MNRAS **000**, 1–10 (2019)

Galactic Double Neutron Star total masses and Gaussian mixture model selection

David Keitel[1,2]★
[1] *University of Glasgow, School of Physics and Astronomy, Kelvin Building, Glasgow G12 8QQ, Scotland, United Kingdom*
[2] *University of Portsmouth, Institute of Cosmology and Gravitation, Portsmouth PO1 3FX, United Kingdom*

LIGO-P1800175 [draft version: 1 February 2019]

ABSTRACT

Huang et al. (2018) have analysed the population of 15 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: Especially for such a small sample size, a pure likelihood ratio test can encourage overfitting. This 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 can be achieved through several simple and well-established statistical tests, including information criteria (AICc, BIC), cross-validation, Bayesian evidence ratios and a penalised EM-test. 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.

Key words: stars: neutron – binaries – methods: statistical – pulsars

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. 2018c 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,b). 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 2σ 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. First, for completeness, (i) visual inspection of the data set (Sec. 2.1) and (ii) GMM fitting and likelihood-ratio tests (Sec. 2.2) are briefly summarised. The additional hypothesis test methods include (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), (v) Bayesian evidence computation through nested sampling (Sec. 2.5), and (vi) a penalised EM-test (Sec. 2.6).

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

★ E-mail: david.keitel@ligo.org

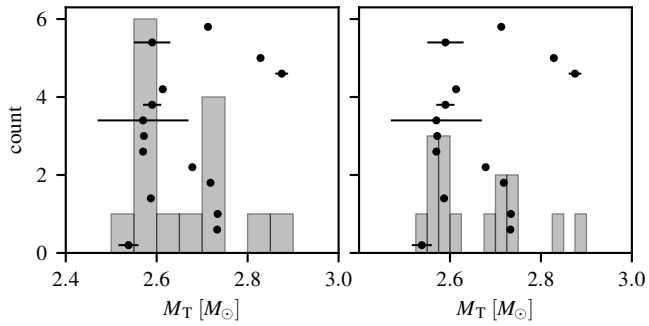

Figure 1. The data set of total masses M_T from Table I of Huang+ (available online as dns_1804.03101v1.txt). The left panel shows a histogram with the same number of bins ($N_{bins} = 12$) between 2.4 and 3.0 $M_\odot$ as in Fig. 1 of Huang+. The right panel uses twice as many bins. The M_T values with error bars for individual DNSs are overlaid as scatter plots; the vertical placement is just for ease of viewing and has no numerical meaning.

component masses (pulsars and companions) are not considered here; this could be a fruitful topic for further study.

Histograms of the M_T data set are shown in Fig. 1. It compares the original binning from Huang+ with the alternative choice of twice as many 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. The overlaid scatter plots illustrate the large range in error bar magnitudes on the M_T measurements, and a concentration of more uncertain measurements near the apparent 'lower peak'.

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

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

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 marginalising over x_n as nuisance variables:

$$\begin{aligned}
&\mathcal{L}(\mu_n, \sigma_n \,|\, \{\mu_k, \sigma_k, C_k\}) \\
&= \int_{-\infty}^{\infty} \frac{\mathrm{d}x_n}{\sqrt{2\pi}\,\sigma_n} \exp\left(-\frac{(x_n - \mu_n)^2}{2\sigma_n^2}\right) \sum_{k=1}^{N_{comp}} \frac{C_k}{\sqrt{2\pi}\,\sigma_k} \exp\left(-\frac{(x_n - \mu_k)^2}{2\sigma_k^2}\right) \\
&= \sum_{k=1}^{N_{comp}} \frac{C_k}{\sqrt{2\pi}\,\sqrt{\sigma_k^2 + \sigma_n^2}} \exp\left(-\frac{(\mu_n - \mu_k)^2}{2(\sigma_k^2 + \sigma_n^2)}\right), \quad (2)
\end{aligned}$$

and again taking the product over data points $\{\mu_n, \sigma_n\}$.

To fit GMMs with $N_{comp} = 1, 2, 3$ 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, Holoien 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 $\mathcal{L}_1/\mathcal{L}_2 \approx 0.014$ is also similar to their reported 0.011, though going to three components provides another factor of $\mathcal{L}_2/\mathcal{L}_3 \approx 0.03$ which is less strong, but already illustrates the danger of model selection by likelihood ratios alone: Adding additional components to the mixture model will generally increase the likelihood until each data point is fit by its own model component. Huang+ estimate significance assuming a χ^2 distribution for the likelihood ratio, which is itself problematic for a GMM on a small data set (see e.g. Ciuperca et al. 2003; Chen et al. 2008; Chen & Li 2009); instead the following sections will describe more robust hypothesis tests.

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 σ_k, 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_{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 Akaike Information Criterion (AIC), originally introduced by Akaike (1974), is given in its modified form (the AICc, Hurvich & Tsai 1989) as

$$\mathrm{AICc} = -2 \ln \mathcal{L} + 2\, N_{coeffs} + \frac{2\, N_{coeffs}\,(N_{coeffs} + 1)}{N_{data} - N_{coeffs} - 1}. \quad (3)$$

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 Criterion (BIC) introduced by Schwarz (1978):

$$\mathrm{BIC} = -2 \ln \mathcal{L} + N_{coeffs} \ln N_{data}. \quad (4)$$

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,

Table 1. GMM results for the total masses M_T of 15 galactic DNS systems from Huang et al. (2018). `sklearnGMM` and `XDGMM` results without measurement errors are identical to the quoted precision. Quantities are defined in sections 2.2–2.6.

