Astronomy & Astrophysics manuscript no. problems2
25th July 2023

A space of parameter spaces in the space sciences: parametric Bayesian inference in astronomy, cosmology and particle physics

Johannes Buchner

Max Planck Institute for Extraterrestrial Physics, Giessenbachstrasse, 85741 Garching, Germany

25th July 2023

ABSTRACT

A sample of parametric Bayesian inference applications from astronomy, cosmology and particle physics is studied, augmented by mock data sets and toy problems. The parameter spaces and posterior distributions are characterized by (1) the number of model parameters, (2) whether the posterior shape is similar to a gaussian, (3) whether the posterior has light or heavy tails, (4) how small the posterior is compared to the prior, i.e., how informative the data are, (5) whether some parameters remain unconstrained while others are highly constrained, (6) whether the posterior has multiple, disconnected modes. These axis define a parameter space of inference problems. We characterize each of the inference problems and observe that inference in astrophysics spans the entire parameter space, from low to high dimensionality, mono- to multi-modal, and a variety of complex distributions that range from uninformative to highly informative. Furthermore, the computational cost of the physical models can range from milliseconds to dozens of seconds. The collated sample of inference problems is proposed as a standard test bed for new samplers. For reproducibility and ease of use, a Docker compute image is provided.

Key words. Bayesian inference; parametric models; astrophysics

1. Introduction

Fitting parametric models to experimental data is one of the key methods to infer physical parameters. In physics, investigation of distant processes is possible by modelling the measurement process accurately. Here, we focus on problems where the model has continuous parameters with predefined prior ranges, and where a likelihood function has been defined to compare the model prediction to data. Some examples include fitting the power spectrum of the Cosmic Microwave Background (CMB) with Dark Energy and Cold Dark Matter (λCDM) cosmologies, fitting time series of the radial velocity of a host star gravitationally pulled by its exoplanets, dissecting multiple components in spectra and population inference from uncertain measurements of many individual objects, such as luminosity or mass functions. The plausible ranges of model parameters that match the data are typically tested in a Bayesian framework once prior and likelihood are specified with Monte Carlo samplers.

Monte Carlo sampling methods of varying complexity have been developed over the last decades. This includes variations of Markov Chain Monte Carlo (MCMC), Particle Monte Carlo (PMC), Importance Samplers (IS) and Nested Samplers (NS). Specific implementations specify the initialisation, exploration strategy (e.g., proposal function) and termination criterion. These are often tuned for the application. Reliable parameter recovery of a method can be tested by Monte Carlo simulating new datasets, and analysing them. An alternative are toy inference problems that approximate features of the real problem. These can be more easily understood and more rapidly analysed. Given the diversity of algorithms and inference problems, it is interesting to consider whether a different algorithm can perform well on the same problem, and whether the currently used algorithm can be transferred to another problem. This work is focusing on the applicability of Monte Carlo Samplers over different types of inference problems.

Inference problems differ substantially by the posterior distribution that a Monte Carlo sampler has to explore. The main characteristics of problems include (1) the number of model parameters, (2) whether the posterior shape is similar to a gaussian, (3) whether the posterior has light or heavy tails, (4) how small the posterior is compared to the prior (i.e., how informative the data are), (5) whether some parameters remain unconstrained while others are highly constrained, (6) whether the posterior has multiple, disconnected peaks. Besides a systematic classification of inference problems based on five characteristics, this work presents a diverse set of real and toy inference problems that cover the entire classification space.

2. Data: collated inference problems

To cover most of the problem space, we collected inference problems published in the literature and available in open-source physics packages. Appendix A introduces real-world problems, which form the main sample in this work. Simplified mock problems with generated data sets that approximate real-world problems are presented in Appendix B. Appendix C introduces artificial toy problems.

For each problem, the likelihood function $L(\theta)$ is defined together with prior distribution $\pi(\theta)$ over the parameter space.

3. Method: Characterization of parametric inference

Various difficulties are encountered by different sub-disciplines. Here we specify six characteristics and give a mathematical definition for each. Finally, we present a visual presentation which characterizes the posterior degeneracies.

3.1. Dimensionality

In astrophysics, fitting problem dimensionalities range from 1 to millions parameters. Examples of extremely high-dimensional problems include pixel reconstructions (e.g.,) and . With more parameters, the possible combinations of parameter values rises exponentially (the curse of dimensionality). This makes the problem complex to explore and distances between parameter space points meaningless.

Here we defined three common sub-groups: low-dimensional ($d = 2 - 9$), mid-dimensional ($d = 10 - 29$) and high-dimension ($d \geq 30$). The boundaries are set near where simple and more sophisticated ideas of geometric sampling start failing. Extremely high dimensions ($d \gg 100$) are not the focus of this work. We note that these virtually always require the derivatives of likelihood functions to effectively navigate the parameter space. The availability of likelihood derivatives could be considered an additional classification category.

3.2. Information gain ("Depth")

Depending on the data quality of the experiment, the posterior may be a tiny region of the prior, or be identical with the prior. This can be quantified by the Kullback-Leibler divergence, or surprise, between the two probability distributions:

$$D_{\mathrm{KL}} = \int \pi(\theta) \ln \frac{\pi(\theta)}{P(\theta)} d\theta$$

Here, π and P give the prior and posterior over the parameter vector space θ. In the case of a base-e logarithm, the unit of $D_{\mathrm{KL}}$is nats, and approximately means how many e-foldings it takes to cut the prior until the posterior is reached. Finding that small region can be a challenge for sampling algorithms (and maximum likelihood minimizers).

In practice, the information gain is already computed by nested sampling algorithms internally for error estimation, and we adopt that method as a measurement. Also, the information gain is related to the number of iterations of the nested sampling algorithms needs to zoom in until the likelihood appears flat.

3.3. Modes

When data can be explained with similar quality by different combinations of processes, the posterior exhibits multiple peaks. This is common in fits of multiple components with (nearly-)interchangable predictions, paired with poor descrimination power of the data. Algorithms based on local jumps can find it difficult to navigate between modes, because a proposal tuned to a single mode may reach another distant mode with vanishing proposal probability. The bottom row of Figure 1 shows examples of multimodal distributions.

To mathematically define multi-modality, a threshold criterion is needed to define disconnectedness. In principle any clustering algorithm can be used. For simplicity and reproducibility, we adopt a simple approach. First, histograms of the marginal posteriors are histogrammed into 20 bins. Bins with density less than $1/5$ of the peak density are considered "empty". Gaps are identified, and thresholds that bracket the peaks extracted. This is repeated for every dimension. Then, all combinations of brackets are computed. These are the clusters. If posterior samples are members of multiple clusters, the clusters are merged. The number of remaining, non-empty clusters is the number of modes of the problem.

3.4. Non-Gaussianity

The Bernstein–von Mises theorem states that when the data are highly informative, the posterior is shaped like a multi-variate Gaussian. Then, Laplace's approximation of the posterior with a log-quadratic density function is justified. This can furthermore occur if the model is linear in its parameters, or a first-order Taylor expansion of the model provides a reasonable approximation at the maximum a posteriori. For this reason, some algorithms are constructed to behave optimally when the posterior is Gaussian (see e.g., Laplace approximation).

The Gaussianity of a posterior can be easily measured through the surprise from a best-fitting multi-variate Gaussian to the posterior. We obtain a suitable gaussian from the mean and covariance of the posterior samples. In multi-modal cases, the posterior is highly non-gaussian.

3.5. Tail weight ("Width")

While the posterior may have ellipsoidal contours like a Gaussian, the posterior density may decline steeper or shallower than a square-exponential, i.e., have thin or heavy tails. For example, when outliers are allowed (e.g., in student-t distribution or explicit outlier modelling), the wings of the posterior can be wide. Some algorithms may be optimized for square-exponential declines. Another cause of heavy tails is when most data points are fitted well by one component, and a minor component relevant for a small data subset improves the fit slightly (for example, in blind spectral line searches on top of a continuum). This leads to a phase transition, where the parameter space to be explored changes rapidly (with small likelihood change) from a wide volume to a narrow volume. This is difficult for many samplers.

To quantify how the weight is distributed over the prior volume, we compute the prior volume above a given likelihood threshold. This is illustrated in Figure 2. We compute the 5% and 95% quantiles of the volume ranges where most of the probability mass resides, and compute the tail weight as:

$$\mathrm{TW} = \log \frac{V_{5\%}}{V_{95\%}}$$

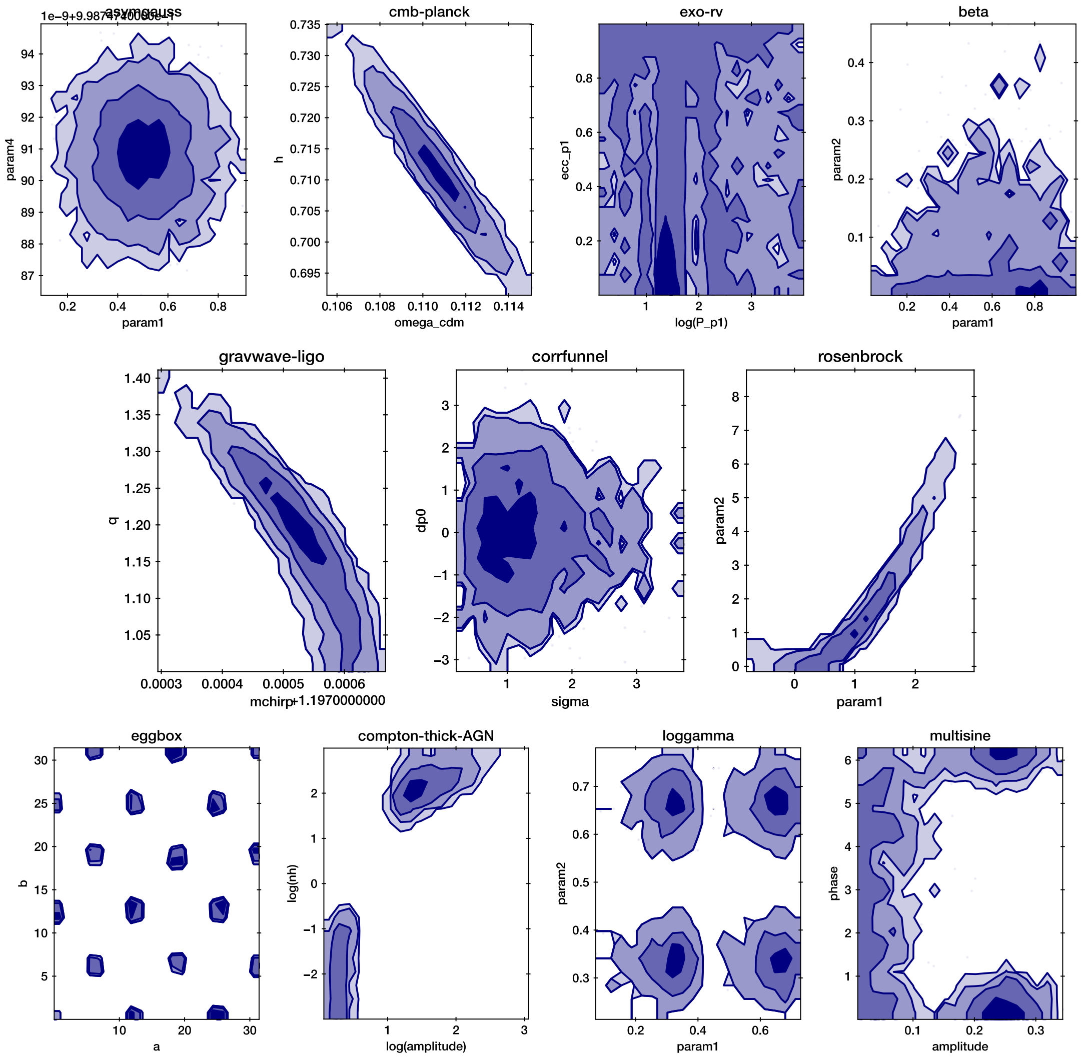

Fig. 1. Selected pair-wise posterior distributions from some of our problems. The 1, 2, 3 and 4-sigma equivalent probability mass contours (made with corner; Foreman-Mackey 2016) illustrate non-linear degeneracies (e.g., middle panels), unequal axes (e.g., left-most and right-most top panels), multi-modality (bottom panels). The loggamma problem (third panel, bottom row) also has heavy tails towards the left. Some parameters are uninformative (e.g., param1 in top right panel) or at the prior parameter edge (middle left panel, ratio $q \geq 1$, top right panel, $0 \leq \text{param2} \leq 1$).

