originally introduced by Akaikes (1974), is given in its modified form (the AICc, Hurvich & Tsai 1989) as

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

(3)

Here the second term is the original Akaikes penalty for complex models, and the third term is a correction to produce more reliable rankings when $N_{\text{data}}$ is small. (The AICc converges to the original AIC for large $N_{\text{data}}$.)

A popular alternative is the Bayesian InformationCriterion (BIC) introduced by Schwarz (1978):

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

(4)

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 (Raftery 1995; Burnham & Anderson 2004; Liddle 2007).

Revisiting the heteroscedastic XDGMM fits for the $M_T$ data set using these three criteria, Fig. 2 provides a comparison against the simple log-likelihood, as a function of $N_{\text{comp}}$. The penalty of the AICc is strong for the present case of small $N_{\text{data}}$, so that despite the likelihood ratio it slightly prefers a single component (by $\Delta \text{AICc} = 3$) and very strongly rejects a third component. The BIC gives very small
Table 1. GMM results for the total masses $M_T$ of 15 galactic DNS systems from Huang+ (2018). *sklearnGMM* and *XDGMM* results without measurement errors are identical to the quoted precision. Quantities are defined in sections 2.2–2.5.

| $N_{\text{comp}}$ | C1   | $\mu_1$ | $\sigma_1$ | C2   | $\mu_2$ | $\sigma_2$ | C3   | $\mu_3$ | $\sigma_3$ | log10 $L$ | AICc | BIC | ln $Z$ |
|-------------------|------|----------|------------|------|----------|------------|------|----------|------------|------------|------|-----|--------|
| Huang+            | 1    | 1.00     | 2.67       | 0.10 | 2.07     | 0.50       | 2.52 | 0.08     | 2.72       | 0.08       | 5.77 | 7.77 |
|                   | 2    | 0.40     | 2.58       | 0.01 | 0.60     | 2.72       | 0.08 | 2.72     | 0.08       | 5.80       | -21.71 | -21.29 |
| *sklearnGMM*      | 1    | 1.00     | 2.66       | 0.10 | 2.52     | 0.02       | 2.72 | 0.02     | 2.85       | 0.02       | 8.15 | -20.88 | -24.00 |
| or *XDGMM* (no errors) | 2   | 0.53     | 2.58       | 0.02 | 0.49     | 2.75       | 0.07 | 2.85     | 0.02       | 9.67       | -4.54  | -22.88 |
| (hetero.)         | 3    | 0.52     | 2.58       | 0.02 | 0.33     | 2.72       | 0.02 | 1.3      | 2.85       | 0.02       | 9.07  | -1.79 | -20.13 |

Figure 2. Model selection criteria for GMMs with $N_{\text{comp}} = 1, 2, 3$ applied to the DNS $M_T$ data set. The first three panels are for heteroscedastic *XDGMM* fits. The last panel gives the average Bayesian evidence from 10 CPNest runs at each $N_{\text{comp}}$ (Sec. 2.5). AICc, BIC and ln $Z$ are plotted as differences $\Delta$ to the $N_{\text{comp}} = 1$ value (e.g. $\Delta_{\text{AICc}}(N_{\text{comp}} = 2) = \text{AICc}(N_{\text{comp}} = 2) - \text{AICc}(N_{\text{comp}} = 1)$). For AICc and BIC, *negative* $\Delta$ would mean a preference for that $N_{\text{comp}}$.

2.4 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 $N_{\text{data}} - 1$ 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 $M_T \lesssim 2.65M_\odot$ range have a large influence on the two-component fits, with the lower-mass peak sometimes even shifting to within the visually apparent ‘gap’. By contrast, in heteroscedastic fits, some of those systems are already downweighted by their large uncertainties, and the leave-one-out fits become somewhat more stable. Three-component fits are very unstable in either case.

Hence, this graphical version of a leave-one-out CV test supports a single-Gaussian fit as stable over data subsets, and clearly cautions against three components. Once measurement errors are taken into account, this approach does not uncover any clearly apparent problems with the two-component fit suggested by Huang+, but it is slightly less stable than one component.

2.5 Bayesian evidence from Nested Sampling

Since the DNS data set is so small ($N_{\text{data}} = 15$), it is computationally cheap to obtain Bayesian posterior estimates and evidences for model selection. Starting from some prior knowledge $I$, a prior distribution $P(\theta | \mathcal{H}, I)$ for the parameters $\theta$ of a model $\mathcal{H}(\theta)$, and the GMM likelihood
Figure 3. Illustration of leave-one-out cross-validation on the $M_T$ data set. Top row: sklearnGMM fits ignoring individual measurement errors. Lower row: heteroscedastic XDGMM fits. Each line corresponds to a fit after leaving out one data point.