	N_{comp}	C_1	μ_1	σ_1	C_2	μ_2	σ_2	C_3	μ_3	σ_3	$\log_{10}\mathcal{L}$	AICc	BIC	$\ln\mathcal{Z}$
Huang+	1	1.00	2.67	0.10							5.77			
	2	0.40	2.58	0.01	0.60	2.72	0.08				7.77			
`sklearnGMM`	1	1.00	2.66	0.10							5.80	-21.71	-21.29	
or `XDGMM`	2	0.51	2.58	0.02	0.49	2.75	0.07				8.15	-20.88	-24.00	
(no errors)	3	0.53	2.58	0.02	0.33	2.72	0.02	0.13	2.85	0.02	9.67	-4.54	-22.88	
`XDGMM`	1	1.00	2.66	0.10							5.74	-21.45	-21.04	
(heterosc.)	2	0.48	2.58	0.02	0.52	2.74	0.07				7.60	-18.35	-21.48	
	3	0.52	2.58	0.02	0.35	2.71	0.02	0.13	2.85	0.02	9.07	-1.79	-20.13	
`CPNest`	1	$1.00^{+0.00}_{-0.00}$	$2.67^{+0.04}_{-0.04}$	$0.10^{+0.03}_{-0.02}$										8.32±0.06
	2	$0.42^{+0.42}_{-0.27}$	$2.58^{+0.07}_{-0.01}$	$0.02^{+0.09}_{-0.02}$	$0.58^{+0.27}_{-0.42}$	$2.72^{+0.08}_{-0.06}$	$0.09^{+0.05}_{-0.04}$							8.63±0.08
	3	$0.38^{+0.49}_{-0.38}$	$2.57^{+0.03}_{-0.18}$	$0.02^{+0.12}_{-0.02}$	$0.35^{+0.32}_{-0.27}$	$2.68^{+0.06}_{-0.10}$	$0.05^{+0.09}_{-0.05}$	$0.26^{+0.37}_{-0.22}$	$2.76^{+0.12}_{-0.07}$	$0.08^{+0.10}_{-0.07}$				8.51±0.08
`MixtureInf`	1	1.00	2.66	0.10										
(no errors)	2	0.51	2.58	0.03	0.49	2.74	0.08							

though values between 3 and 5 are usually quoted (Raftery 1995; Burnham & Anderson 2004; Liddle 2007). Note also that these criteria are formally motivated by asymptotic considerations (see e.g. Burnham & Anderson 2002, 2004 and references therein) which cannot be invoked for the small-N_{data} problem under consideration here. Hence, for now let us consider them as heuristic criteria, and investigate how they compare with other tests. (See also appendix A for simulations with larger N_{data}.)

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_{comp}. The penalty of the AICc is strong for the present case of small N_{data}, so that despite the likelihood ratio it slightly prefers a single component (by ΔAICc ≈ 3) and very strongly rejects a third component. The BIC gives very small differences, telling us that the data are indecisive. From Table 1, note also that the no-errors fits give a lower BIC for $N_{comp} = 2$, and hence indeed the full error treatment is important in obtaining a robust model selection – the difference is easily understood by the clustering of wide-uncertainty measurements near lower M_T.

Overall, these criteria (unsurprisingly) agree rather clearly that there is no justification for adding a third GMM component. However for the main question of Huang+, whether there are two components or only one, the situation is still indecisive. As we will see from the alternative examples in the appendix, 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.

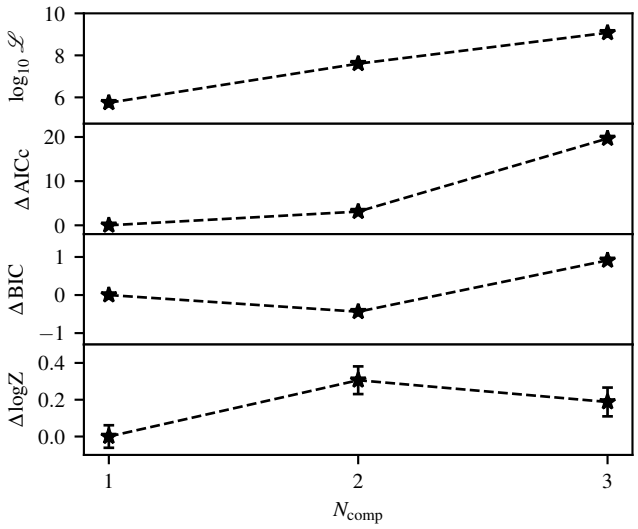

Figure 2. Model selection criteria for GMMs with $N_{comp} = 1, 2, 3$ applied to the DNS M_T data set. The first three panels are for heteroscedastic `XDGMM` fits. The last panel gives the average Bayesian evidence from 10 `CPNest` runs at each N_{comp} (Sec. 2.5). AICc, BIC and $\ln\mathcal{Z}$ are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