In practice, the mapping of volume and likelihood, as well as the normalising constant, the marginal likelihood integrated over volume shrinkages, is already computed internally in nested sampling. We cap the value at $\mathrm{TW} = \log 10 \times d$.

3.6. Inequality

Some model parameters may alter the model prediction strongly, while others have more minute implications. Because of this, the posterior of some parameters can be consistent with the prior with no information learnt. For other parameters, its plausible range may have diminished by several orders of magnitude. Proposals that are isotropic over the parameter space directions then may perform poorly. The top row of Figure 1 shows such examples in pairs of parameters.

We quantify the inequality of parameters in problems, as:

$$\mathrm{IE} = \frac{\max_i \left\{\max\left(1\,\mathrm{bit}, \mathrm{IG}_i\right)\right\}}{\min_i \left\{\max\left(1\,\mathrm{bit}, \mathrm{IG}_i\right)\right\}}.$$

Here, the information gain IG_i in bits for each parameter θ_i is computed from the prior $\pi(\theta_i)$ and posterior $P(\theta_i|D)$ marginal distributions as:

$$\mathrm{IG}_i = \int \pi(\theta_i) \log_2 \frac{\pi(\theta_i)}{P(\theta_i|D)} d\theta_i$$

If all parameters have similar posterior uncertainties, $\mathrm{IE} \approx 0$. We cap this value to at most $\mathrm{IE} = 100$.

3.7. Phase transitions

Phase transitions are sudden, unexpected changes in volume when the likelihood is changed continuously. An example is shown in Figure 3. This is common in multi-component fitting, where the rough overall data is first approximated, and then fine tuning a second component onto

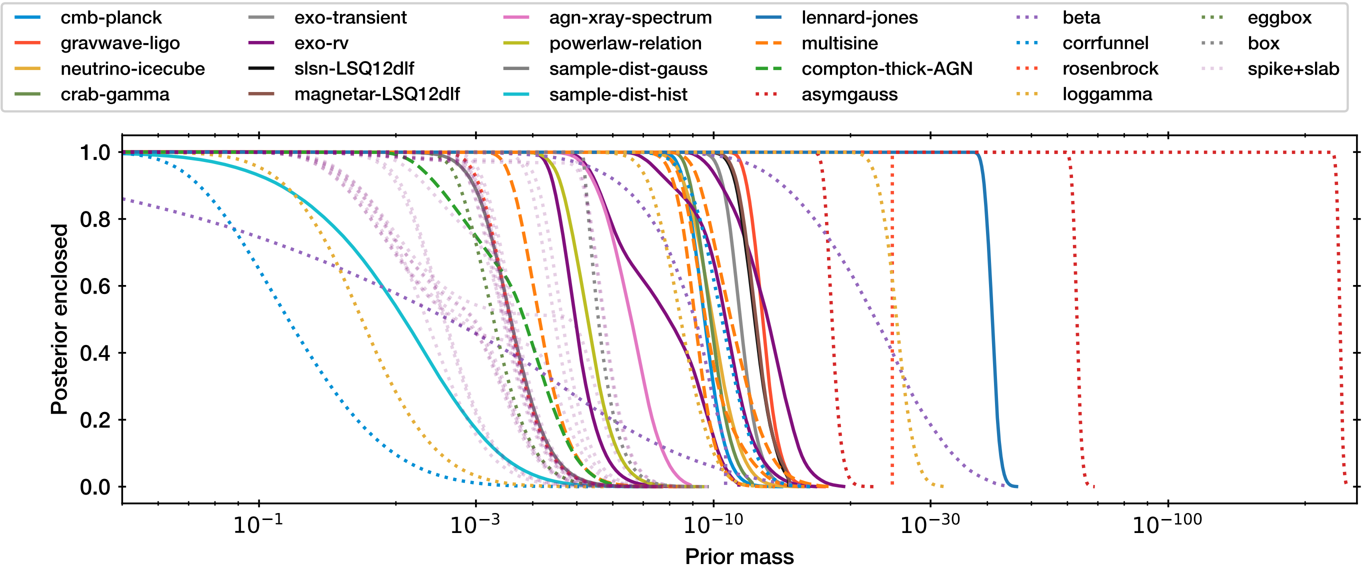

Fig. 2. Probability fraction enclosed as a function of prior volume. From the highest likelihood regions outwards, the volume is increased from left to right until the entire prior space ($V = 1$) is enclosed. The median (cross) indicates how large the posterior volume is relative to the prior, and is related to the *information gain*. Quantiles at 5% and 95% indicated as circles indicate the shell where most probability mass is enclosed. This is related to the *tail weight*. Some problems show wide transitions (e.g., green dashed) relative to a gaussian (blue dashed).

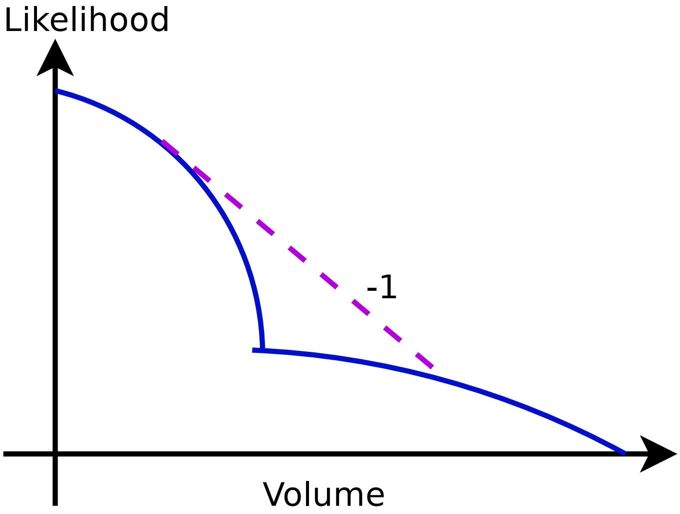

Fig. 3. In the prior volume - likelihood plot of nested sampling, phase transitions are visible as non-convexities. The dashed curve illustrates our technique for closing the curve.

a subset of the data leads to a rapid further increase in the likelihood. The situation has been likened in Skilling (2009) to condensation of a water vapor into liquid water, where small temperature changes leads to a sudden change in volume. This can affect some inference techniques, because the relevant geometry of the (proposal) space before and after the transition is radically different.

We measure the presence and the impact of phase transitions. Skilling (2009) discussed that the slope of the likelihood-volume curve deviates strongly from a powerlaw with index -1. Figure 3 illustrates a phase transition in blue, with a wide plateau followed by a spike of small volume. Here, we use an extrapolating powerlaw (purple in Figure 3) with index -1 to raise all points to the right of the curve. From both the modified curve and the original curve, we compute the evidence, and consider the ratio as our phase transition indicator:

$$P = \frac{\int L'(V)\,dV}{\int L(V)\,dV}$$

Here the denominator with L is the normal nested sampling computation of the evidence, while the numerator contains the modified curve, L', as described above.

3.8. Compact visualisation of posterior structure

For a visual impression of the experienced parameter space and its linear and non-linear degeneracies, a common tool are corner plots. These are also known in the statistics literature as pairs plots, and an example is shown in Figure 4. Here, we quantify the further interaction between pairs of parameters and present a compact visualisation as a $d \times d$ matrix. The diagonal entries are filled with the information gain H expressed in bits. The upper and lower triangles are populated with correlation coefficients of pairs of parameters.

Corner plots only consider the posterior distribution. However, to express the experience of a sampler fairly, the characterization should consider also the prior. Here we work with the posterior samples expressed with coordinates on the prior cumulative probability distribution (probability integral transform). In the case of nested samplers working with prior transforms based on the unit hypercube, this merely means we take the un-transformed posterior samples. First, we characterize the linear degeneracies. the pair-wise Pearson correlation coefficient ρ_{Pearson} is calculated for each posterior parameter pair. Secondly, we characterize residual non-linear degeneracies. We apply an affine transform which standardizes the parameter space to be linearly independent, and then compute the pair-wise Spearman correlation coefficient ρ_{Spearman}.

The obtained matrix visualisation is demonstrated in Figure 5, corresponding to the corner plot in Figure 4. The diagonal shows the parameter information gain, the lower triangle indicates the linear degeneracies and the upper triangle indicates the non-linear degeneracies. In the case of a un-correlated gaussian, illustrated in the right panel of Figure 5, the off-diagonal elements are zero.

4. Results

The collated problems are listed in Table 1, together with the derived characteristics. The model evaluation cost is also listed, and ranges from less than 1ms to almost 1s per likelihood evaluations. Figure 7 presents the location of each inference problem in a corner plot. Among the "real" physics inference problems (solid circles), there does not appear to a trend or preferred region of the parameter space; all regions are covered. The mock problems (crosses) cover the full parameter space as well.

The exoplanet radial velocity fitting with one planet shows a phase transition, and heavy-tailed posteriors. These properties are also produced by the spike and slab toy problem variations, calculated here in only two dimensions. Fitting for 3 exoplanets yield 9 modes, and the LSQ12dlf analyses show a similar number of modes. The multisine mock problem mimics the properties of the exoplanet radial velocity fitting in terms of multi-modality, width, parameter asymmetry, phase transition and heavy-tailed posteriors, suggesting it is a useful approximation.

As already shown in Figure 2, the information gain differs substantially. Additionally, the asymmetry column of Table 1 indicates how much some parameters are learned while others are uninformative. This parameter can be very high (>30).

Beyond those summary statistics, Figure 8 presents a visual impression of the posterior structure. Based on the visualisation developed in section 3.8, this presents a diversity in dimensionality, as indicated by the number of entries, the parameter information gain (diagonal entries), the linear parameter degeneracies (bottom left triangles) and non-linear degeneracies (upper right entries). Some inference problems, such as cmb-planck (top left panel) show no non-linear degeneracy, while others (agn-xray-spectrum, bottom left panel) show strong non-linear degeneracies.