Figure 4. Posteriors from CPNest runs for a single Gaussian (left panel) and a two-component GMM (remaining panels). Note that $\sigma_1$ appears twice on the x-axis of the last two panels, to better illustrate the posterior structures discussed in detail in appendix B.

\[ P(x \mid \mathcal{H}, I) = \mathcal{L}(x_n \mid \mu_k, \sigma_k, C_k) \] from Eq. 1 or 2, the posterior distribution for $\theta$ follows from Bayes’ theorem:

\[ P(\theta \mid x, \mathcal{H}, I) = \frac{P(\theta \mid \mathcal{H}, I) P(x \mid \theta, \mathcal{H}, I)}{P(x \mid \mathcal{H}, I)} . \]

The Bayesian evidence is defined as the likelihood of a model $\mathcal{H}$ marginalised over its whole prior support,

\[ Z = P(x \mid \mathcal{H}, I) = \int d\theta P(\theta \mid \mathcal{H}, I) P(x \mid \theta, \mathcal{H}, I) . \]

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. See Liddle (2007); Heavens (2009); Jaynes (2003) and references therein for the underlying theory.

To evaluate $Z$ for GMMs of different $N_{\text{comp}}$, we can use CPNest (Veitch et al. 2017), a python implementation of the nested sampling algorithm by Skilling (2004), with the heteroscedastic likelihood function (Eq. 2) and $N_{\text{live}} \approx 1024$ sampler live points. The code also provides an estimate for the uncertainty on the evidence, $\Delta \ln Z = \sqrt{H/N_{\text{live}}}$, with $H$ the information gain from prior to posterior.

Bayesian inference in general depends on the choice of priors $P(\theta \mid \mathcal{H}, I)$; the following results are obtained from weakly informative priors which are discussed in appendix B along with some robustness tests for different choices. Overall, the CPNest posterior estimates appear stable under reasonable prior changes, and evidence ratios remain similar.

CPNest results are also included in Table 1. The listed parameter estimates are posterior medians $\pm$ 10% and 90% quantiles. While these consistently include the previous estimates, it is interesting to note that for $N_{\text{comp}} = 2$ the posterior uncertainties on $\mu_k$ and $\sigma_k$ are also almost compatible with a vanishing separation between the two components, and those on the $C_k$ and component weights $C_k$ are rather large. The $N_{\text{comp}} = 3$ case is not well constrained and hence posterior estimates are very broad, with strongly overlapping components. The $N_{\text{comp}} = 1, 2$ posteriors are also illustrated in Fig. 4 and discussed in detail in appendix B. In addition, Fig. 5 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, $\log_{10} \mathcal{L}$ is generally close to the previous fit results, while higher values can be found in some overall less favoured parts of parameter space. (See appendix B.) The main quantity of interest for model comparison, the ev-
Z

idence $Z_1$ is not derived from a point estimate, but as seen in Eq. 6 it takes into account the whole sample volume. At $Z_2/Z_1 \lesssim 1.4$, $Z_2/Z_1 \approx 1.1$, the evidence ratios are indicative, meaning that the increased prior volume of GMMs with higher $N_{\text{comp}}$ just about makes up for the higher likelihoods achieved, and no clear preference for either model can be found. Priors would need to be severely tightened to bring $Z_2/Z_1$ even above 2.

### 3 CONCLUSIONS

The distribution of total masses $M_T$ of Galactic DNS systems shows some apparent bimodality, which can be fit with a two-component GMM as shown by Huang+. A pure likelihood ratio test prefers those two components over one, with Huang+ estimating the significance of this preference as 2σ. As a first step towards testing if this indeed points to two distinct underlying populations of astrophysical objects, while it is my understanding that Huang+ are also working on a more sophisticated analysis, in this note I have kept their initial GMM assumption, but considered more robust model selection criteria: Neither the frequentist information criteria (AICc and BIC) considered in Sec. 2.3, 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 (Sec. 2.5) find any robust preference for more than one component. The various GMM fitting methods employed here still all agree with Huang+ that a two-component GMM certainly provides ‘a good fit’ to the data; the scenario is not ruled out either and, as pointed out by Huang+, 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.

It will be interesting to revisit this model selection problem once additional DNS systems are observed, as expected in great numbers from upcoming surveys e.g. with MeerKAT (Bailes et al. 2018) and the SKA (Smits et al. 2009); or to combine the Galactic sample with GW observations of extragalactic mergers, as suggested by Huang+ in the second half of their paper. (Though Pankow (2018) suggests that GW170817 (Abbott et al. 2017a) might not be consistent with the same population as the galactic DNSs.)

