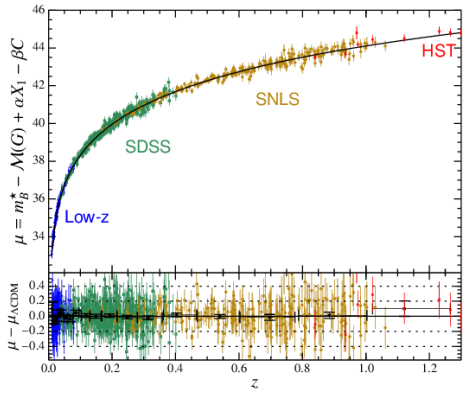
A fast track to Bayesian statistics

Costantino Pacilio

March 13, 2025

1 Motivation



The figure above shows the luminosity distance (y-axis) versus the redshift (x-axis) of type Ia supernovae (credits: [1] and irfu.cea.fr). The colored bars are measurements from individual supernovae (the x errors are subdominant with respect to the y errors), while the black line is the best fit for the relation $y = f(x)$. The problem of fitting the relation between two measured quantities across their respective domains is frequent in science. These (informal!) notes try to offer a first introduction to this problem, from the perspective of Bayesian statistics.

Since these notes are aimed at graduate students to complement a theoretical class on gravitational waves, and since gravitational-wave data analysts make massive usage of Bayesian statistics, understanding the basic concepts of Bayesian statistics is an essential part of the training (even for theoreticians, if they want to understand how their theories are tested and the significance of such tests).

Bayesian statistics is popular in astrophysics, as opposed to frequentist statistics. Unfortunately, undergraduate classes rarely offer sufficient training on this subject. These notes cover the most basic concepts (fitting a model, model comparison, nested models) by focusing on a much simpler problem than gravitational-wave data, namely regression between two observables x and y.

Ref. [2] offers a textbook introduction to (frequentist and Bayesian) statistics for astronomers and astrophysicists (the book is freely available at www.astroml.org. The book also covers machine learning applications and `Python` tutorial.

This is a living document and you are currently reading the first version. Do not use it like a textbook excerpt, but merely as an informal invitation to the subject!

2 Fitting a model

Consider a set of $N_{\rm obs}$ independent observations, $\boldsymbol{d} = \{d_1, \ldots, d_i, \ldots, d_{N_{\rm obs}}\}$. For each observation we measure two variables x and y, with the aim of reconstructing the functional dependence of y from x

$$y \stackrel{?}{=} f(x, \boldsymbol{\lambda}) \tag{1}$$

where f is the model and $\boldsymbol{\lambda}$ denotes the set of the model hyperparameters. While we, as observes, ignore the underlying truth, assume that the latter is given by a linear relation

$$y_i = \alpha_{\rm true} x_i + \beta_{\rm true}\,. \tag{2}$$

with $\alpha_{\rm true} = 2$ and $\beta_{\rm true} = 0.5$. *Given a model* f, how do we estimate $\boldsymbol{\lambda}$?

First, we need to know the likelihood of the data with respect to the observed quantities x and y. For the generic observation d_i, we assume that x is measured with infinite precision, while only the first two momenta of y are measured, namely its mean $y_{i,\rm obs}$ and standard deviation σ_y. For simplicity, we also assume that σ_y is the same for all observations. Therefore, the likelihood is

$$p(d_i|x, y) = \delta(x - x_{i,\rm obs})\, \mathcal{N}(y|y_{i,\rm obs}, \sigma_y)\,. \tag{3}$$

Since the observations are independent, the total likelihood is the product of the individual ones

$$p(\boldsymbol{d}|x, y) = \prod_{i=1}^{N_{\rm obs}} p(d_i|x, y)\ . \tag{4}$$

Now, we apply Bayes' theorem to derive an expression for the posterior probability density of $\boldsymbol{\lambda}$:

$$p(\boldsymbol{\lambda}|\boldsymbol{d}) = \frac{p(\boldsymbol{d}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})}{p(\boldsymbol{d})}\,. \tag{5}$$