Finally, we have a deeper look at the evolving likelihood surface as the sampler moves from the prior to the posterior mass. To this end, we observe the live point distribution at regular snapshots. We apply the same procedure as in Figure 3.8 and note the largest Pearson and (linearly whitened) Spearman correlation coefficient. Figure 9 shows the evolution of these values for each inference problem. For problems with no parameter interactions, such as the asymgauss or beta, the curves remain in the bottom left corner of the plot. Other problems however show strong linear (Pearson) and/or non-linear (whitened Spearman) degeneracies over the parameter space. This behaviour is stronger for real problems (dashed and solid curves) than toy problems (dotted curves).

5. Discussion & future work

5.1. Strengths and limitations of the collected inference problems

This work has assembled a list of parametric Bayesian inference applications from astronomy, cosmology and particle physics. This is augmented by mock applications and toy problems. A reproducible installation is available at [1], allowing researchers to test with identical priors and likelihoods. The parameter spaces and posterior distributions have been characterized and cast into a space of parametric inference spaces.

A major result of this work is that the parameter space structure of these inference spaces is highly diverse. This is illustrated in Figure 7. They span from low to high dimensionality, can be mono-, bi- or multi-modal. A variety of complex posterior distribution morphologies appear as illustrated in Figure 8, and inference on the parameters ranges from uninformative to extremely informative. Furthermore, the computational cost of the physical models can range from milliseconds to seconds.

To test samplers in realistic, but controlled conditions with known posteriors and evidence, toy inference problems are commonly used. Standard ones include a Gaussian, Neal's funnel, log-gamma, eggbox, rosenbrock and spike+slab, which are also included here. However, toy inference problems may not represent the experience of a sampler exploring the parameter space of a real data analysis problem. In particular, Figure 9 demonstrates that real inference problems can have stronger degeneracies than toy problems. There is more work to be done to create toy inference problems that closely mimick real physics inference. In this vain, we provide several small data sets which are easy to work with yet represent real physics data analysis (e.g., multisine, compton-thick-AGN, sample-dist, powerlaw-relation, lennard-jones). In the future, perhaps the more complex data analysis pipelines can be replaced by auxiliary functions, for example based on Gaussian mixtures or normalising flows, which closely approximate the real likelihood function, but have known, analytic properties (integral Z, marginal distributions). Despite the limitations of artificial toy problems, Figure 7 demonstrates that these cover the same (vast) parameter space as the real inference problems (compare the filled circles to crosses).

5.2. Performance of samplers

In recent years, a variety of Bayesian inference sampling packages have been published. MCMC-based algorithms include emcee (Foreman-Mackey et al. 2013), Stan (Carpenter et al. 2017), zeus (Karamanis & Beutler 2020) and pocoMC (Karamanis et al. 2022). Nested sampling-based algorithms include multinest (Feroz et al. 2009), polychord (Handley et al. 2015), DNest4 (Brewer & Foreman-Mackey 2018), dynesty (Speagle 2020), ultranest (Buchner 2020), nautilus (Lange 2023) and i-nessai (Williams et al. 2023). For reference snowline[2] provides a Laplace's approximation implementation[3], which can be improved with Gaussian

[1] `https://github.com/JohannesBuchner/space-of-inference-spaces/`

[2] `https://johannesbuchner.github.io/snowline/`

[3] based on `https://iminuit.readthedocs.io/`, see Dembinski & et al. (2020); James & Roos (1975)

field	name	dim	cost	depth	width	modes	asym	!gauss	phase
cosmology	cmb-planck	6	659	1.59	2.4	1	1.34	0.01	0.003
gravitational waves	gravwave-ligo	7	4	1.83	2.6	1	2.14	0.05	0.004
astroparticle	neutrino-icecube	16	146	0.62	3.2	1	29.68	0.12	0.004
supernova remnants	crab-gamma	6	76	1.64	2.6	1	3.27	0.05	0.006
exoplanets	exo-transient	9	4	1.27	2.5	1	33.09	0.08	0.003
	exo-rv-0	2	3	2.48	1.8	1	1.19	0.00	0.000
	exo-rv-1	7	3	1.14	5.0	3	33.09	0.44	0.225
	exo-rv-2	12	2	0.88	5.9	6	33.07	0.30	0.015
	exo-rv-3	17	3	0.77	6.3	9	6.16	0.06	0.031
transients	slsn-LSQ12dlf	12	16	1.03	2.8	6	14.22	0.15	0.023
	magnetar-LSQ12dlf	12	14	1.03	2.8	7	17.31	0.17	0.020
extragalactic	agn-xray-spectrum	4	4	1.65	2.7	1	3.86	0.20	0.007
	powerlaw-relation	3	3	1.76	2.0	1	1.18	0.01	0.001
	sample-dist-gauss	2	1	1.79	2.1	1	1.03	0.08	0.001
	sample-dist-hist	11	3	0.19	2.4	1	11.51	0.08	0.005
materials	lennard-jones-6	12	0	3.42	4.2	1	1.00	0.11	0.003
mock	multisine-0	2	1	2.07	1.8	1	1.51	0.00	0.000
	multisine-1	5	0	1.82	2.5	2	2.00	0.02	0.002
	multisine-2	8	0	1.20	4.3	20	32.43	0.50	0.002
	multisine-3	11	0	0.99	5.2	66	33.04	0.54	0.006
	compton-thick-AGN	5	1	0.77	2.7	2	7.42	0.02	0.037
toy	asymgauss-4d	4	0	4.55	2.1	1	3.07	0.02	0.001
	asymgauss-16d	16	0	3.93	3.8	1	3.79	0.08	0.002
	asymgauss-100d	100	0	2.38	10.4	1	3.01	0.31	0.012
	beta-2d	2	0	1.31	11.9	1	11.78	12.88	0.010
	beta-10d	10	0	0.91	4.8	8	4.50	0.10	0.014
	beta-30d	30	0	0.76	26.4	1024	12.19	1.04	0.039
	corrfunnel-2d	2	0	0.58	1.7	1	4.82	0.02	0.001
	corrfunnel-10d	10	14	1.04	4.5	1	1.65	0.28	0.004
	rosenbrock-2d	2	0	1.78	1.9	1	1.16	0.02	0.001
	rosenbrock-20d	20	0	1.24	-1.3	1	1.00	1.27	0.000
	loggamma-2d	2	0	0.84	1.9	4	1.15	0.01	0.001
	loggamma-10d	10	0	0.85	3.3	4	1.76	0.01	0.005
	loggamma-30d	30	6	0.84	5.9	4	1.82	0.12	0.007
	eggbox-2d	2	0	1.64	1.8	18	1.00	0.01	0.001
	box-5d	5	0	1.12	1.2	1	1.00	0.35	0.000
	spikeslab-1-2d-4	2	2	1.17	2.2	1	1.01	0.19	0.001
	spikeslab-1-2d-40	2	2	1.49	3.1	1	1.01	0.84	0.190
	spikeslab-1-2d-400	2	2	1.34	4.2	1	1.03	1.71	0.358
	spikeslab-1-2d-4000	2	2	2.41	5.0	1	1.01	2.03	0.510
	spikeslab-40-2d-4	2	2	1.27	1.9	1	1.01	0.01	0.001
	spikeslab-40-2d-40	2	2	1.79	2.0	1	1.00	0.04	0.005
	spikeslab-40-2d-400	2	2	2.32	2.2	1	1.02	0.10	0.024
	spikeslab-40-2d-4000	2	2	2.86	4.4	1	1.02	0.29	0.053
	spikeslab-1000-2d-4	2	2	1.29	1.9	1	1.00	0.01	0.001
	spikeslab-1000-2d-40	2	2	1.83	1.9	1	1.00	0.00	0.002
	spikeslab-1000-2d-400	2	2	2.23	1.9	1	1.00	0.07	0.001
	spikeslab-1000-2d-4000	2	2	2.87	1.9	1	1.00	0.22	0.003
	spikeslab-1-2d-40-off1	2	2	1.48	3.0	1	1.01	0.82	0.173
	spikeslab-1-2d-40-off2	2	2	1.36	3.1	2	1.00	0.93	0.187
	spikeslab-1-2d-40-off4	2	2	1.43	3.0	2	1.00	0.88	0.193
	spikeslab-1-2d-40-off10	2	2	1.74	2.7	1	1.01	0.41	0.028
	spikeslab-40-2d-40-off1	2	2	1.78	2.2	1	1.01	0.06	0.006
	spikeslab-40-2d-40-off2	2	2	1.81	2.0	1	1.00	0.03	0.007
	spikeslab-40-2d-40-off4	2	2	1.76	2.1	1	1.01	0.11	0.006
	spikeslab-40-2d-40-off10	2	2	1.73	1.9	1	1.00	0.13	0.001
	spikeslab-1000-2d-40-off1	2	7	1.78	1.9	1	1.01	0.02	0.001
	spikeslab-1000-2d-40-off2	2	7	1.72	1.9	1	1.01	0.02	0.001
	spikeslab-1000-2d-40-off4	2	7	1.77	1.9	1	1.01	0.11	0.001
	spikeslab-1000-2d-40-off10	2	7	1.76	1.9	1	1.01	0.15	0.001

Table 1. List of problems analysed. The columns describe (1) research field, (2) name, (3) dimensionality, (4) model evaluation cost in milliseconds, (5) information gain, (6) tail weight, (7) parameter inequality, (8) Gaussian approximation information loss, (9) phase transition.