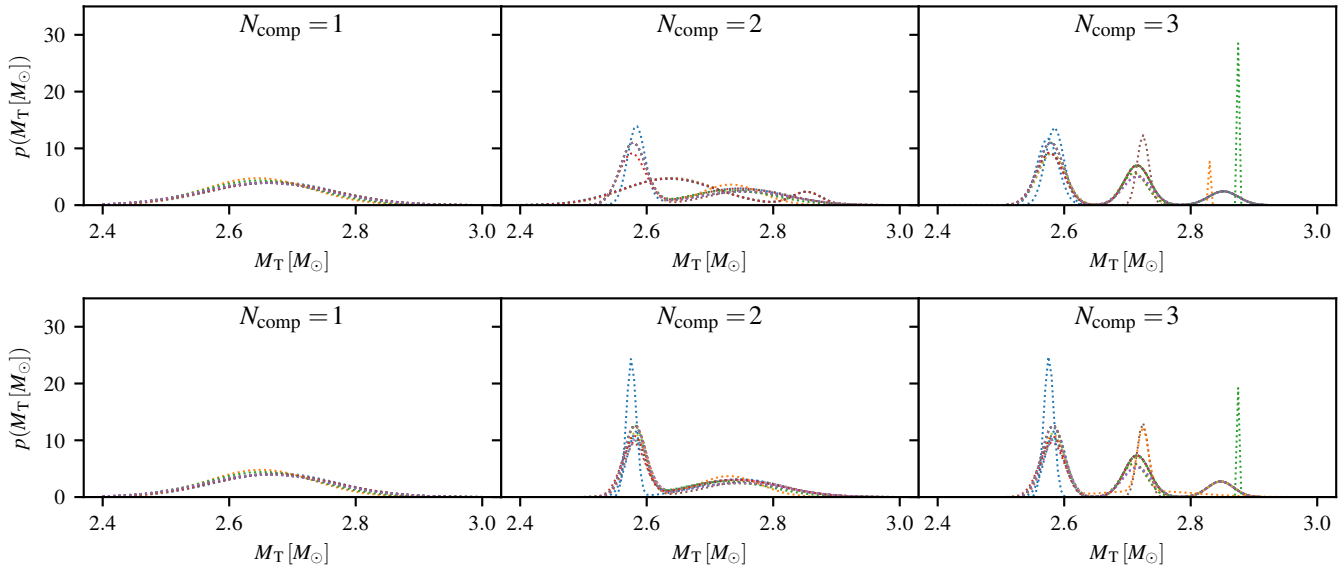

Figure 3. Illustration of leave-one-out cross-validation on the M_{T} data set. Top row: `sklearnGMM` fits ignoring individual measurement errors. Lower row: heteroscedastic `XDGMM` fits. Each line corresponds to a fit after leaving out one data point.

dividual systems in the $M_{\mathrm{T}} \lesssim 2.65 M_\odot$ 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 ($N_{\mathrm{data}} = 15$), it is computationally cheap to obtain Bayesian posterior estimates and evidences for model selection. Starting from some prior knowledge $\mathcal{I}$, a prior distribution $P(\theta \mid \mathcal{H}, \mathcal{I})$ for the parameters θ of a model $\mathcal{H}(\theta)$, and the GMM likelihood $P(x \mid \theta, \mathcal{H}, \mathcal{I}) = \mathcal{L}(x_n \mid \{\mu_k, \sigma_k, C_k\})$ from Eq. 1 or 2, the posterior distribution for θ under that model follows from Bayes' theorem:

$$P(\theta \mid x, \mathcal{H}, \mathcal{I}) = \frac{P(\theta \mid \mathcal{H}, \mathcal{I})\, P(x \mid \theta, \mathcal{H}, \mathcal{I})}{P(x \mid \mathcal{H}, \mathcal{I})} \,. \tag{5}$$

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

$$\mathcal{Z}_{\mathcal{H}} = P(x \mid \mathcal{H}, \mathcal{I}) = \int \mathrm{d}\theta\, P(\theta \mid \mathcal{H}, \mathcal{I})\; P(x \mid \theta, \mathcal{H}, \mathcal{I}) \,. \tag{6}$$

Note that this is still dependent on the model $\mathcal{H}$, whereas the total evidence $P(x \mid \mathcal{I})$ 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. See Gregory (2005); Liddle (2007); Heavens (2009); Jaynes (2003) and references therein for the underlying theory.

To evaluate $\mathcal{Z}_{\mathcal{H}}$ for GMMs of different N_{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 $N_{\mathrm{live}} = 1024$ sampler live points. The code also provides an estimate for the uncertainty on the evidence, $\Delta \ln \mathcal{Z}_{\mathcal{H}} \approx \sqrt{H_{\mathcal{H}}/N_{\mathrm{live}}}$, with the information gain $H_{\mathcal{H}}$ from prior to posterior.