The term $p(\boldsymbol{\lambda})$ expresses the prior belief. In the absence of a well-motivated prior, we will opt for a uniform prior. Note, however, that a uniform prior is not an *uninformative* prior and, strictly speaking, no prior is uninformative. Note also that a uniform prior cannot simply be dropped from the expression as an irrelevant proportionality factor, because *in practice* a uniform prior is uniform within a bounded region (usually but not necessarily a rectangular

box); therefore the boundary limits the allowed values of $\boldsymbol{\lambda}$. This also shows that a uniform prior can be informative, for example, by restricting the allowed deviations from general relativity to be within 10% around the GR value, or by restricting the tidal deformability in the range $[0, 5000]$; these values might be informed by prior experiments or by theoretical considerations, for example, the lack of beyond-GR theories predicting large deviations or the lack of theoretically viable neutron-star model predicting large tidal deformabilities.

The term $p(\boldsymbol{d}|\boldsymbol{\lambda})$ is the hyper-likelihood. We can expand it as

$$\begin{aligned} p(\boldsymbol{d}|\boldsymbol{\lambda}) &= \prod_{i=1}^{N_{\rm obs}} p(d_i|\boldsymbol{\lambda}) = \prod_i \int dx \int dy\ p(d_i|x,y)\ p(x,y|\boldsymbol{\lambda}) \\ &= \prod_i \mathcal{N}(f(x_{i,\rm obs},\boldsymbol{\lambda})|y_{i,\rm obs},\sigma_y)\ p(x_{i,\rm obs})\,. \end{aligned} \tag{6}$$

The term $p(x)$ is subtle. For the moment, we neglect it (equivalently, we assume a wide uniform prior over x).

The term $p(\boldsymbol{d})$ is called the "evidence" of the data, and it is explicitly given by

$$p(\boldsymbol{d}) = \int d\boldsymbol{\lambda}\ p(\boldsymbol{d}|\boldsymbol{\lambda})\ p(\boldsymbol{\lambda})\,. \tag{7}$$

Since it is independent of $\boldsymbol{\lambda}$, we can assume that it is a normalization factor when using equation (5) to sample $\boldsymbol{\lambda}$.

Now, let us specify a model f. We start from a linear model

$$y = \alpha x + \beta\,. \tag{8}$$

We use nested sampling and equation (5) to sample from $p(\alpha,\beta|\boldsymbol{d})$, assuming $\sigma_y = 0.25$ and uniform priors

$$p(\alpha,\beta) = U(\alpha|0,5)\ U(\beta|0,5)\,. \tag{9}$$

Note that, under our assumptions, and since the uniform prior is wide when compared to the posterior support, the problem is analytically tractable and reduces to linear least squares regression. The maximum a posteriori (MAP) estimator is obtained by maximizing the likelihood

$$\frac{\partial\mathcal{L}}{\partial\alpha} \overset{\rm MAP}{=} 0, \quad \frac{\partial\mathcal{L}}{\partial\beta} \overset{\rm MAP}{=} 0 \tag{10}$$

while the covariance matrix is obtained from the inverse of the Fisher matrix

$$\Sigma = \Gamma^{-1}, \quad \Gamma_{ab} = -\left.\frac{\partial^2\mathcal{L}}{\partial\theta^a\partial\theta^b}\right|_{\rm MAP}, \quad \theta^{a,b} \in \{\alpha,\beta\}\,. \tag{11}$$

We obtain the results shown in figure 1. You see that the posterior does not peak at the true of the hyper-parameters. This is simply due to the noise scattering induced by our noise model, and it is the general behavior (which is,