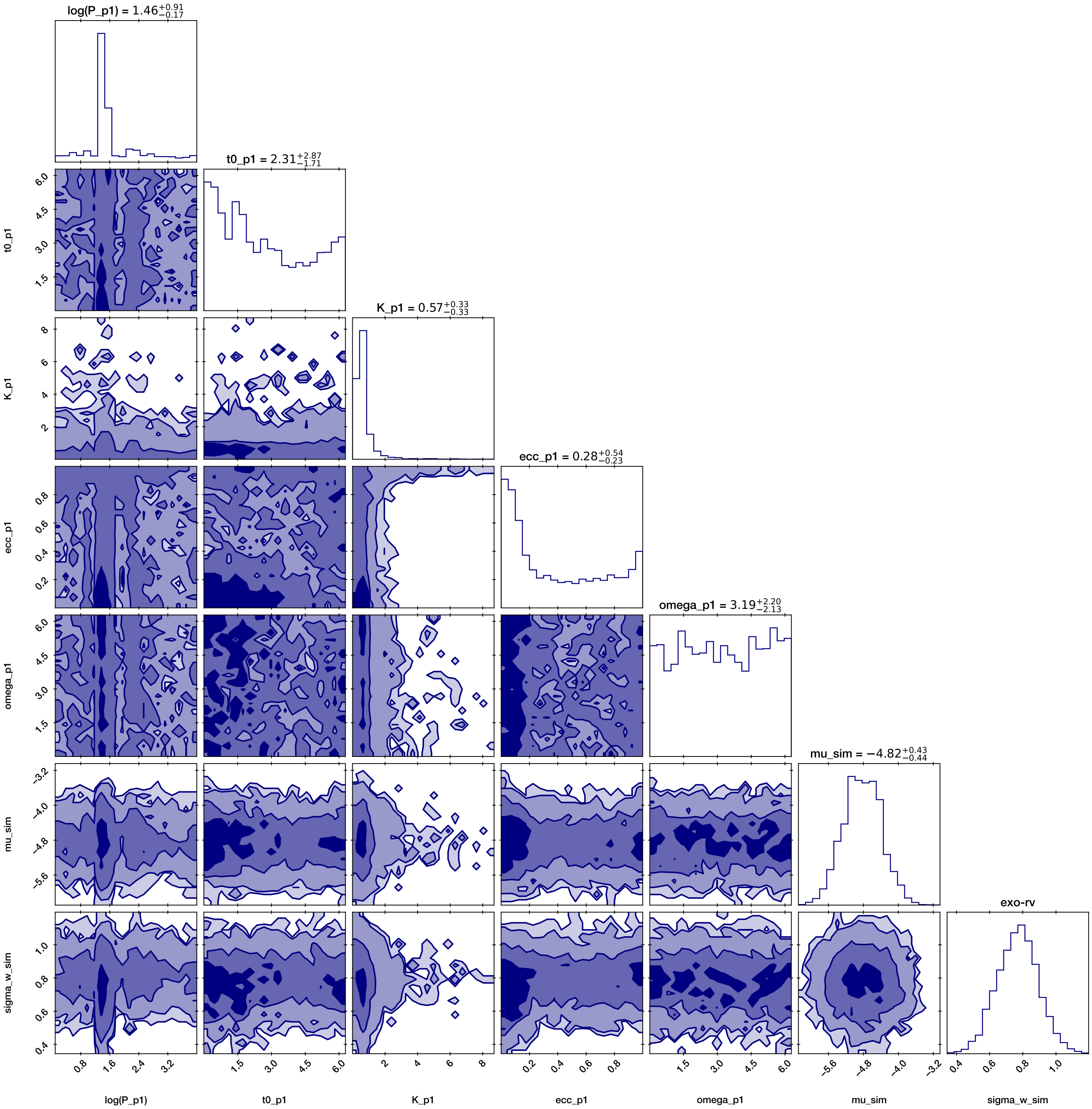

Fig. 4. Corner plot (created with corner.py; Foreman-Mackey 2016) for the exo-rv problem with 1 planet. The degeneracies can be non-linear (see e.g., ecc_p1 and K-p1). Some parameters are uninformative (omega_p1), while others are very well constrained.

mixtures learned by variational Bayes and importance sampling [4].

It is interesting to firstly try to understand in which regions of the parameter space an inference algorithm gives reliable results. This is achievable with problems where the truth is known (toy problems or generated data sets). Only a subsequent step is performance comparison among reliable algorithms. To this end, we point out a few guidelines for comparing Bayesian samplers fairly. Initially, the full sample of problems presented here should be considered. If only a subset is of interest, this should be clearly and transparently stated (e.g., focus on low-dimensional inference problems, on mono-modal inference problems, etc.). It is a trivial statement that with more compute budget, a better accuracy and reliability can be achieved. Therefore, we recommend plotting bias or accuracy against computational cost for each algorithm and runtime. One quantifier of computational cost is the number of (likelihood) model evaluations or wall-clock time. The former is suitable for computationally challenging likelihoods, the latter is more

[4] based on `https://github.com/pypmc/pypmc`, see Beaujean & Caldwell (2013)

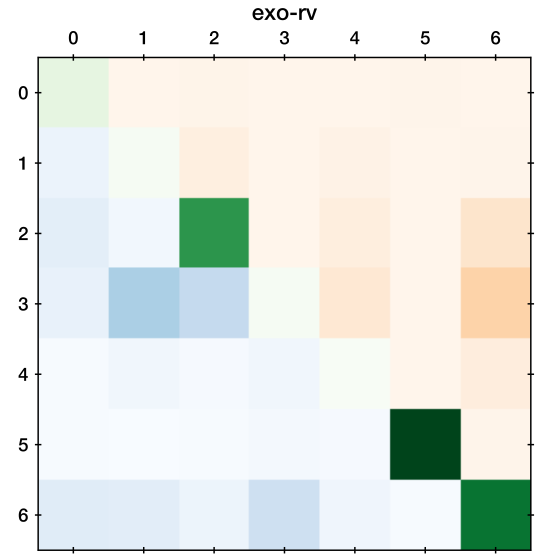

Fig. 5. Proposed visualisation of the degeneracies, corresponding to Figure 4. The lower left triangle in blue shows linear correlation between parameter posteriors ($|\rho_{\mathrm{Pearson}}|$ from 0 to 1), with darker color indicating stronger correlation. Similarly, the upper right traingle in orange shows Spearman rank correlation between parameter posteriors, after removing the linear correlations. The diagonal entries in green indicate the information gain on the parameter, increasing from white to darker colors. Here, some parameters are uninformative while others are highly inform

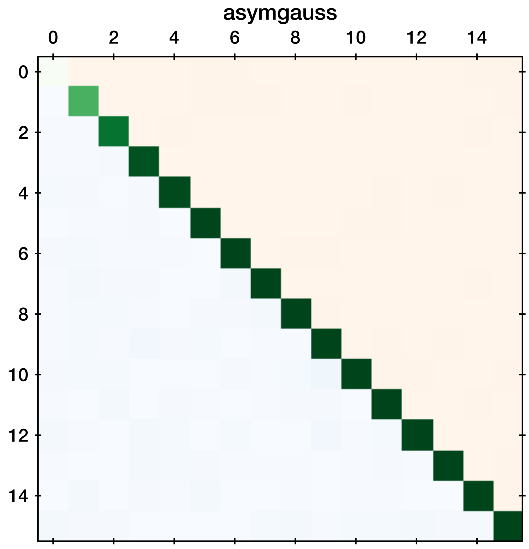

Fig. 6. Same as Figure 5, but for our 16 dimensional Gaussian toy problem. Here, the non-diagonal entries are near zero, indicating (correctly) that there are no parameter degeneracies. The diagonal elements vary, with the first parameter being much less informative than the others.

relevant for computationally cheap models paired with algorithms that require costly training (such as deep neural networks).

It remains to be defined how to quantify the fidelity of the computation. To quantify posterior fidelity, see for example the probability-probability plots and Jensen-Shannon divergence quantification in Romero-Shaw et al. (2020). However, the reliable retrieval of 3σ upper or lower parameter limits may also be of interest. Finally, when the true marginal likelihood Z is known and can be tested against, recovering $\ln Z$ to an accuracy much better than ~ 0.3 may not be useful in the context of Bayes factors, as it only mildly affects the interpretation.

5.3. Future work

Breakthrough progress in machine learning is typically driven by (1) open, large, high-quality data sets and (2) a clear formulation of a meaningful objective. Examples span from MNIST's challenge of digitizing hand-drawn numbers to the Critical Assessment of Structure Prediction (CASP) (Moult et al. 1995) protein-folding challenge, recently met by AlphaFold (Jumper et al. 2021). The "Learning to learn by gradient descent by gradient descent" paper (Andrychowicz et al. 2016) demonstrated that optimization algorithms can be derived by machine learning. Perhaps a similar breakthrough can be accomplished for optimal Bayesian inference sampling procedures, by providing a open, large data base and a clear objective. To this end, the survey of inference problems encountered across cosmology, particle physics, astrophysics and astronomy is presented. The representative database includes fully specified likelihood and priors in a runnable docker image with python interfaces. While similar previous work has focusd on providing simplified models that serve as unit-tests (e.g., Sountsov et al. 2020; Magnusson et al. 2021), this work uses real-world inference with the software pipelines employed by researchers.

As a first step, existing and novel algorithms can be judged across the parameter space of problems for their empirical behavior and robustness, and to make well-founded recommendation for specific applications. For more rapid testing and a unified interface, fast model emulators for the provided real-world examples would be useful. This is left for future work.

A step further into the future is comparable to Atari computer game playing artificial intelligences (Bellemare et al. 2012) that learn optimal game playing strategies with reinforcement learning: A playground for learning optimal Bayesian inference algorithms.

Acknowledgements. I thank Frederik Beaujean for insightful conversations. I thank Gana Moharram, who studied the IceCube neutrino analysis for a bachelor thesis with Philipp Eller, for insightful conversations.

References

Aartsen, M. G., Ackermann, M., Adams, J., et al. 2018, Phys. Rev. Lett., 120, 071801
Aartsen, M. G., Ackermann, M., Adams, J., et al. 2020, European Physical Journal C, 80, 9
Aartsen, M. G., Ackermann, M., Adams, J., et al. 2019, Phys. Rev. D, 99, 032007
Abbott, B. P., Abbott, R., Abbott, T. D., et al. 2017, ApJ, 848, L12
Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., & O'Neil, M. 2014
Andrychowicz, M., Denil, M., Gomez, S., et al. 2016, arXiv e-prints, arXiv:1606.04474
Baronchelli, L., Nandra, K., & Buchner, J. 2018, ArXiv e-prints
Beaujean, F. & Caldwell, A. 2013, ArXiv e-prints
Bellemare, M. G., Naddaf, Y., Veness, J., & Bowling, M. 2012, arXiv e-prints, arXiv:1207.4708
Betancourt, M. J. & Girolami, M. 2013, arXiv e-prints, arXiv:1312.0906
Brahm, R., Hartman, J. D., Jordán, A., et al. 2018, AJ, 155, 112
Brewer, B. J. 2016, A Rosenbrock challenge for MCMC folks...