In the meantime, the simple reanalysis in this note has certainly not exhausted the full potential of the present data set. One could also consider distribution functions beyond the GMM family. (Huang+ already suggested a GMM plus uniform distribution.) And since the cross-validation analysis suggests that, for the current small data set size, a few systems can have a large effect on any inference of the underlying distribution, revisiting individual systems’ mass measurements – or even their identity as DNSs – could also improve the situation. For example, J1811–1736 has the widest uncertainty in the Huang+ data set ($M_T = 2.57 \pm 0.10$; referring back to the original studies of Lyne et al. (2000) and Corongiu et al. (2006), its rather low companion mass means that while it is generally accepted as a DNS, this identification might not be completely iron-clad. A combined reanalysis of total and component masses could also be promising in constrainning the model selection problem, and more sophisticated statistical techniques could be applied to deal with possible selection effects.

Data sets used in this note (reproduced from Huang et al. (2018) and Patruno et al. (2017)) and CPNest posterior samples are provided as ancillary files of the arXiv preprint.

### ACKNOWLEDGEMENTS

The author was funded under the EU Horizon2020 framework, Marie Skłodowska-Curie grant agreement 704094 GRANITE. Thanks to Zhu Xingjiang and Paul Lasky (P.L.) for inspirational discussions on the $M_T$ topic; to both of them, to John Veitch (J.V.) and Graham Woan for comments on the manuscript; to J.V. for an introduction to CPNest; and to P.L. for suggesting Fig. 5.

### REFERENCES

Abbott B. P., et al., 2017a, Phys. Rev. Lett., 119, 161101
Abbott B. P., et al., 2017b, Apj, 851, L16
Abbott B. P., et al., 2018a, preprint (arXiv:1805.11581)
Abbott B. P., et al., 2018b, Living Rev. Rel., 21:3
Akaike H., 1974, IEEE Trans. Autom. Control, 19, 716
Bailes M., et al., 2018, in PoS MeerKAT2016, Stellenbosch, South Africa, p. 011 (arXiv:1803.07424)
Baiotti L., Rezzolla L., 2017, Rept. Prog. Phys., 80, 096901
Bovy J., Hogg D. W., Roweis S. T., 2011, Ann. Appl. Stat., 5, 1657
Burnham K. P., Anderson D. R., 2002, Model Selection and Multimodel Inference: A practical information-theoretic approach. Springer-Verlag
Burnham K. P., Anderson D. R., 2004, Soc. Meth. & Res., 33, 261
Corongiu A., Kramer M., Stappers B. W., Lyne A. G., Jessner A., Possenti A., D’Amico N., Loehmer O., 2006, A&A wrapper class, https://github.com/tholoien/xdgmm
Holoien T. W.-S., Marshall P. J., Wechsler R. H., 2017a, XDGMM
Holoien T. W.-S., Marshall P. J., Wechsler R. H., 2017b, AJ, 153, 249
Huang Y.-J., Jiang J.-L., Li X., Jin Z.-P., Fan Y.-Z., Wei D.-M., 2018, preprint (arXiv:1804.03101)
Hurvich C. M., Tsai C.-L., 1989, Biometrika, 76, 297

Figure 5. GMM distribution functions $p(M_T)$ inferred with CPNest for one (blue solid line) or two components (orange dashed line), superimposed on a data histogram. The solid foreground lines correspond to the median posterior parameters and the background ‘haze’ is a superposition of $p(M_T)$ evaluated at parameters within the 90% confidence region. See appendix B for details.
APPENDIX A: GMMS ON SIMULATED DNS POPULATIONS

To further illustrate the scaling and robustness of various model selection criteria, let us consider some simulated populations where the ‘true’ distribution is known. Fig. A1 shows results from simulated single Gaussians (left column) or two-component mixtures (right column) with parameters matching those reported by Huang+, as a function of the number of randomly drawn samples matching those reported by Huang+. 50 independent populations are drawn for each N_{data} step. Each difference ΔQ is taken as Q_{N_{comp}=2} − Q_{N_{comp}=1}. Hence, positive Δlog_{10}L imply a better fit of the two-component model, while for AIC and AICc (shown together in the second row) as well as BIC a negative value indicates preference for the two-component GMM. Error bars correspond to one standard deviation.