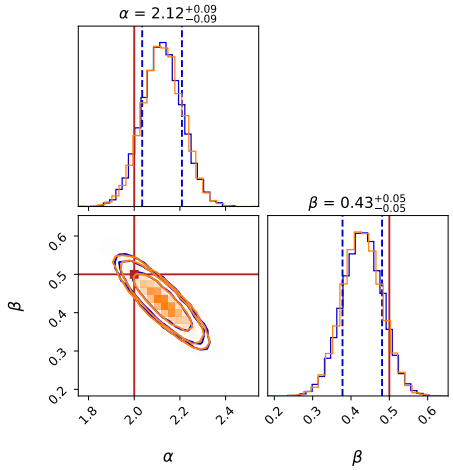


Figure 1: *Blue:* Corner plot for the hyper-parameters $\{\alpha, \beta\}$ defined in equation (8), and sampled with nested sampling from the box-uniform prior (9). The blue dashed lines and the plot titles indicate the 68% credible intervals, while the two-dimensional contours indicate the 68%, 90% and 95% credible regions of the joint posterior. The red lines correspond to the true injected values $\{\alpha_{\rm true}, \beta_{\rm true}\}$. *Orange:* the closed form expression obtained by maximizing the log-likelihood with respect to α and β.

of course, the reason why we quote credible intervals and not just median/MAP estimators).

Sometimes, we are interested in performing a so-called "injection campaign". For example, we want to test whether a sampling pipeline works correctly. A necessary (but not sufficient!) condition for a pipeline to work correctly is expressed as a quantile-quantile plot, sometimes also called a p-p plot. A p-p plot is a 2-dimensional plot with quantiles p from 0 to 1 on the x-axis, while on the y-axis you find the cumulative ratio of injections such that the true injected value lies within the p-th quantile. A pipeline works correctly if the p-p plot is consistent with a diagonal line, within its Bernoulli uncertainty band. It should be noted that this condition is only guaranteed to hold if the injection parameters are sampled from the same prior distribution used for individual posterior recoveries. You see that, in order to check this necessary condition for a pipeline to work, it is essential to retain the statistical biases induced by the noise realizations.

Some other times, we are interested in the "typical" posterior recovery of our pipeline and data. The notion of "typical" is blurred, but intuitively we can conjecture that in half of the experiment realizations the bias will overestimate the hyper-parameters, while in the other half it will underestimate them. Therefore, "on average" we expect that the pipeline will return posteriors centered on the true values of the hyper-parameters. This intuition is formalized

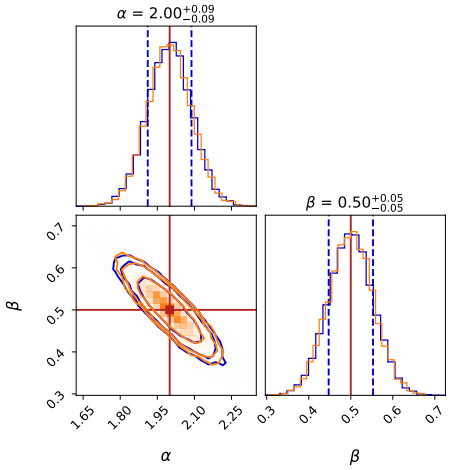


Figure 2: Similar to figure 1, but for a zero-noise realizations.

by the notion of zero-noise realization of the noise [5, 4]. Although this may seem contradictory, keep in mind that a noise realization is a statistical draw, namely, for each observation d_i, the noise is sampled independently from a normal distribution centered in 0 with standard deviation σ_y. Therefore, we are allowed to draw a noise realization $\boldsymbol{n} = \{0, \ldots, 0\}$. In a "macroscopic" sense this realization of the noise is unlikely, for the same reason why drawing ordered cards from a shuffled deck is "macroscopically" unlikely. But from the "microscopic" point of view, all configurations are equally probable, and we ascribe unlikelyness to those configurations that satisfy a macroscopic condition of low entropy. It can be shown that a zero-noise realization is equivalent to considering individual likelihoods $p(d_i|x, y)$ that are averaged *geometrically* over infinite noise realizations — equivalently, to consider the ordinary average of the individual log-likelihoods $\log p(d_i|x, y)$ over infinite noise realizations (see [3] for a proof). If we rerun the experiment with a zero-noise realization, we find that the MAP estimator is centered in the true values of the hyper-parameters — see figure 2.