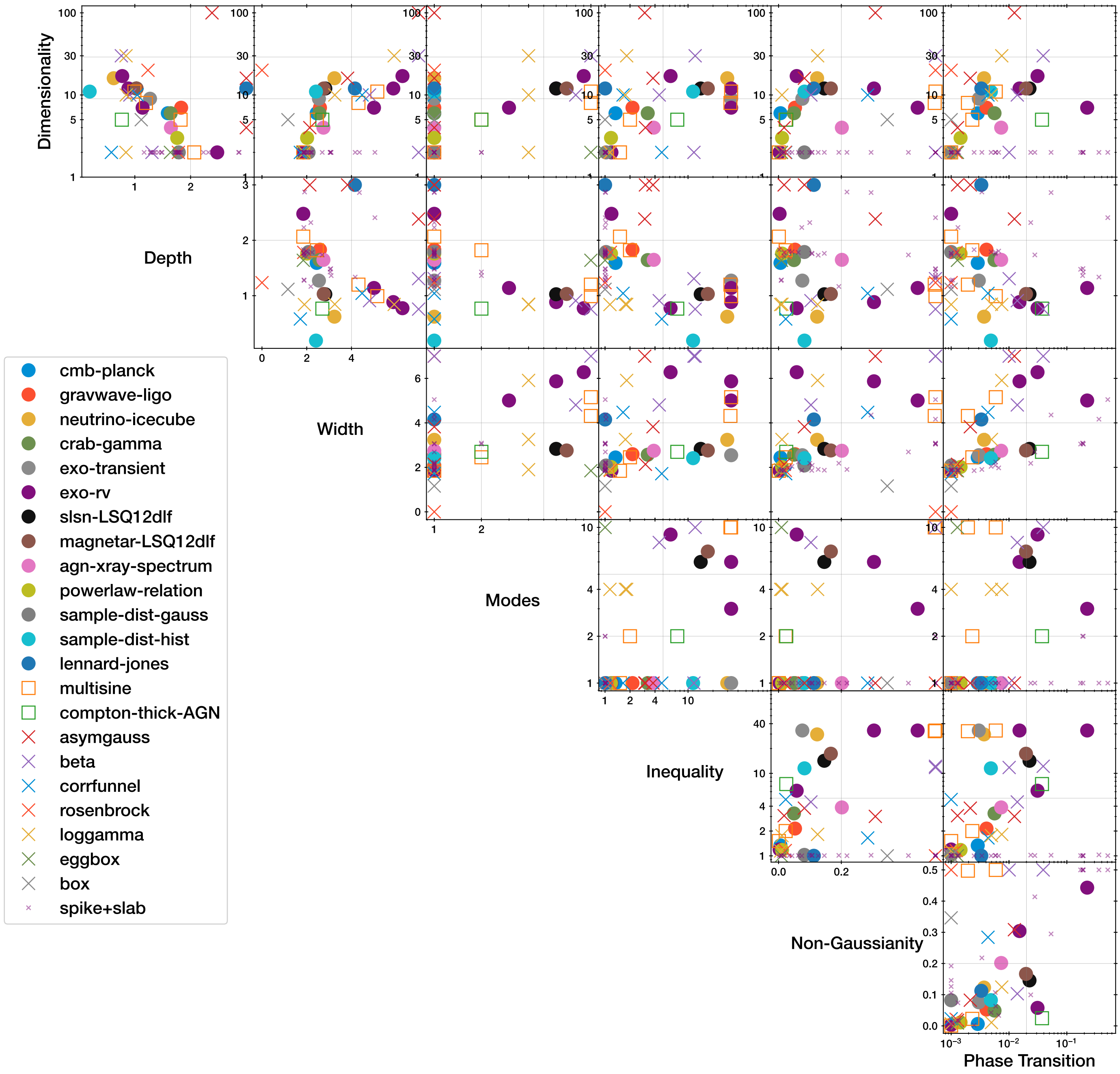

Fig. 7. Parameter space of the problems over the defined six characteristics.

Brewer, B. J. & Foreman-Mackey, D. 2018, Journal of Statistical Software, Articles, 86, 1

Buchner, J. 2020, UltraNest v2.2.1, `https://johannesbuchner.github.io/UltraNest/`

Buchner, J. 2021, The Journal of Open Source Software, 6, 3045

Buchner, J., Brightman, M., Nandra, K., Nikutta, R., & Bauer, F. E. 2019, A&A, 629, A16

Buchner, J., Georgakakis, A., Nandra, K., et al. 2014, A&A, 564, A125

Carpenter, B., Gelman, A., Hoffman, M. D., et al. 2017, Journal of statistical software, 76, 1

Dembinski, H. & et al., P. O. 2020

Espinoza, N., Kossakowski, D., & Brahm, R. 2019, MNRAS, 490, 2262

Feroz, F. & Hobson, M. P. 2008, MNRAS, 384, 449

Feroz, F., Hobson, M. P., & Bridges, M. 2009, MNRAS, 398, 1601

Foreman-Mackey, D. 2016, The Journal of Open Source Software, 1, 24

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306

Freeman, P., Doe, S., & Siemiginowska, A. 2001, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 4477, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, ed. J.-L. Starck & F. D. Murtagh, 76–87

Fulton, B. J., Petigura, E. A., Blunt, S., & Sinukoff, E. 2018, PASP, 130, 044504

Handley, W. J., Hobson, M. P., & Lasenby, A. N. 2015, MNRAS, 453, 4384

James, F. & Roos, M. 1975, Comput. Phys. Commun., 10, 343

Jumper, J., Evans, R., Pritzel, A., et al. 2021, Nature, 596, 583

Karamanis, M. & Beutler, F. 2020, arXiv e-prints, arXiv:2002.06212

Karamanis, M., Nabergoj, D., Beutler, F., Peacock, J., & Seljak, U. 2022, The Journal of Open Source Software, 7, 4634

Kipping, D. M. 2013, MNRAS, 435, 2152

Kormendy, J. & Ho, L. C. 2013, ARA&A, 51, 511

Kreidberg, L. 2015, Publications of the Astronomical Society of the Pacific, 127, 1161

Lange, J. U. 2023, arXiv e-prints, arXiv:2306.16923

Magnusson, M., Bürkner, P., & Vehtari, A. 2021

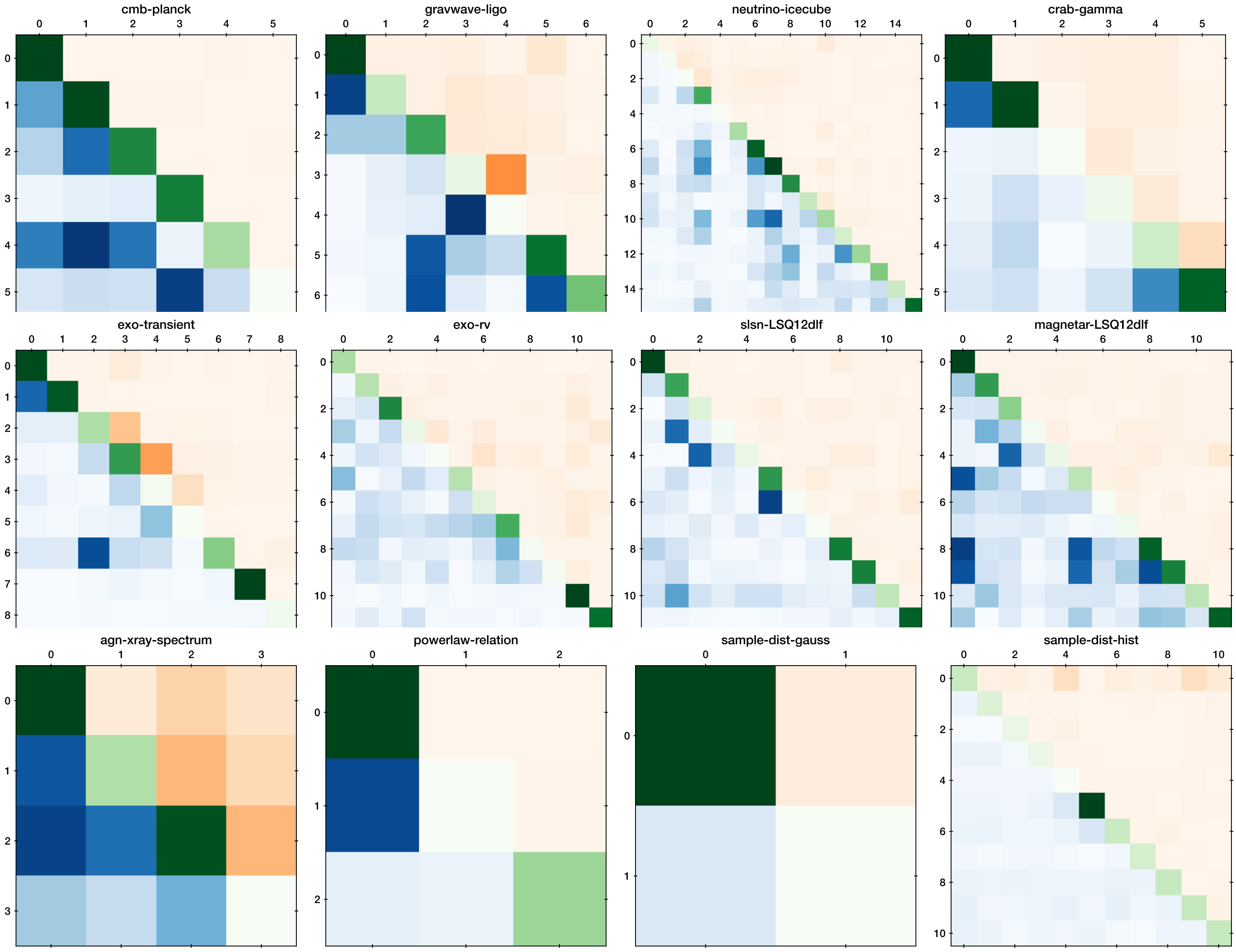

Fig. 8. Posterior structure visualisations (see section 3.8) for our real physics inference problems. The number of entries indicates the dimensionality, diagonal entries indicate parameter information gain, the bottom left triangle entries indicate linear parameter degeneracies and the upper right triangle indicates non-linear degeneracies. For example, the top left panel shows no non-linear degeneracy, while the bottom left panel shows strong non-linear degeneracies.

Moult, J., Pedersen, J. T., Judson, R., & Fidelis, K. 1995, A large-scale experiment to assess protein structure prediction methods
Nelson, B. E., Ford, E. B., Buchner, J., et al. 2020, AJ, 159, 73
Nicholl, M., Guillochon, J., & Berger, E. 2017, ApJ, 850, 55
Nitz, A., Harry, I., Brown, D., et al. 2023, gwastro/pycbc: v2.1.2 release of PyCBC
Romero-Shaw, I. M., Talbot, C., Biscoveanu, S., et al. 2020, MNRAS, 499, 3295
Skilling, J. 2009, in BAYESIAN INFERENCE AND MAXIMUM ENTROPY METHODS IN SCIENCE AND ENGINEERING: The 29th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Vol. 1193, AIP Publishing, 277–291
Sountsov, P., Radul, A., & contributors. 2020, Inference Gym
Speagle, J. S. 2020, MNRAS, 493, 3132
van Dokkum, P., Danieli, S., Cohen, Y., et al. 2018, Nature, 555, 629
van Dyk, D. A., Connors, A., Kashyap, V. L., & Siemiginowska, A. 2001, ApJ, 548, 224
Williams, M. J., Veitch, J., & Messenger, C. 2023, arXiv e-prints, arXiv:2302.08526

Appendix A: Real inference problems

Appendix A.1: Cosmology with the Cosmic Microwave Background

(if you are a cosmologist, please help me with references!)

The Cosmic Microwave Background (CMB) is the oldest electromagnetic signal observable. It originated when the Universe underwent a transition from being so dense that photons would be constantly scattered to the current state where photons can travel freely. These photons allow us to measure the temperature of their regions of origin approximately 379,000 years after the Big Bang. A map of their emission over the sky gives information about the temperature correlation. This carries information of how the Big Bang inflation proceeded, which is dependent on important constituents of the Universe, such as the baryonic matter and dark matter content of the Universe. The CMB remains one of the most important experiments to measure cosmological parameters in the dark energy and cold dark matter cosmology framework (ΛCDM). One difficulty in the fitting of cosmological models like ΛCDM to the CMB is that the prediction of angular correlations is computationally expensive (of the order of a few seconds per likelihood evaluation).