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

`CPNest` results are also included in Table 1. The listed parameter estimates are posterior medians ± 10% and 90% quantiles. While these consistently include the previous estimates, it is interesting to note that for $N_{\mathrm{comp}} = 2$ the posterior uncertainties on μ_k and σ_k are also almost compatible with a vanishing separation between the two components, and those on the σ_k and component weights C_k are rather large. The $N_{\mathrm{comp}} = 3$ case is not well constrained and hence posterior estimates are very broad, with strongly overlapping components. The $N_{\mathrm{comp}} = 1, 2$ 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 model evidence $\mathcal{Z}_{\mathcal{H}}$, is not derived from a point estimate, but as seen in Eq. 6 it takes into account the whole sampled

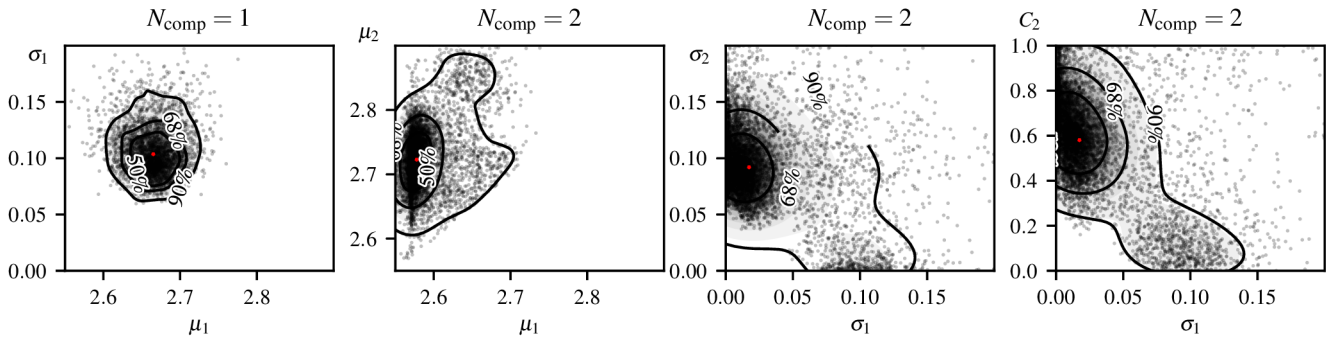

Figure 4. Posteriors from `CPNest` runs for a single Gaussian (left panel) and a two-component GMM (remaining panels). Note that σ_1 appears twice on the x-axis of the last two panels, to better illustrate the posterior structures discussed in detail in appendix B.

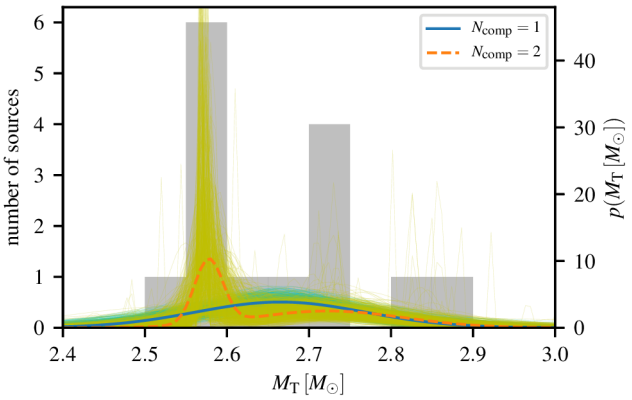

Figure 5. GMM distribution functions $p(M_{\rm T})$ inferred with `CPNest` for one (blue solid line) or two components (orange dashed line), superimposed on a data histogram. The solid foreground lines correspond to the median posterior parameters and the background 'haze' is a superposition of $p(M_{\rm T})$ evaluated at parameters within the 90% confidence region. See appendix B for details.

volume. At $\mathcal{Z}_2/\mathcal{Z}_1 \lesssim 1.4$, $\mathcal{Z}_2/\mathcal{Z}_3 \approx 1.1$, the evidence ratios are indecisive, meaning that the increased prior volume of GMMs with higher $N_{\rm comp}$ just about makes up for the higher likelihoods achieved, and no clear preference for either model can be found. The alternative prior choices considered in appendix B do not change $\mathcal{Z}_2/\mathcal{Z}_1$ far away from unity, indicating that much tighter priors would be needed to obtain clear preference for a multi-component model, which would then however be driven by that prior choice and not by the data.

2.6 EM test

For finite mixture models, the EM-test (Chen & Li 2009; Chen et al. 2012) is based on a penalised maximum-likelihood estimator (MLE) and the expectation-maximisation (EM) algorithm (Dempster et al. 1977). An `R` implementation is available in the package `MixtureInf` (Li et al. 2016). In alternating iteration steps, EM assigns data samples to the proposed mixture components according to their relative probability and then the MLE is updated. (See Do & Batzoglou 2008 for a didactic introduction.)

`MixtureInf` parameter estimates for $N_{\rm comp} = 1, 2$ are included in Table 1. These are reasonably close to the `sklearnGMM` and `XDGMM` results and consistent with `CPNest` within that method's uncertainty intervals. For $N_{\rm comp} = 3$, `MixtureInf` returns 2 components identical to the $N_{\rm comp} = 2$ model and a completely negligible third, so this is not listed separately.

The EM-test statistic is a type of penalised likelihood. The standard p-value assigned to the $N_{\rm comp} = 1$ null hypothesis by `MixtureInf` is computed under an asymptotic large-$N_{\rm data}$ assumption; nominally this returns $p = 0.087$ for the DNS $M_{\rm T}$ data set but due to the low $N_{\rm data} = 15$ this should be interpreted carefully. A better understanding of the actual hypothesis test power can be achieved through repeating the EM-test on simulated data, see appendix A, indicating that at a nominal p-value threshold (e.g. $p = 0.05$) this test rejects too few data sets generated with $N_{\rm comp} = 1$. Still, compared with those simulations, the obtained EM-test result does not allow a confident rejection of the $N_{\rm comp} = 1$ hypothesis either. Another caveat is that `MixtureInf` by default does not include measurement uncertainties.