Figure A1. Model selection criteria for GMMS with N_{comp} = 1, 2 applied to simulated M_T populations drawn from one- and two-component GMMs with the fiducial values reported by Huang+. 50 independent populations are drawn for each N_{data}. For a true single-component Gaussian, the likelihood ratio test tends to stay inconclusive: a two-component GMM can always fit the data slightly, but not much, better. At N_{data} ≳ 20, and continue to the case any further due to the fixed complexity penalty of AIC and AICc (shown together in the second row) as well as BIC a negative value indicates preference for the two-component GMM. Error bars correspond to one standard deviation.

APPENDIX B: NESTED SAMPLING: PRIOR CHOICE AND ADDITIONAL DETAILS

This section gives some additional details on the CPNest runs and interpretation of the resulting posteriors and evidences.

Prior choice: The CPNest results in Sec. 2.5 use uniform priors in [0, 1] for the GMM weights C_k, truncated log-uniform priors in [0.001, 1] for the σ_k and Gaussian priors for the μ_k with widths 0.3 and means spaced uniformly in the range of M_T. Constraints ensure ∑_k C_k = 1 and μ_k+1 ≥ μ_k ∀k.

In Bayesian inference, it is generally wise to test the effect of different prior choices. For example, keeping the same priors on σ_k and C_k but changing the μ_k priors to uniform within [1.8,4.0]M_{c,⊙} yields almost unchanged posterior estimates for N_{comp} = 1 and slightly broader, but consistent estimates for N_{comp} = 2. Alternatively, keeping the μ_k and C_k priors but narrowing the log-uniform range for the σ_k to [0.005,0.2] cuts off the third minor peak in the posterior estimates, but only marginally influences the overall estimates. The evidence values change somewhat with the prior volume, but the ratios remain indicative: Z_2/Z_1 ≈ exp(8.63 − 8.32) ≈ 1.4 for the first set of priors, Z_2/Z_1 ≈ exp(6.28 − 7.04) ≈ 0.5 for the uniform μ_k priors and Z_2/Z_1 ≈ exp(9.55 − 8.94) ≈ 1.8 for the narrower σ_k priors.
**Posterior estimates:** Let us consider the obtained posterior distributions a bit more closely, especially the $N_{\text{comp}} = 1, 2$ cases illustrated in Fig. 4. The posterior parameter estimates given in Table 1 are medians ± 90% quantiles. However, the $N_{\text{comp}} = 2$ posteriors have some asymmetric, and for $\sigma_2$ even multimodal, structure. In general there are various arguments for or against quoting posterior medians vs. means, see e.g. Jaynes (2003), but for such cases medians tend to be more robust. In any case, we can understand these features as not just due to technical issues, e.g. insufficient convergence of the sampler but to features of the underlying $M_T$ data set and the fact that a single-component Gaussian is effectively included in the parameter space of a two-component GMM. In the $N_{\text{comp}} = 2$ posteriors, the smaller secondary peak at high $\sigma_1$ and low $\sigma_2$ corresponds to very low $C_2$: in this part of parameter space, the posterior $p(M_T)$ GMM function is effectively almost a single, broad Gaussian with only a small localised bump at higher masses. Conversely, there is a third, even smaller peak in the posterior for low $\sigma_1$, high $\sigma_2$ and high $C_2$, where the result is a dominant broad Gaussian with a small localised bump near the observed low-$M_T$ excess. These subdominant solutions with localised bumps can also be seen in some of the distribution functions from within the 90% quantiles plotted as background lines in Fig. 5. These observations are consistent with the indecisive evidence ratio between $N_{\text{comp}} = 1$ and 2, but they do also tell us to stay cautious of overfitting even when additional data points will become available.

Along similar lines, for $N_{\text{comp}} = 3$ the degeneracies become worse and the posteriors more complicated: results in Table 1 are just quoted for completeness and it is not particularly edifying to analyse the posteriors in detail.

**Likelihoods:** The nontrivial posterior structure for $N_{\text{comp}} \geq 2$ leads to an ambiguity when trying to quote ‘the likelihood’ from a nested-sampling run, because in general neither the maximum-likelihood (ML) point nor the mode of the posterior (MAP) need to be particularly close to where the main mass of posterior probability is concentrated in parameter space. For $N_{\text{comp}} = 1$, both the ML and MAP agree with the XDGMM results, and with the likelihoods evaluated at the median or mean posterior parameter estimates, to within $\log_{10} L \approx 0.05$. For $N_{\text{comp}} \geq 2$ the likelihoods at median or mean still are similar to the XDGMM results, while the ML and MAP values can be significantly higher (up to a factor of 10); but these tend to come from extreme-$C_k$ parts of the parameter space corresponding to the ‘dominant broad component plus small bump’ over-fitting cases discussed above.