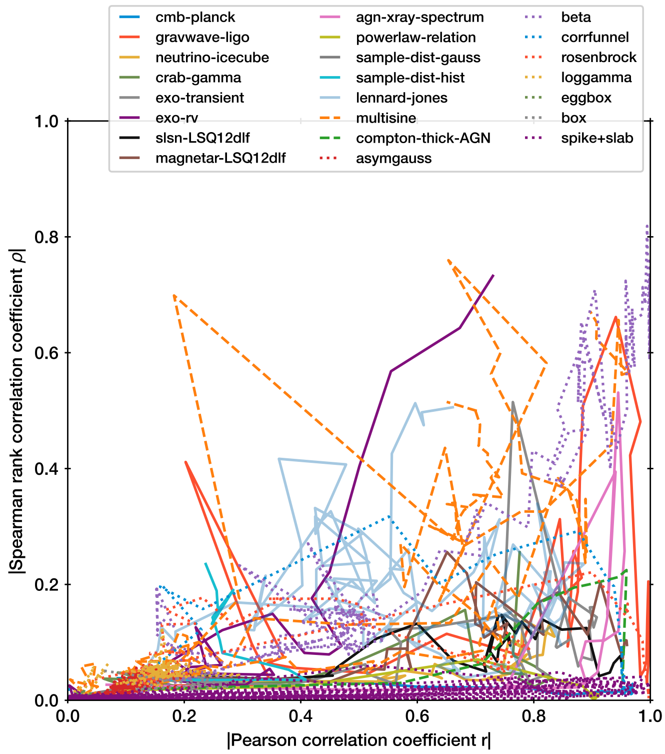

Fig. 9. For each problem, the evolution of the Pearson correlation coefficient (x-axis) and the Spearman correlation coefficient as nested sampling raises the likelihood threshold is shown as a curve. Low coefficients (bottom left corner) indicate independent parameters, high x-axis values indicate strong linear degeneracies. High y-axis values indicate strong non-linear degeneracies, such as multiple modes, or bananas in the parameter space.

Here we adopt an example[5] of the MontePython cosmology fitting package[6]. We use the fake_planck_bluebook likelihood which emulates a Planck measurement. The free parameters are ΛCDM cosmological parameters, namely the baryonic density Ω_b (between 1.8 and 3), the dark matter density Ω_cdm (between 0.1 and 0.2), the scalar spectral index n_s (between 0.9 and 1.1), A_s (between 1.8 and 3), the hubble parameter value, relative to 100 km/s/Mpc, h (between 0.6 and 0.8), and the time of reionisation τ_reion (between 0.004 and 0.12). All priors are uniform within the mentioned bounds.

Appendix A.2: Gravitational wave analysis

Gravitational waves originate from distortions of space-time by compact objects. Recently, the development of multiple, extremely sensitive instruments have allowed the observation of two black holes merging. GW170817 (Abbott et al. 2017) was the first gravitational event detected by three detectors and allowed for the first time localisation on the sky, albeit with substantial parameter degeneracies.

Here we adopt a tutorial example[7] of the PyCBC gravitational wave analysis package (Nitz et al. 2023) . The merging system GW170817 is described by the mass ratio q, the chirp mass m_chirp, which is a combination of the two masses that influences the signal amplitude, the inclination i of the system relative to the observer, the time of coalescence t_c in seconds, the distance d and position on the sky (RA, DEC). The priors adopted are:

$$\begin{aligned}
q &\sim && \mathrm{Uniform}(1,2)\\
m_\mathrm{chirp} &\sim && \mathrm{Uniform}(1,2)\\
\sin i &\sim && \mathrm{Uniform}(0,1)\\
t_c - t_m &\sim && \mathrm{Uniform}(0.02,0.05)\\
d &\sim && \mathrm{Uniform}(10,100)\\
RA &\sim && \mathrm{Uniform}(0,2\pi)\\
\cos DEC &\sim && \mathrm{Uniform}(-1,1)
\end{aligned}$$

where t_m is the time automatically associated by an automated pipeline searching for candidate events in the noise.

Appendix A.3: Atmospheric Neutrino Oscillations from IceCube

IceCube is a neutrino detector near the south pole, which has collected atmospheric neutrino data (Aartsen et al. 2018,?, 2019, 2020) that can be studied for neutrino oscillation. Here we adopt a tutorial example of the PISA IceCube analysis package[8] applied to three-year IceCube data, where the muon and neutrinos model predicts a 2d-histogrammed signal, which is compared to collected data. The parameters are adopted from the PISA defaults:

$$\begin{aligned}
\text{nue\_umu\_ratio} &\sim && \text{Gauss}(1,\ 0.05)\\
\text{Barr\_uphor\_ratio} &\sim && \text{Gauss}(0,\ 1)\\
\text{Barr\_nu\_nubar\_ratio} &\sim && \text{Gauss}(0,\ 1)\\
\text{delta\_index} &\sim && \text{Gauss}(0,\ 0.1)\\
\text{theta13} &\sim && \text{Gauss}(8.5^\circ,\ 0.205^\circ)\\
\text{theta23} &\sim && \text{Uniform}(31^\circ,\ 59^\circ)\\
\text{deltam31} &\sim && \text{Uniform}(0.001\,\text{eV},\ 0.007\,\text{eV})\\
\text{aeff\_scale} &\sim && \text{Uniform}(0,\ 3)\\
\text{nutau\_norm} &\sim && \text{Uniform}(-1,\ 8.5)\\
\text{nu\_nc\_norm} &\sim && \text{Gauss}(1,\ 0.2)\\
\text{opt\_eff\_overall} &\sim && \text{Gauss}(1,\ 0.1)\\
\text{opt\_eff\_lateral} &\sim && \text{Gauss}(25,\ 10)\\
\text{opt\_eff\_headon} &\sim && \text{Uniform}(-5,\ 2)\\
\text{ice\_scattering} &\sim && \text{Gauss}(0,\ 10)\\
\text{ice\_absorption} &\sim && \text{Gauss}(0,\ 10)\\
\text{atm\_muon\_scale} &\sim && \text{Uniform}(0,\ 5)
\end{aligned}$$

[5] `https://github.com/JohannesBuchner/montepython_public/blob/3.5/input/example_ns.param`

[6] hosted at `https://github.com/brinckmann/montepython_public/`

[7] based on `https://github.com/gwastro/PyCBC-Tutorials/blob/7f5ff8fdd40c5dce2237b6082e1755b4aba9b989/tutorial/inference_1_ModelsAndPEByHand.ipynb`

[8] based on `https://github.com/icecube/pisa/blob/f91224b58360ee9ecefe4bdb232249263c9eee17/pisa_examples/IceCube_3y_oscillations_example.ipynb`

Appendix A.4: Gamma-rays from the Crab pulsar-wind nebula

The Crab nebula is the brightest astrophysical source in the sky at high energies. The Large-Area Telescope (LAT) onboard Fermi has observed emission from the region, which is a super-position of the Crab Pulsar (PSR J0534+2200), and synchrotron and Inverse Compton emission from the Crab Nebular. We follow the tutorial [9] of the 3ML multi-messenger package .

All parameters are assigned uninformative priors. The parameters include the normalisation (uniform from 0 to $1.4 \times 10^{10}\mathrm{keV}^{-1}\mathrm{s}^{-1}\mathrm{cm}^{-2}$) and spectral index (uniform from -10 to 10) of the "super_cutoff_powerlaw" model for PSR J0534+2200, the normalisation for the power law spectrum of NVSS J052622+224801 (log-uniform from 10^{-20} to 1.1×10^{-14}) and 4FGL J0544.4+2238 (log-uniform from 10^{-20} to 1.39×10^{-13}), and the isotropic background normalization and the galactic background factor.

Appendix A.5: Exoplanet transit observations

Photometric light curves of stars can show small, periodic dips in brightness, which are interpreted as transits of exoplanets in front of the star. We follow the tutorial of the juliet package[10] to analyse HATS-46 light curve observed with the Transiting Exoplanet Survey Satellite (see Brahm et al. 2018, for details). The model is essentially a constant light curve modified by a U-shaped dip that repeats with some period. We use juliet (Espinoza et al. 2019) for the implementation, which in turn uses batman (Kreidberg 2015) for the light curve modelling together with limb-darkening (Kipping 2013).

The model parameters include the properties of the planet (period P in days, time-of-transit center t0, planet-to-star radius ratio p, eccentricity ecc, argument of periastron passage omega), as well as nuisance parameters (dilution factor mdilution, offset relative flux mflux, jitter sigma, Limb-darkening parameters q1 and q2):

$$\begin{aligned}
\mathrm{P\_p1} &\sim & \mathrm{Gauss}(4.7,\ 0.1)\\
\mathrm{t0\_p1} &\sim & \mathrm{Gauss}(1358.4,\ 0.1)\\
\mathrm{r1\_p1} &\sim & \mathrm{Uniform}(0,\ 1)\\
\mathrm{r2\_p1} &\sim & \mathrm{Uniform}(0,\ 1)\\
\mathrm{q1\_TESS} &\sim & \mathrm{Uniform}(0,\ 1)\\
\mathrm{q2\_TESS} &\sim & \mathrm{Uniform}(0,\ 1)\\
\mathrm{ecc\_p1} &= & 0\\
\mathrm{omega\_p1} &= & 90\\
\mathrm{rho} &\sim & \mathrm{LogUniform}(100,\ 10000)\\
\mathrm{mdilution\_TESS} &= & 1.0\\
\mathrm{mflux\_TESS} &\sim & \mathrm{Gauss}(0,\ 0.1)\\
\mathrm{sigma\_w\_TESS} &\sim & \mathrm{LogUniform}(0.1,\ 1000)
\end{aligned}$$

Appendix A.6: Exoplanet detection from Radial Velocity data

One of the most efficient ways to detect exoplanets is through changes in the line-of-sight (radial) velocity of individual stars, as the tug of planets gravitationally accelerates them (Doppler shift). A planetary system leads to a complex overlay of periodic velocity changes. Additionally, the measurement of velocities is uncertain, in part because of the instrument accuracy and precision, and in part because the spectral emission lines used to measure Doppler shifts can be unstable due to stellar activity. That latter process has been modeled with Gaussian processes in recent years.

Here we adopt the problem setup of the Extreme Precision Radial Velocity III challenge[11] (Nelson et al. 2020). They simulated several artificial exoplanet systems containing two planets, with a Gaussian process and realistic observation time sampling. The challenge participants were asked to compute the Bayesian marginal likelihoods of each data set assuming that the true number of planets was 0, 1, 2 or 3. Participants did not know the true simulation input, but they were told the exact Gaussian noise properties, which follows a Gaussian process. Here we repeat this exercise, for their dataset 5 (shown in Figure A.1). The exact specification of this problem is defined in (Nelson et al. 2020). Essentially, it is similar to a sine time series fit, except that the periodic signal can be asymmetric due to ellipticity, giving 5 parameters per planet (signal amplitude, period, pericenter time, eccentricity and mean anomaly) that describe a Keplerian orbit. Additionally, the white noise amplitude σ_j and the systematic velocity C are free parameters, giving $2 + N_{\mathrm{planets}} \times 5$ free parameters for $N_{\mathrm{planets}} = 0, 1, 2, 3$. We use juliet (Espinoza et al. 2019) for the implementation, which in turn uses RadVel (Fulton et al. 2018) for Kepler orbits and george (Ambikasaran et al. 2014) for modelling Gaussian processes.