3 CONCLUSIONS

The distribution of total masses $M_{\rm T}$ 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 ($M_\mathrm{T} = 2.57 \pm 0.10$); 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 constraining 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 ASCII tables in the online material for this paper; see also appendix D.

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`; to P.L. for suggesting Fig. 5; and to the anonymous MNRAS referee for suggesting the EM-test.

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:1810.02581)
Abbott B. P., et al., 2018b, Phys. Rev. Lett., 121, 161101
Abbott B. P., et al., 2018c, 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
Chen J., Li P., 2009, Ann. Statist., 37, 2523
Chen J., Tan X., Zhang R., 2008, Statistica Sinica, 18, 443
Chen J., Li P., Fu Y., 2012, J. Am. Stat. Assoc., 107, 1096
Ciuperca G., Ridolfi A., Idier J., 2003, Scand. J. Stat., 30, 45
Corongiu A., Kramer M., Stappers B. W., Lyne A. G., Jessner A., Possenti A., D'Amico N., Loehmer O., 2006, A&A, 462, 703
Dempster A. P., Laird N. M., Rubin D. B., 1977, J. Royal Stat. Soc. B, 39, 1
Do C. B., Batzoglou S., 2008, Nature biotechnology, 26, 897
Gregory P. C., 2005, Bayesian Logical Data Analysis for the Physical Sciences. Cambridge University Press
Heavens A., 2009, preprint (arXiv:0906.0664)
Holoien T. W.-S., Marshall P. J., Wechsler R. H., 2017a, XDGMM wrapper class, https://github.com/tholoien/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
Jaynes E. T., 2003, Probability Theory. The Logic of Science. Cambridge University Press
Kiziltan B., Kottas A., De Yoreo M., Thorsett S. E., 2013, ApJ, 778, 66
Li S., Chen J., Li P., 2016, MixtureInf: Inference for Finite Mixture Models, https://cran.r-project.org/package=MixtureInf
Liddle A. R., 2007, MNRAS, 377, L74
Lyne A. G., et al., 2000, MNRAS, 312, 698
Ma P.-X., Jiang J.-L., Wang H., Jin Z.-P., Fan Y.-Z., Wei D.-M., 2018, ApJ, 858, 74
Margalit B., Metzger B. D., 2017, ApJ, 850, L19
Özel F., Psaltis D., Narayan R., Villarreal A. S., 2012, ApJ, 757, 55
Pankow C., 2018, Astrophys. J., 866, 60
Patruno A., Haskell B., Andersson N., 2017, ApJ, 850, 106
Pedregosa F., et al., 2011, J. Machine Learning Res., 12, 2825
Raftery A. E., 1995, Soc. Meth., pp 111–163
Schwab J., Podsiadlowski P., Rappaport S., 2010, ApJ, 719, 722
Schwarz G. E., 1978, Ann. Stat., 6, 461
Skilling J., 2004, in Fischer R., Preuss R., Toussaint U. V., eds, AIP Conf. Proc. Vol. 735, 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering. pp 395–405, doi:10.1063/1.1835238
Smits R., Kramer M., Stappers B., Lorimer D. R., Cordes J., Faulkner A., 2009, A&A, 493, 1161
VanderPlas J. T., Connolly A. J., Ivezic Z., Gray A., 2014, in Proc. CIDU 2012. IEEE (arXiv:1411.5039), doi:10.1109/CIDU.2012.6382200
Veitch J., Del Pozzo W., et al., 2017, CPNest 0.1.4 - Parallel nested sampling in python, doi:10.5281/zenodo.835874
Zhang C. M., et al., 2011, A&A, 527, A83

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 N_data in a data set. For each step of 10 in N_data, 50 data sets are drawn, and the differences in $\log_{10}\mathcal{L}$, AIC, AICc and BIC evaluated between one- and two-component models fitted with `sklearnGMM` and without measurement errors. (AIC is added here to demon-

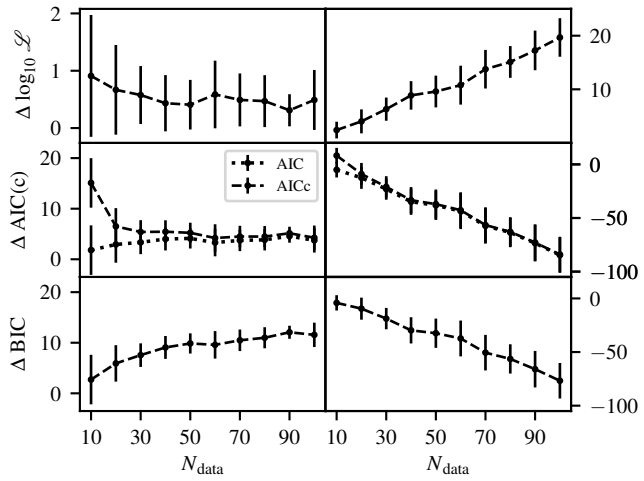

Figure A1. Model selection criteria for GMMs with $N_{\rm comp} = 1, 2$ applied to simulated $M_{\rm T}$ populations drawn from one- and two-component GMMs with the fiducial values reported by Huang+. 50 independent populations are drawn for each $N_{\rm data}$ step. Each difference ΔQ is taken as $Q_{N_{\rm comp}=2} - Q_{N_{\rm comp}=1}$. Hence, positive $\Delta \log_{10} \mathcal{L}$ 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.

strate the convergence with AICc; Bayesian evidences are omitted due to their higher computational cost.)