NOTE: the MAP of the marginalized 1-d posterior is not necessarily coincident with the MAP of the posterior as a whole! In this case, they are the same because the posterior is a multivariate normal with an excellent approximation.

Until now, we have assumed that the noise level σ_y is known prior to estimation of the parameters. This is a very common assumption and, actually, it is often necessary to obtain an estimate of the noise before estimating the hyper-parameters of the model. For example, in gravitational-wave data-analysis, you observe a time series $\{s(t_1), \ldots, s(t_i), \ldots, s(t_{N_{\rm obs}})\}$, where the signal at each time stamp is the sum of the gravitational waveform h and a noise realization n, $\boldsymbol{s} = \boldsymbol{h} + \boldsymbol{n}$. Since you observe only $\boldsymbol{s}$, and not $\boldsymbol{h}$ and $\boldsymbol{n}$ individually, you

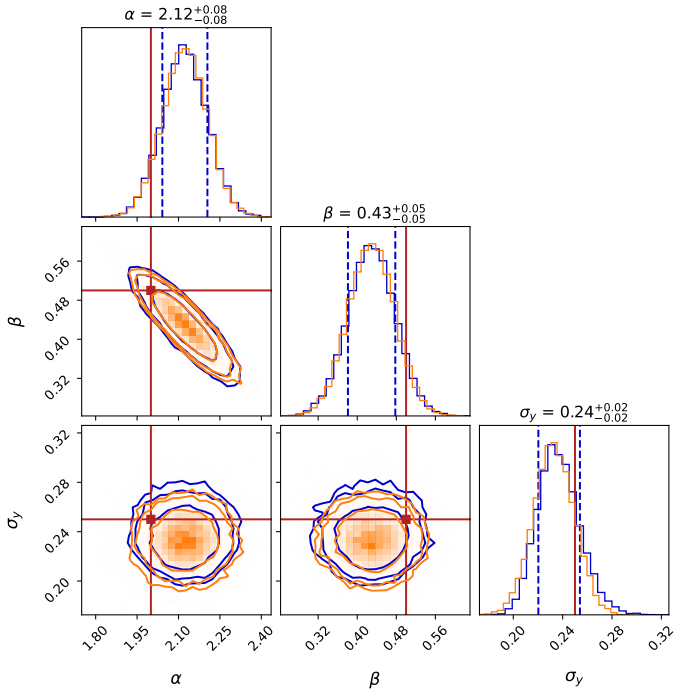


Figure 3: Similar to figure 1, but we estimate also the noise level σ_y as a free parameter.

cannot really estimate the noise from the data to analyze. What is usually done in practice is to consider data segments immediately before and after the signal and use them to estimate the statistical properties of the noise. This strategy works because the duration and frequency of the signals that stand out above the detector noise is such that there are noise-only data segments between consecutive signals. However, this strategy will not work in general: for example, the next-generation space-based gravitational-wave detector LISA will receive a continuous stream of signals, therefore the noise must be estimated together with the model parameters. This requires the development of cutting-edge techniques and is part of the so-called "LISA Global Fit". In our simple toy problem, we can easily extend the likelihood to infer the noise level σ_y alongside α and β.

If we repeat the above experiment, but also estimate the noise level σ_y with a uniform prior in the range $[0, 1]$, we obtain the corner plot in figure 3. The orange contours represent the corresponding approximation, obtained by maximizing the likelihood. Note that, also in this case, a closed form solution exists: the posterior for σ is uncorrelated to α and β, and it is approximated by a normal distribution $p(\sigma|\boldsymbol{d}) \approx \mathcal{N}(\sigma|\mathrm{RMSE}, \mathrm{RMSE}/\sqrt{2N_{\mathrm{obs}}})$, where RMSE