**APPENDIX C: COMPARISON EXAMPLE:**

**LMXB SPIN FREQUENCIES**

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 $f_{\text{spin}}$ for a population of 29 neutron stars in Low Mass X-ray Binary (LMXB) systems. The data set is given in Table 2 of Patruno et al. (2017). Those authors also fit one- and two-component GMMs (with an R implementation of the expectation-maximization algorithm) and found a BIC difference of $\gtrsim 7$ in preference of two components.

Ignoring measurement errors in this case (which for NS spin frequencies should be much smaller than for masses), the sklearnGMM and XDGMM implementations return consistent parameters as reported in Table C1. The two-component results are fully consistent with those reported by Patruno et al. (2017) within their confidence intervals.

Looking at the various statistical criteria as a function of number of components $N_{\text{comp}} = 1, 2, 3$ (also illustrated in Fig. C1), these are also fully consistent with the findings of Patruno et al. (2017): Again the likelihood ratio alone is already in favour of two components ($L_1/L_3 \approx 10^{-4}$), but only additional criteria can yield a robust decision, and on its own, the additional gain of $L_2/L_3 \approx 0.07$ could tempt us to use an even more complex model. Fortunately, in this case AICc and BIC agree in clearly preferring two components over one (with $\Delta \gtrsim 7$–11), and also prefer two over three by $\Delta$ of almost 5. A leave-one-out CV test (see Fig. C2) also shows the two-component fit to be almost as stable as a single component in this case.

CPNest results for this data set (with uniform priors in $C_k$, log-uniform in $\sigma_2$ within [10, 350] Hz and Gaussian $\mu_k$ priors of width 70 Hz) are very consistent with the other GMM fits, except for some differences in the more degenerate $N_{\text{comp}} = 3$ case. The evidence ratio is in favour by $\gtrsim 16$ of two components over one, with three components preferred by a marginal factor of only 3 and much less robust results.

Overall, the LMXB data set seems to be a good example for information criteria and Bayesian evidence consistently backing up the selection of a more complex model: in contrast to the $M_T$ of DNSs data set, the preference for two components over a single Gaussian distribution seems statistically robust. Three components cannot be quite as confidently excluded, but two yield the most robust fit.

---

Figure C1. Model selection criteria for GMMs with $N_{\text{comp}} = 1, 2, 3$ applied to the Patruno+ data set of $f_{\text{spin}}$ for NSs in LMXBs. The first three panels are for sklearnGMM fits. The last panel gives the average Bayesian evidence from 10 CPNest runs at each $N_{\text{comp}}$. AICc, BIC and $\ln L$ are plotted as differences to the $N_{\text{comp}} = 1$ value.
Table C1. GMM results for the $f_{\text{spin}}$ data set for NSs in LMXBs from Patruno et al. (2017), assuming negligible measurement errors.

| $N_{\text{comp}}$ | sklearnGMM | or XDGMN | CPNest | sklearnGMM | or XDGMN | CPNest | sklearnGMM | or XDGMN | CPNest |
|-------------------|------------|-----------|--------|------------|-----------|--------|------------|-----------|--------|
|                   | $C_1$ | $\sigma_1$ | $C_2$ | $\mu_2$ | $\sigma_2$ | $C_3$ | $\mu_3$ | $\sigma_3$ | $\log_{10} L$ | AICc | BIC | $\ln Z$ |
| 1                 | 1.00 | 414 | 153 | -81.20 | 378.42 | 380.69 | -190.26 | ±0.05 |
| 2                 | 0.60 | 308 | 99  | 0.40 | 576 | 29 | 377.23 | ±0.08 |
| 3                 | 0.22 | 201 | 30  | 0.36 | 565 | 55 | 378.07 | ±0.08 |

Figure C2. Illustration of leave-one-out cross-validation on the Patruno+ $f_{\text{spin}}$ data set. Each dotted line corresponds to a fit after leaving out one data point. In this case, the fitted distributions are relatively stable not just for $N_{\text{comp}} = 1$, but also for the preferred fit with $N_{\text{comp}} = 2$, and only for even more components they start becoming unstable.

Figure C3. Posteriors from CPNest runs on the LMXB data set for a single Gaussian (left panel) and a two-component GMM (remaining panels). Note that $\sigma_1$ appears twice on the x-axis of the last two panels.

Figure C4. Median GMM distribution functions $p(f_{\text{spin}})$ inferred with CPNest for one (blue solid line) or two components (orange dashed line), superimposed on a data histogram, and with 90% quantiles background ‘haze’.

This paper has been typeset from a TeX/LaTeX file prepared by the author.