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 small $N_{\rm data}$, the AICc strongly prefers the one-component model due to its correction term, while for large $N_{\rm data}$ it converges to the AIC at Δ of +3 to +5, and does not strengthen the case any further due to the fixed complexity penalty of $2\,N_{\rm coeffs}$. On the other hand, the BIC continues to develop a stronger preference for the 'true' model for increasing $N_{\rm data}$ due to its different penalty term.

On the other hand, if the 'true' model is a two-component GMM, all selection criteria agree in collecting very strong preference ($\Delta < -10$) for it by $N_{\rm data} \gtrsim 20$ to 30, and continue to strengthen this preference as $N_{\rm data}$ increases.

Intuitively, this dichotomy makes sense: For a 'true' single Gaussian and even for large $N_{\rm data}$, a two-component GMM fit can still always approximate the observed distribution by simply having the components overlap almost completely. The likelihood is then almost the same, and information criteria can only decide by their penalty term. However, for draws from a two-component GMM, large $N_{\rm data}$ will make any one-component fit disagree strongly with the data, and the fit improvement of two components will easily make the $\ln \mathcal{L}$ contribution dominate over any penalty term.

A similar set of simulations is also useful to study the EM-test for low $N_{\rm data}$, putting the nominal results obtained in Sec. 2.6 into proper perspective in the non-asymptotic regime. From 6000 simulations of $N_{\rm data} = 15$ samples from 'true' single-component $M_{\rm T}$ distributions with the nominal Huang+ parameters, only 2.8% produce p-values below the nominal 5% threshold. This indicates that, taking the p-value estimated from the asymptotic expressions at face value, the null hypothesis would not be rejected often enough, and hence the $p = 0.087$ result for the DNS data set cannot necessarily be taken as failure to reject $N_{\rm comp} = 1$. The empirical 95% quantile of EM-test statistics from these simulations is 4.82, corresponding to an asymptotic p-value of 0.09, so that the DNS result is suggestively close to this threshold. However with $p = 0.05$ being a rather lenient threshold for the physical sciences to begin with, and the caveat of the `MixtureInf` EM-test implementation not considering measurement errors, this borderline result cannot be confidently interpreted as evidence for $N_{\rm comp} > 1$ either.

For comparison, on simulations with $N_{\rm data} = 15$ and 'true' $N_{\rm comp} = 2$, an EM-test with chosen threshold of $p = 0.05$ on the asymptotic p-value would reject the $N_{\rm comp} = 1$ hypothesis in about 48% of cases. And for $N_{\rm data} = 150$, the empirical rejection rate indeed becomes $\approx$ 5% for $N_{\rm comp} = 1$ simulations and $\approx$ 100% for $N_{\rm comp} = 2$ simulations.

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 $M_{\rm T}$. Constraints enforce $\sum_k C_k = 1$ and $\mu_{k+1} \geq \mu_k \forall k$.

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 $[1.8, 4.0]\,M_\odot$ yields almost unchanged posterior estimates for $N_{\rm comp} = 1$ and slightly broader, but consistent estimates for $N_{\rm comp} = 2$. 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 $N_{\rm comp} = 2$ posteriors, but only marginally influences the overall estimates. The evidence values change somewhat with the prior volume, but the ratios remain indecisive: $\mathcal{Z}_2/\mathcal{Z}_1 \approx \exp(8.63 - 8.32) \approx 1.4$ for the first set of priors, $\mathcal{Z}_2/\mathcal{Z}_1 \approx \exp(6.28 - 7.04) \approx 0.5$ for the uniform μ_k priors and $\mathcal{Z}_2/\mathcal{Z}_1 \approx \exp(9.55 - 8.94) \approx 1.8$ for the narrower σ_k priors.

Posterior estimates: Let us consider the obtained posterior distributions a bit more closely, especially the $N_{\rm comp} = 1, 2$ cases illustrated in Fig. 4. The posterior parameter estimates given in Table 1 are medians ± 90% quantiles. However, the $N_{\rm comp} = 2$ posteriors have some asymmetric, and for σ_k 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 $M_{\rm T}$ data set and the fact that a single-component Gaussian is effectively included in the paramer space of a two-component GMM. In the $N_{\rm comp} = 2$ posteriors, the smaller secondary peak at high σ_1 and low σ_2 corresponds to very low C_2: in this part of parameter space, the posterior $p(M_{\rm T})$ 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 σ_1, high σ_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 $N_\mathrm{comp} = 1$ 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 $N_\mathrm{comp} = 3$ 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 $N_\mathrm{comp} \geq 2$ 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 $N_\mathrm{comp} = 1$, 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 $\log_{10}\mathcal{L} \pm 0.05$. For $N_\mathrm{comp} \geq 2$ 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_spin for a population of 29 neutron stars in Low Mass X-ray Binary (LMXB) systems. The data set from Table 2 of Patruno et al. (2017) is reproduced in Table D2 and as an ASCII file lmxbs_1705.07669v1.txt provided online with this paper. Patruno et al. have also fit one- and two-component GMMs to this data set (with an `R` implementation of the EM 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`, `XDGMM` and `MixtureInf` implementations return consistent parameters as reported in Table C1. The $N_\mathrm{comp} = 2$ 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 $N_\mathrm{comp} = 1, 2, 3$ (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 ($\mathcal{L}_1/\mathcal{L}_2 \approx 10^{-4}$), but only additional criteria can yield a robust decision, and on its own, the additional gain of $\mathcal{L}_2/\mathcal{L}_3 \approx 0.07$ could tempt us to use an even more complex model. Fortunately, in this case