Appendix A.7: X-ray spectral analysis of Active Galactic Nuclei

Active Galactic Nuclei (AGN) are regions in the centres of massive galaxies where super-massive black holes grow. As gas swirls into the black hole, enormous amounts of radiation are produced by release of gravitational energy, sometimes shining brighter than all host galaxy stars together. Close to the black hole, X-rays are also produced and are an important tracer of the mass inflow into the black hole. They also allow identifying AGN in the sky, even when the black hole is surrounded by thick gas and dust, as most of the energetic X-rays penetrate through any obscurers. Space-based X-ray focusing instruments allow measurements of the X-ray spectra, which carry information on the AGN luminosity, obscuring column density and properties of the X-ray emitter (photon index, energy turn-over). The detection of X-ray radiation is performed by counting photon events and capturing an estimate of their energy, time of arrival and location on the sky. However, the energy response and energy-dependent sensitivity of the instrument adds some analysis complexity. Addi-

[9] based on `https://threeml.readthedocs.io/en/stable/notebooks/Fermipy_LAT.html`

[10] based on `https://juliet.readthedocs.io/en/latest/tutorials/transitfits.html#transit-fits`

[11] hosted at `https://github.com/EPRV3EvidenceChallenge/Inputs`

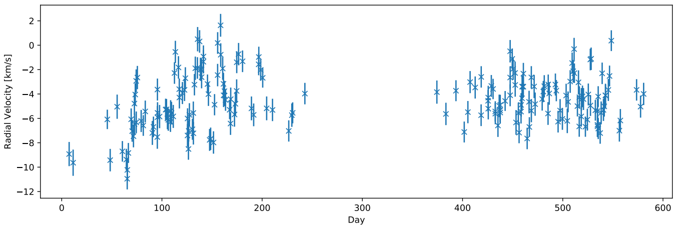

Fig. A.1. Exoplanet Doppler shift time series data (blue).

tionally, when few counts are detected per energy bin, as is commonly the case, the process needs to be modelled with a Poisson likelihood. For more details, see Buchner et al. (2014) and van Dyk et al. (2001).

Here, we include the source of a Chandra deep observation (source ID 179) presented Buchner et al. (2014). The data are available online[12] and the model is defined by the xagnfitter script[13] of the Bayesian X-ray analysis package (BXA; Buchner 2021) based on sherpa (Freeman et al. 2001). The intrinsic X-ray emission is a powerlaw, which is modulated by a physical obscurer model simulation of photo-electric absorption, Compton-scattering and fluorescent line-emission (Buchner et al. 2019). This model is available as a table[14]. To this obscured AGN model, a soft, unobscured powerlaw is added. The background spectrum is added and its normalisation determined simultaneously with a joint fit of the spectrum extracted in a background sky region and the source+background spectrum extracted at the source location. The model has four parameters. This includes a wide log-uniform prior (between 10^{-8} and 10^{-3}) on the powerlaw normalisation, a wide log-uniform prior (between 10^{-7} and 10^{-1}) on the soft power law normalisation relative to the primary powerlaw normalisation, intrinsic power law, a Gaussian prior on the power law photon index with mean 1.95 and standard deviation 0.15, a log-uniform prior on the obscurer column density between 10^{20} and $10^{26}/\mathrm{cm}^2$.

Appendix A.8: Superluminous supernova

Super-Luminous Supernovae (SLSN) are extreme explosions at the end of stellar evolution. By fitting photometric observations over time, the explosion mechanism and its physical parameters can be determined (and distinguished from other transients). We follow the tutorial of MOSFiT [15], to analyse LSQ12dlf, with two models described in Nicholl et al. (2017): a magnetar engine with a simple spectral energy distribution (‘magnetar‘ model) and a magnetar with a modified spectrum and additional constraints (‘slsn‘ model). The parameters are:

nhhost ∼	LogUniform(10^{16}, 10^{23})
Pspin ∼	Uniform(1, 10)
Bfield ∼	LogUniform(0.1, 10)
Mns ∼	Uniform(1, 2)
thetaPB ∼	Uniform(0, 1.5708)
texplosion ∼	Uniform(−500, 0)
kappa ∼	Uniform(0.05, 0.2)
kappagamma ∼	LogUniform(0.1, 10000)
mejecta ∼	LogUniform(0.001, 100)
vejecta ∼	Uniform(5000, 20000)
temperature ∼	LogUniform(1000, 100000)
variance ∼	LogUniform(0.001, 100)
codeltatime ∼	LogUniform(0.001, 100)
codeltalambda ∼	LogUniform(0.1, 10000)

In case of ‘slsn‘, the following parameters differ:

Bfield ∼	Uniform(0.1, 10)
mejecta ∼	LogUniform(0.1, 100)
vejecta ∼	Uniform(5000, 20000)
temperature ∼	Uniform(3000, 10000)

Appendix A.9: Lennard-Jones potential

In material science, the group behaviour of atoms gives rise to structures and phases of matter (solids, liquids, gases, crystals, metals, etc). A standard example is the Lennard-Jones potential, which we adopt here with 6 particles in a box. Pairs of particles feel two forces, a repulsive force at short distances (Pauli repulsion) and an attractive force at longer distances (van der Waals force). This are formulated as:

$$\mathcal{L} = \prod_{i} \prod_{j>i} \left(\frac{\sigma}{r_{ij}}\right)^{6} - \left(\frac{\sigma}{r_{ij}}\right)^{12}$$

where r_{ij} is the Euclidean distance between particles with indices i and j, but at least σ. We set $\sigma = 10^{-3}$, which defines the scale at which particles feel attraction or repulsion.

Each particle’s three-dimensional location (x,y,z) is a free parameter between -1 and +1. To avoid identical modes, we set the likelihood to zero when $|z_1| < |z_2| < ... < |z_n|$ is not satisfied. Due to translation symmetry, we assume the first particle is placed in the coordinate centre (0,0,0). From rotation symmetry, we assume the second

[12] https://github.com/JohannesBuchner/BXA/tree/master/examples/sherpa

[13] https://johannesbuchner.github.io/BXA/xagnfitter.html

[14] https://github.com/JohannesBuchner/xars/blob/master/doc/uxclumpy.rst

[15] based on https://mosfit.readthedocs.io/en/latest/fitting.html#public-data

particle is placed along the positive z direction (0,0,z_2). From a second rotation symmetry, the third particle is placed on the (0,y_3,z_3) plane. The remaining particles have all (x,y,z) coordinates as free parameters.

Appendix A.10: Power-law line fit: Tully-Fisher relation

Stars in the centres of galaxies move akin to a ideal gas, so that the velocity dispersion of galactic central elliptical components (bulges) correlates with the bulge mass. Kormendy & Ho (2013) presented a compilation of measurements of velocity dispersions and masses, both annotated with asymmetric (and heteroscedastic) error bars. Following the "Fitting a line" tutorial[16] of UltraNest, we assume these correspond to Gaussian tails, scaled according to the error bar size. The measurements are fit with a powerlaw, with the intrinsic scatter along the powerlaw accounted for by a log-Normal distribution. The parameters are

$$
\begin{aligned}
\text{slope} \sim &\quad \text{Uniform}(-3,\, 3)\\
\text{offset} \sim &\quad \text{LogUniform}(10,\, 1000)\\
\text{scatter} \sim &\quad \text{LogUniform}(0.001,\, 10)
\end{aligned}
$$

The likelihood is integrating the Log-Normal distribution over the data point distribution, i.e., it is a hierarchical Bayesian model. This integration is performed numerically.

Appendix A.11: Sample distributions: A galaxy without dark matter

Appendix A.11.1: Gaussian Sample distribution

van Dokkum et al. (2018) observed the velocity of globular clusters in a low-mass, ultra-diffuse galaxy. The width of the (Gaussian) distribution of velocities, i.e., the velocity dispersion, is directly related to the total galaxy mass. From comparing the stellar light to the total light, they inferred that the dark matter mass in this galaxy is negligible. Following [17], we analyse their velocity measurements with Gaussian error bars, assuming a Gaussian sample distribution (see e.g., Baronchelli et al. 2018 for such a Gaussian hierarchical model in astronomy). The parameters are the mean, for which we assume a uniform distribution between -100 and 100 km/s, and the scatter, for which we assume a log-uniform distribution between 1 and 1000 km/s. The parameter for each data point's true value is integrated out, as implemented in [18].

Appendix A.11.2: Dirichlet sample distribution

The above example is repeated, but with a more flexible sample distribution. Following the PosteriorStacker tutorial[19], a uniform Dirichlet distribution is adopted, with 11 uniformly spaced bins between -80 km/s and +80 km/s. This analysis allows checking whether a Gaussian was a reasonable model.

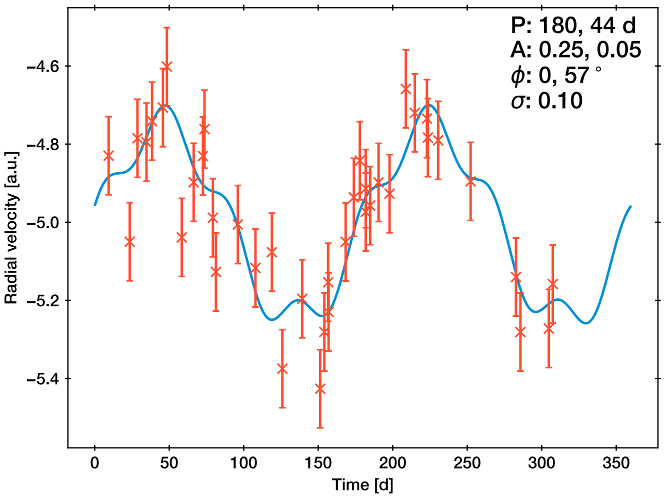

Fig. B.1. Sine time series data (orange) with generating two-component model.

Appendix B: Mock problems

Appendix B.1: Sine time series

Analyses of inhomogeneously sampled light curves is a common problem in astrophysics, sometimes with complex noise processes and semi-periodic signals. Here we present a simple multi-component sine signal as a toy problem, and uniformly randomly sample observing times. The problem is defined as:

$$
\begin{aligned}
\log L = &\quad \prod_{i=1}^{M} \text{Normal}(m(t_i, \theta) - d_i, \sigma^2)\\
m(t, \theta) = &\quad m_0 + \sum_{j=1}^{n_\text{comp}} A_c \times \sin\left(\frac{2\pi t}{P_j} + \phi_j\right)\\
m_0 \sim &\quad \text{Uniform}(-50, 50)\\
\log \sigma \sim &\quad \text{Uniform}(-1.5, -0.5)\\
\log A \sim &\quad \text{Uniform}(-2, 2)\\
\phi \sim &\quad \text{Uniform}(0, 2\pi)\\
\log P \sim &\quad \text{Uniform}(-1, 3)
\end{aligned}
$$

Figure B.1 presents a graph (t_i, d_i) of $M = 40$ data points, generated by sampling $n_\text{comp} = 2$ components with offset $m_0 = -5$, measurement uncertainty $\sigma = 0.1$, amplitudes $A_1 = 0.25$, $A_2 = 0.05$, periods $P_1 = 180$, $P_2 = 44$ and phases $\phi_1 = 0$, $\phi_2 = 1\,\text{rad}$. The same data are analysed with $n_\text{comp} = 0, 1, 2, 3$, producing two to eleven dimensional problems with completely exchangable components.