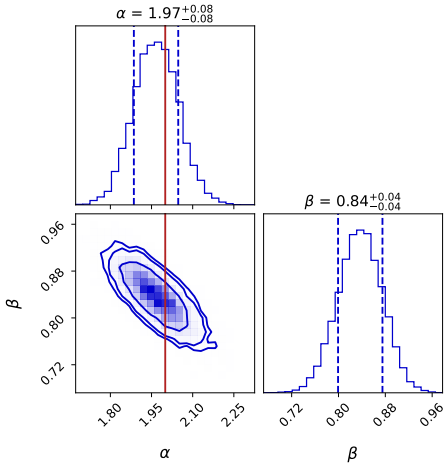


Figure 4: Similar to figure 1, but for the quadratic model H_1 in equation (14).

is the root mean squared error

$$\mathrm{RMSE} = \sqrt{\sum_i \frac{(y_{i,\mathrm{obs}} - \alpha_* x_{i,\mathrm{obs}} - \beta_*)^2}{N_{\mathrm{obs}}}}\,. \tag{12}$$

When we estimate the noise level, we are not allowed to perform zero-noise experiments. Why?

3 Model comparison

In the previous section, we assumed a linear model f with the same functional form as the true underlying model. Let us call it the hypothesis H_0

$$H_0 : y = \alpha x + \beta\,. \tag{13}$$

Suppose that we instead assume a quadratic dependence of the form

$$H_1 : y = \alpha x^2 + \beta\,. \tag{14}$$

What is the result of the posterior recovery? From figure 4, we see that the sampler has still converged to a well-defined posterior density $p(\alpha, \beta|\boldsymbol{d})$, although different from the linear model H_0. We ask: which is the best model?

One might be tempted to answer this question by comparing the predictions of each posterior density with the data. To do this, we construct the posterior functional distribution

$$p(y|x, \boldsymbol{d}, H_0) = \int d\alpha \int d\beta\; p(y|x, \alpha, \beta)\; p(\alpha, \beta|\boldsymbol{d}, H_0) \equiv \mathbb{E}\left[p(y|x, \alpha, \beta)\right] \tag{15}$$

where the expectation value is taken over the samples from $p(\alpha, \beta|\boldsymbol{d}, H_0)$. An equivalent expression holds for H_1. In practice, equation (15) is equivalent to construct the set

$$\{y_{\text{pred}}\} = \{f(x, \boldsymbol{\lambda}) \quad \forall \boldsymbol{\lambda} \sim p(\boldsymbol{\lambda}|\boldsymbol{d}, H_0)\} \tag{16}$$

and sampling from it with repetitions. The results are shown in figure 5.

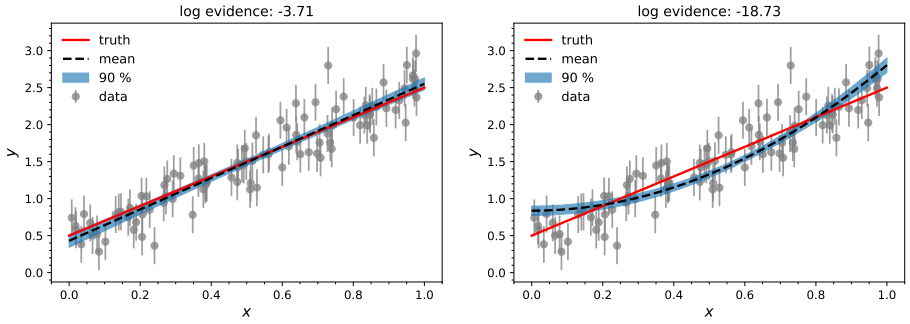


Figure 5: Posterior functional distributions for the linear model H_0 (left) and the quadratic model H_1 (right). The titles show the respective log-evidences obtained from the nested sampler.