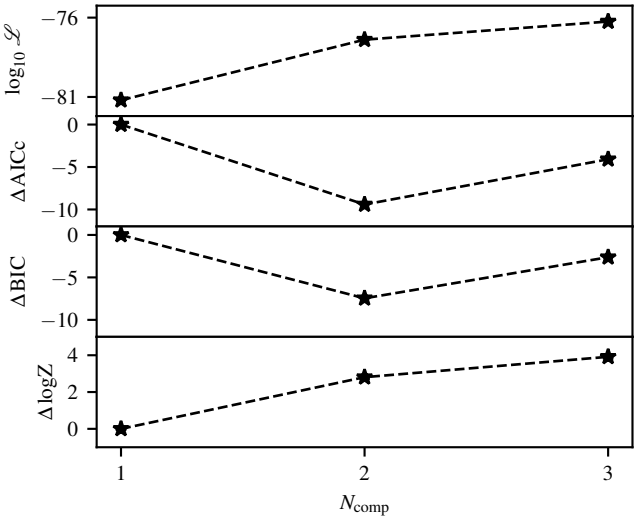

Figure C1. Model selection criteria for GMMs with $N_\mathrm{comp} = 1, 2, 3$ applied to the Patruno+ data set of f_spin for NSs in LMXBs. The first three panels are for `sklearnGMM` fits. The last panel gives the average Bayesian evidence from 10 `CPNest` runs at each N_comp. AICc, BIC and $\ln\mathcal{Z}$ are plotted as differences to the $N_\mathrm{comp} = 1$ value.

AICc and BIC agree in clearly preferring two components over one (with Δ of 7–11), and also prefer two over three by Δ 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. A `MixtureInf` EM-test reports a nominal p-value of 0.008 which also seems to reject a single-component hypothesis much more clearly than in the DNS case. As in Sec. 2.6 the exact value needs to be interpreted with caution due to the low N_data, but this being only a comparison example, no further simulation tests have been conducted in this case.

`CPNest` results for this data set (with uniform priors in C_k, log-uniform in σ_k within $[10, 350]$ Hz and Gaussian μ_k priors of width 70 Hz) are very consistent with the other GMM fits, except for some differences in the more degenerate $N_\mathrm{comp} = 3$ case. The evidence ratio is in favour by ≥ 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, EM-test 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.

Table C1. GMM results for the $f_{\rm spin}$ data set for NSs in LMXBs from Patruno et al. (2017), assuming negligible measurement errors.

	$N_{\rm comp}$	C_1	μ_1	σ_1	C_2	μ_2	σ_2	C_3	μ_3	σ_3	$\log_{10}\mathcal{L}$	AICc	BIC	$\ln\mathcal{Z}$
`sklearnGMM`	1	1.00	414	153							-81.20	378.42	380.69	
or `XDGMM`	2	0.60	308	99	0.40	576	29				-77.39	369.00	373.23	
	3	0.22	201	30	0.36	365	55	0.41	574	30	-76.25	374.33	378.07	
`CPNest`	1	$1.00^{+0.00}_{-0.00}$	410^{+34}_{-34}	157^{+31}_{-24}										-190.26±0.05
	2	$0.62^{+0.14}_{-0.15}$	315^{+42}_{-38}	110^{+39}_{-28}	$0.38^{+0.15}_{-0.14}$	574^{+15}_{-25}	31^{+27}_{-11}							-187.45±0.08
	3	$0.31^{+0.40}_{-0.31}$	255^{+64}_{-67}	76^{+62}_{-58}	$0.33^{+0.26}_{-0.27}$	381^{+60}_{-50}	91^{+66}_{-59}	$0.35^{+0.14}_{-0.14}$	577^{+15}_{-19}	30^{+19}_{-12}				-186.35±0.08
`MixtureInf`	1	1.00	414	155										
	2	0.60	307	100	0.40	575	41							

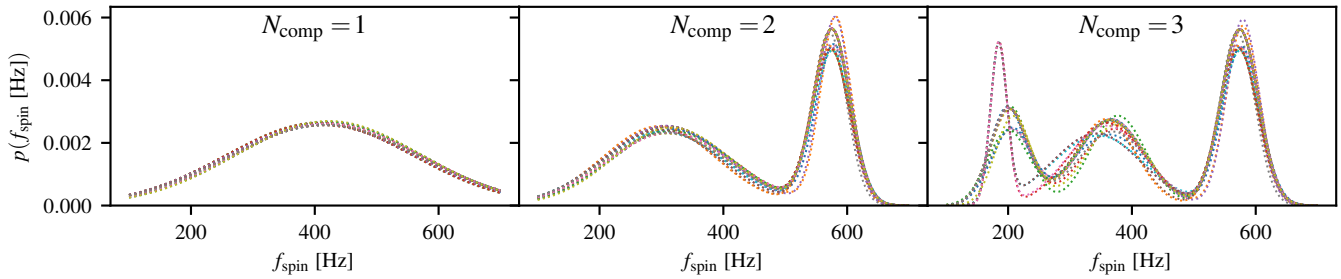

Figure C2. Illustration of leave-one-out cross-validation on the Patruno+ $f_{\rm spin}$ data set. Each dotted line corresponds to a fit after leaving out one data point. In this case, the fitted distributions are relatively stable not just for $N_{\rm comp} = 1$, but also for the preferred fit with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