Appendix B.2: Low X-ray count observations of a Compton-thick AGN

This mock problem ('xrayspectrum') describes a 5-dimensional degenerate physical parameter space discussed in Buchner et al. (2014) of a heavily obscured active galactic nucleus. The model components and data are described in Figure B.2. The emission spectrum over 200 energy channels from 0.5 to 8 keV is described by:

$$F(E) = B + A_{LE}(E) \times E^{-\Gamma} \times \left[\exp\left(-N_\text{H} \times E^{-3}\right) + f_\text{scat}\right]$$

[16] `https://johannesbuchner.github.io/UltraNest/example-line.html`
[17] `https://johannesbuchner.github.io/UltraNest/example-intrinsic-distribution.html`
[18] `https://github.com/JohannesBuchner/PosteriorStacker`
[19] `https://github.com/JohannesBuchner/PosteriorStacker`

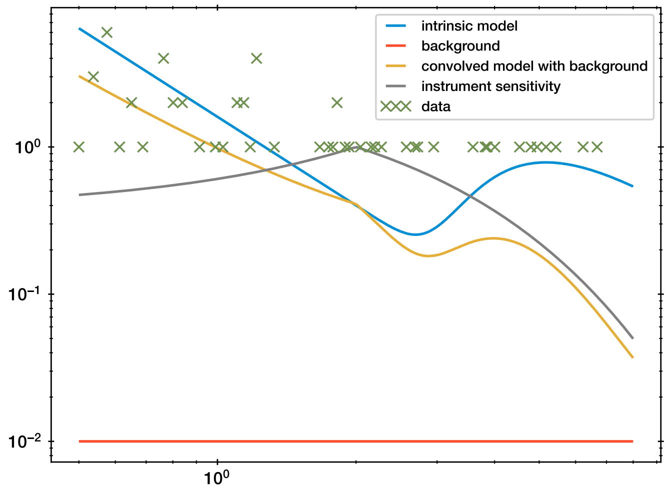

Fig. B.2. X-ray spectrum with the true generating model. The total model (blue curve) is a sum of two powerlaws, one altered by a exponential truncation towards the low-energy side. That model is multiplied by the instrument sensitivity (gray curve), and background contamination (red) is added. The final model is shown in blue, from which poisson data are drawn (green crosses).

with background amplitude B, signal amplitude A, spectral photon index Γ, obscuring column density N_H, and fraction of the powerlaw that escapes unobscured $f_{\rm scat}$. The instrument sensitivity peaks near 2keV: $A_{LE}(E) = \exp\left\{-\left|\frac{E-2\text{keV}}{2\text{keV}}\right|\right\}$.

This model shows a degeneracy when A is low and comparable to B, because high N_H and high $f_{\rm scat}$ look like $N_H = 0$ (a simple powerlaw). The model is illustrated in Figure B.2. The priors on the parameters are:

$$\begin{aligned} \log A \sim & \quad \text{Uniform}(-5, +5) \\ \Gamma \sim & \quad \text{Normal}(2, 0.2^2) \\ \log N_{\rm H} \sim & \quad \text{Uniform}(-3, +3) \\ \log f_{\rm scat} \sim & \quad \text{Uniform}(-7, -1) \\ \ln B \sim & \quad \text{Normal}(0.2, 0.1^2) \end{aligned}$$

Appendix C: Toy problems

In this section, the prior is the unit hypercube ($0 < \theta_i < 1$ for $1 \leq i \leq d$), unless specified otherwise. If analytically known, the marginal parameter posterior distribution $p(\theta_i|D)$ and marginal likelihood Z are given.

Appendix C.1: Asymmetric Gaussian

Integration of a Gaussian distribution is a standard test problem. The variation here introduces some parameter inequality and spreads the means in a sine pattern, so that the posterior is not at the center of the prior range.

$$\begin{aligned} L = & \quad \prod_{i=1}^{d} \text{Normal}(\mu_i, \sigma_i^2) \\ \sigma_i = & \quad 0.1 \times 10^{-\left(-9-\frac{\sqrt{d}}{2}\right)\times\frac{i-1}{d-1}} \\ \mu_i = & \quad \frac{1}{2} + \frac{1-5\sigma_i}{2} \times \sin\frac{i-1}{2d} \end{aligned}$$

This problem is evaluated using uniform priors on $d = 4$ (making σ_i range from 10^{-9} to 10^{-1}), 16 ($10^{-8} < \sigma_i < 10^{-1}$) and 100 dimensions ($10^{-5} < \sigma_i < 10^{-1}$). The four-dimensional case is shown in the top left panel of Figure 1. Note the different axes ranges. The true posterior is $p(\theta_i|D) = \text{Normal}(\mu_i, \sigma_i^2)$, and the marginal likelihood $Z \approx 1$.

Appendix C.2: Beta product

Diverse test problems can be generated by combining standard one-dimensional probability distributions. Here, the Beta distribution is used to represent diverse posterior shapes in each parameter:

$$L = \prod_{i=1}^{d} \text{Beta}(a_i, b_i)$$

with fixed, known a,b, randomly generated as:

$$\begin{aligned} \log a_i \sim & \quad \text{Uniform}(-1, 1) \\ \log b_i \sim & \quad \text{Uniform}(-1, 1) \end{aligned}$$

This distribution can produce multiple modes (where $a_i < 1$ and $b_i < 1$), non-Gaussian tails. The likelihood is relatively uninformative on each parameter (completely non-informative when $a_i = b_i = 0$). We test this problem in 2, 10 and 30 dimensions. The true marginal posteriors are given by $P(\theta_i|D) = \text{Beta}(a_i, b_i)$, and the marginal likelihood is $Z = 1$.

Appendix C.3: Correlated Funnel

Neil's funnel is a standard test problem that represents features of hierarchical Bayesian models. It is a normal distribution with the standard deviation also a free parameter. This causes a funnel shape (see middle panel of Figure 1) involving all parameters. Such non-affine correlations can sometimes be eased significantly by reparametrizations which scale the unknown mean parameters by the standard deviation parameter Betancourt & Girolami (2013). Here we use the correlated version of Karamanis & Beutler (2020):

$$
\begin{aligned}
L &= \prod_i \mathrm{Normal}(\mu_i - \gamma \times \mu_{i-1}, \Sigma^2) \\
\ln\sigma &\sim \mathrm{Normal}(0,1) \\
\mu_i &\sim \mathrm{Uniform}(-100,100) \\
\Sigma_{ij} &= \begin{cases} \sigma & \text{if } i=j \\ \gamma\times\sigma & \text{otherwise} \end{cases}
\end{aligned}
$$

This problem is tested in 2, 10 and 50 dimensions with correlation strength $\gamma = 0.95$. The true marginal posteriors are $p(\mu_i|D) = \mathrm{Normal}(0,1)$, $p(\ln\mu_i|D) = \mathrm{Normal}(0,1)$, and the marginal likelihood $Z \approx 1$.

Appendix C.4: Rosenbrock function

The Rosenbrock function is a standard test problem in optimization. It exhibits a non-linear, narrow degeneracy that can be difficult to navigate (right middle panel of Figure 1). We adopt a probabilistic formulation suggested by Brewer (2016) (see a similar version in **?**):

$$
\begin{aligned}
\log L &= -2 \times \sum_{i=1}^{d-1} 100 \times \left(\theta_{i+1} - x_i^2\right)^2 + (1-\theta_i)^2 \\
\theta_i &\sim \mathrm{Uniform}(-10,10)
\end{aligned}
$$

We test this problem in 2, 20 and 50 dimensions.

Appendix C.5: LogGamma

The LogGamma problem Beaujean & Caldwell (2013) exhibits multi-modality and heavy tails, which lead to non-elliptical contours.

$$
\begin{aligned}
g_a &\sim \mathrm{LogGamma}\left(1, \frac{1}{3}, \frac{1}{30}\right) \\
g_b &\sim \mathrm{LogGamma}\left(1, \frac{2}{3}, \frac{1}{30}\right) \\
n_c &\sim \mathrm{Normal}\left(\frac{1}{3}, \frac{1}{30}\right) \\
n_d &\sim \mathrm{Normal}\left(\frac{2}{3}, \frac{1}{30}\right) \\
d_i &\sim \mathrm{LogGamma}\left(1, \frac{2}{3}, \frac{1}{30}\right) \quad \text{if } 3 \le i \le \frac{d+2}{2} \\
d_i &\sim \mathrm{Normal}\left(\frac{2}{3}, \frac{1}{30}\right) \quad \text{if } \frac{d+2}{2} < i \\
L_1 &= \frac{1}{2}\left(g_a(x_1) + g_b(x_1)\right) \\
L_2 &= \frac{1}{2}\left(n_c(x_2) + n_d(x_2)\right) \\
L &= L_1 \times L_2 \times \prod_{i=3}^{d} d_i(x_i)
\end{aligned}
$$

We test this problem in 2, 10 and 30 dimensions. The true marginal posteriors are given by $P(x_1|D) = L_1(x_1)$, and the marginal likelihood is $Z = 1$.

Appendix C.6: Eggbox

The eggbox function is a two-dimensional extremely multi-modal function (bottom left panel of Figure 1) proposed by Feroz & Hobson (2008), defined as:

$$
\begin{aligned}
\log L &= (2 + \cos(5\pi\cdot\theta_1)\cdot\cos(5\pi\cdot\theta_2))^5 \\
\theta_i &\sim \mathrm{Uniform}(0, 10\pi)
\end{aligned}
$$

Appendix C.7: Box

This is a flat distribution at the corner of the parameter space, placed on top of a wide, unimportant Gaussian distribution:

$$
\ln L = -\frac{1}{2} \times \left(\frac{\theta}{0.1}\right)^2 + 100 \times I[\delta < 0.1]
$$

Here, $\delta = \max_i |\theta_i|$ is the parameter with the largest deviation from zero. The priors are the standard uniform distribution on all parameters θ_i.

We test this problem in $d = 5$ dimensions. The true marginal posteriors are given by $P(\theta_i|D) = \mathrm{Uniform}(0, 0.1)$, and the marginal likelihood is $\ln Z \approx 100 + (0.1)^d$.

Appendix C.8: Spike and slab

The spike and slab problem is a mixture of two gaussians, one with a wide standard deviation, one with a narrow standard deviation. This leads to a strong phase transition. The narrower gaussian has standard deviation $\sigma_2 = f^{-\frac{1}{d}}$, and is shifted by Δ in each direction. The wider gaussian has standard deviation $\sigma_1 = 1$. Then the likelihood is:

$$
\begin{aligned}
L &= L_1 + L_2 \\
L_1 &= \frac{w_1}{1+w_1} \prod_i \frac{1}{2\pi\sigma_1^2} \exp\left\{-\frac{1}{2} \times \left(\frac{\theta_i - \Delta\times\sigma_1}{\sigma_1}\right)^2\right\} \\
L_2 &= \frac{1}{1+w_1} \prod_i \frac{1}{2\pi\sigma_2^2} \exp\left\{-\frac{1}{2} \times \left(\frac{\theta_i}{\sigma_2}\right)^2\right\}
\end{aligned}
$$

We use a two-dimensional setup, with the two Gaussians co-located $\Delta = 0$. For the relative weights we adopt w_1 values of 1, 40, 1000, and for σ_2 adopt 4, 40, 400 or 4000. This gives 16 mono-modal toy problems with phase transitions. Secondly, we adopt offset values for Δ of 1, 2, 4, 10 with weights w_1 of 1, 40 or 1000. This gives another 8 (bimodal) toy problems. The problems are named "spikeslab-w_1-dd-σ_2(-offΔ)". The parameters are uniformly distributed from -10 to +10.

The true evidence is $Z = 20^{-d} \approx e^{-6}$ when σ_2 is small, and the true marginals are the superposition of the two weighted Gaussians.