From the comparison between the data and the predictions, both models seem to be compatible with (namely, not excluded by) the data. Therefore, we need a more quantitative comparison. This is provided to us by the so-called Bayes factor between H_0 and H_1. The relative probability of H_0 to H_1 can be expanded as

$$\frac{p(H_0|\boldsymbol{d})}{p(H_1|\boldsymbol{d})} = \frac{p(\boldsymbol{d}|H_0)}{p(\boldsymbol{d}|H_1)} \frac{p(H_0)}{p(H_1)} \,. \tag{17}$$

The last factor on the right-hand side expresses the ratio of our prior belief to H_0 and H_1. This must be assessed on a case-by-case basis and there is no general argument to determine it. The second term is the ratio of the evidences and it is called the Bayes factor between H_0 and H_1

$$B_{01} = \frac{p(\boldsymbol{d}|H_0)}{p(\boldsymbol{d}|H_1)} \,. \tag{18}$$

We see that the log Bayes factor between H_0 and H_1 is ≈ 15, meaning that H_0 is approximately $e^{15} \approx 3 \times 10^6$ more likely than H_1!

Note that the Bayes factor alone is not enough to confute a model with respect to another. This is because we might have prior beliefs such that $p(H_0)/p(H_1) \approx 0$ and still prefer H_1 over H_0. For example, observations of tidal effects within gravitational waves are not only compatible with neutron stars, but also with exotic forms of stars called "boson stars"; moreover, sometimes a gravitational wave can be better fitted (in the sense of the Bayes factor)

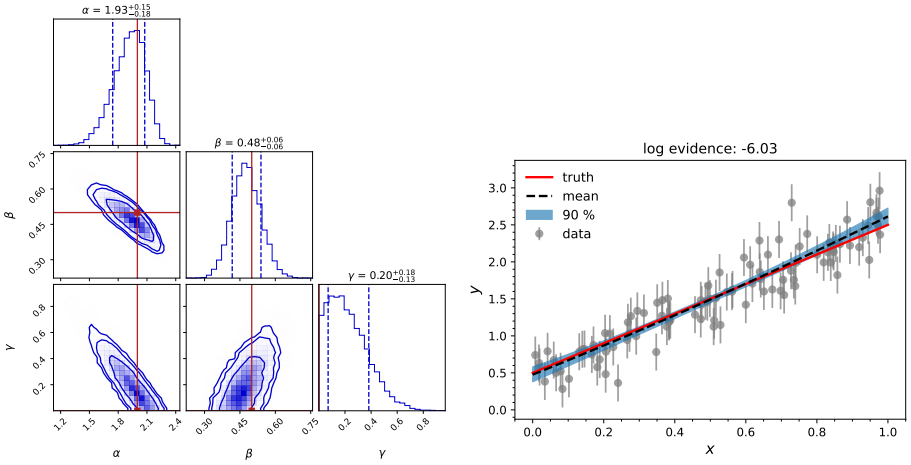


Figure 6: Corner plot (left) and posterior functional distribution (right) for the model H_2 in equation (19).

by modeling the source as a head-on collisions of boson stars (say H_0), rather than a circular inspiral of two neutron stars (say H_1). However, it is fair to state that the astrophysical evidence for the existence of neutron stars orbiting in binaries and the theoretical foundations for such systems are so strong that we are very strongly led to consider H_1 as the only realistic hypothesis on the table. Naturally, the situation might change in the future with more theoretical advancements and/or novel astrophysical observations. Therefore, it is always useful to compute Bayes factors, as they do not depend on the prior but only on the data, and their interpretation can change in time.

4 Nested models

Let us refine our model comparison example by proposing a more viable alternative hypothesis

$$H_2 : y = \alpha x + \beta + \gamma x^2 \,. \tag{19}$$

We see that H_0 can be recovered from H_2 by slicing it at $\gamma = 0$. When this happens, we will say that H_0 is nested in H_2. We sample the posterior density by imposing the additional uniform prior on γ

$$p(\gamma|H_2) = U(\gamma|0, 5) \,. \tag{20}$$

The corresponding corner plot and posterior plot is displayed in figure 6. As expected, the posterior for γ rails against the lower bound of the prior. (By the way, notice that in this case the credible quantiles quoted in the corner plot for γ are meaningless, and you should quote only an upper bound). We see that H_0 and a log Bayes factor of ≈ 2.32 compared to H_2; however, their posterior

predictions are very similar to each other. The reason is that the model H_2 is more complex, because it contains an additional parameter, but its explanatory power is the same. According to Occam's razor, given the same explanatory power, the simpler model shall be preferred. We see that Bayes factors take Occam's razor into account automatically.