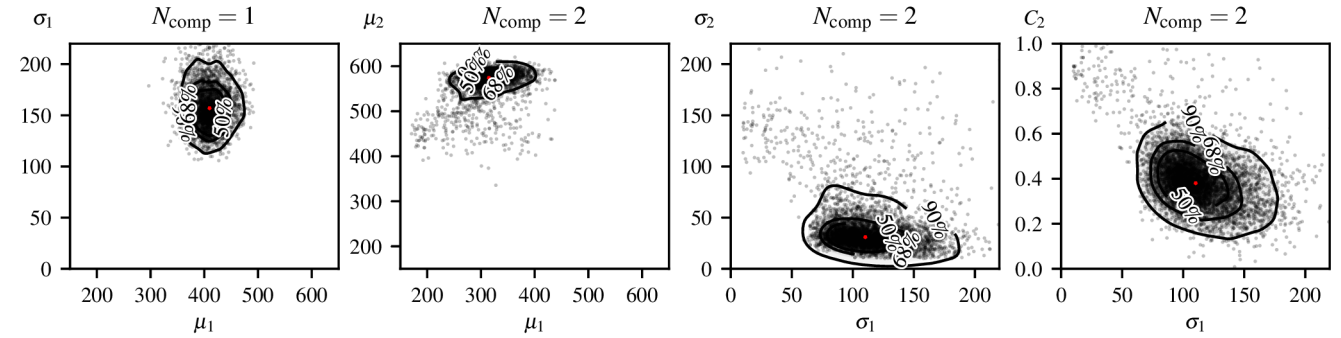

Figure C3. Posteriors from `CPNest` runs on the LMXB data set for a single Gaussian (left panel) and a two-component GMM (remaining panels). Note that σ_1 appears twice on the x-axis of the last two panels.

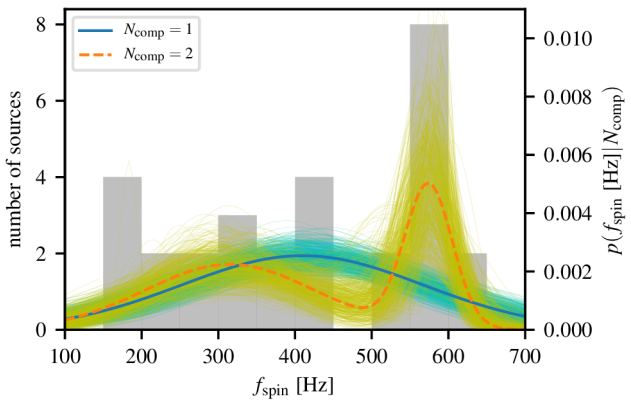

Figure C4. Median GMM distribution functions $p(f_{\rm spin})$ inferred with `CPNest` for one (blue solid line) or two components (orange dashed line), superimposed on a data histogram, and with 90% quantiles background 'haze'.

Table D1. Data set of DNS total mass measurements M_{T} with errors δM_{T}, reproduced from Huang et al. (2018). Also available online as dns_1804.03101v1.txt.

system	M_{T}	δM_{T}
J1411+2551	2.538	0.022
J1757-1854	2.73295	0.00009
J0453+1559	2.734	0.003
J0737-3039	2.58708	0.00016
J1518+4904	2.7183	0.0007
B1534+12	2.678428	0.000018
J1756-2251	2.56999	0.00006
J1807-2500B	2.57190	0.00073
J1811-1736	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
J1930-1852	2.59	0.04
B2127+11C	2.71279	0.00013

APPENDIX D: DATA SETS AND ONLINE MATERIAL

The DNS and LMXB data sets used in this study are given in Tables D1 and D2 and these are also available online as ASCII files. See the original publications for details on data provenance.

The `CPNest` posterior sample chains discussed in Sec. 2.5 and appendices B and C are available as online-only ASCII tables in the following files:

- CPNest_posterior_DNS_Mt_Nc1.dat for a $N_{\mathrm{comp}} = 1$ model on DNS M_{T} data, containing 2880 samples with columns [mu1,logsigma1,logL,logPrior];
- CPNest_posterior_DNS_Mt_Nc2.dat for $N_{\mathrm{comp}} = 2$ on DNS M_{T} data, containing 6159 samples with columns [mu1,logsigma1,mu2,logsigma2,weight2,logL,logPrior];
- CPNest_posterior_LMXB_fspin_Nc1.dat for $N_{\mathrm{comp}} = 1$ on LMXB f_{spin} data, containing 2661 samples with columns [mu1,logsigma1,logL,logPrior];
- CPNest_posterior_LMXB_fspin_Nc2.dat for $N_{\mathrm{comp}} = 2$ on LMXB f_{spin} data, containing 4727 samples with columns [mu1,logsigma1,mu2,logsigma2,weight2,logL,logPrior].

This paper has been typeset from a TEX/LATEX file prepared by the author.

Table D2. Data set of LMXB spin frequencies f_{spin}, reproduced from Patruno et al. (2017). Also available online as lmxbs_1705.07669v1.txt.

source	f_{spin}
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
NGC6440_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