In the case of nested models, we can compute the Bayes factor using the Savage-Dickey ratio (see statproofbook.github.io for a proof)

$$B_{02} = \frac{p(\gamma = 0|\boldsymbol{d}, H_2)}{p(\gamma = 0|H_2)} \,. \tag{21}$$

With this method, you do not need a parameter estimation under H_0, because the posterior sample under H_2 is sufficient. In practice, you need to perform a kernel density estimate (KDE) of $p(\gamma|\boldsymbol{d}, H_2)$ and evaluate it at zero. This presents a difficulty when the posterior does not fall off at zero at the boundaries: the reason is that the KDE induces density leakage outside the boundaries, when the posterior rails towards them. There is no unique way of dealing with this issue, but a popular working way is to correct the original KDE using reflective boundary conditions

$$p_{\mathrm{KDE,corr}}(x) = p_{\mathrm{KDE}}(x) + p_{\mathrm{KDE}}(2x_{\mathrm{bound}} - x) \,. \tag{22}$$

With this corrected KDE and using the Savage-Dickey ratio (21), we estimate a log Bayes factor $B_{02} \approx 2.38$, in line with the one obtained using the log evidences provided by the nested sampling.

5 Population distribution of the independent variable

When we derive the likelihood (6), we neglected an additional term $p(x)$. In fact, this term expresses the prior imposed on the population distribution of x. In this section, we are going to reinstate the term in the likelihood and explore its effect.

In all previous examples we have assumed that the observations were uniformly distributed in x (see, for example, figure 5). Now, we assume that x is observed non-uniformly within its domain, but rather that observations follow the following normal distribution

$$x \sim \mathcal{N}(x|\mu_{x,\mathrm{true}}, \sigma_{x,\mathrm{true}}) \tag{23}$$

with $\mu_{x,\mathrm{true}} = 0.5$ and $\sigma_{x,\mathrm{true}} = 0.1$.

However, in a real experiment, you ignore the true underlying distribution $p(x)$, suggesting that the correct procedure is to estimate it concurrently with the model variables $\boldsymbol{\lambda}$. Therefore, let us introduce two additional hyperparameters $\boldsymbol{\zeta} = \mu_x, \sigma_x$ and parametrize the population of x by a normal distribution

$$p(x|\boldsymbol{\zeta}) = \mathcal{N}(x|\mu_x, \sigma_x) \,. \tag{24}$$

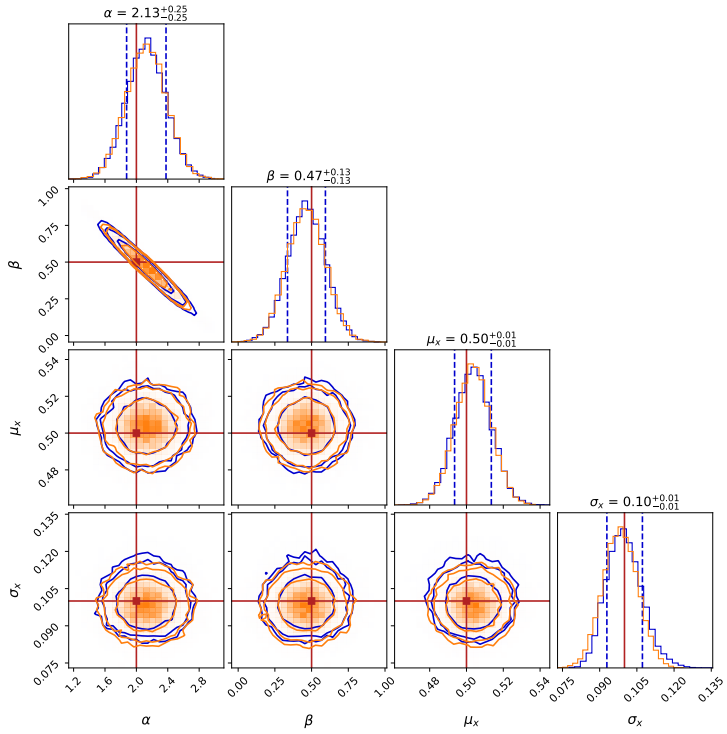


Figure 7: Joint recovery of the model parameters $\boldsymbol{\lambda}$ (see equation (13)) and of the population parameters $\boldsymbol{\zeta}$ (see equation (24)). The blue (resp. orange) contour represent samples from a nested sampler (resp. from a closed-form Fisher expansion around the maximum likelihood estimator).

You can easily show that the likelihood (6) can be generalized to include these additional hyper-parameters, finding

$$p(\boldsymbol{d}|\boldsymbol{\lambda}, \boldsymbol{\zeta}) = \prod_i \mathcal{N}(f(x_{i,\mathrm{obs}}, \boldsymbol{\lambda})|y_{i,\mathrm{obs}}, \sigma_y)\ p(x_{i,\mathrm{obs}}|\boldsymbol{\zeta})\,. \tag{25}$$

We recover the joint posterior on $\{\boldsymbol{\lambda}, \boldsymbol{\zeta}\}$ with box-uniform priors. The result is shown in figure 7. One striking feature is that the population parameters $\boldsymbol{\zeta}$ are uncorrelated with the model parameters $\boldsymbol{\lambda}$;in other words, we could have equivalently estimated $p(\boldsymbol{\zeta}|\boldsymbol{d})$ and $p(\boldsymbol{\lambda}|\boldsymbol{d})$ independently from each other. This factorization is also apparent within the likelihood (25). However, this arises from neglecting uncertainties on x: in the general case, the two estimates cannot

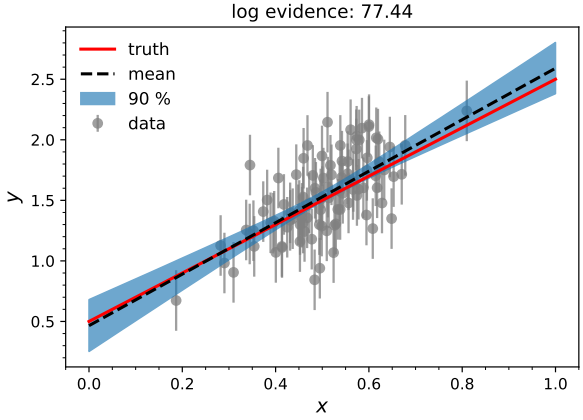


Figure 8: Predictive distribution of $y = f(x)$, when observations of x are not distributed uniformly within the domain.

be disentangled:

$$\begin{aligned} p(\boldsymbol{d}|\boldsymbol{\lambda},\boldsymbol{\zeta}) &= \prod_i \int dx \int dy\; p(d_i|x,y)\; p(y|x,\boldsymbol{\lambda})\; p(x|\boldsymbol{\zeta}) \\ &= \prod_i \int dx\; p(d_i|x, y=f(x,\boldsymbol{\lambda}))\; p(x|\boldsymbol{\zeta})\,. \end{aligned} \tag{26}$$

Note also that, when imposing the ansatz (24), we guessed the correct shape of $p(x)$. However, in a real-life experiment we don't have previous knowledge of the population distribution, which will inevitably lead to biased estimates of $\boldsymbol{\zeta}$, similarly to what we observed when using H_1 instead of H_0 for $\boldsymbol{\lambda}$. This problem arises in gravitational-wave population inference, where educated guesses are made. One strategy to mitigate this problem is to propose several possible priors, and then select the one that gives the largest evidence. Alternatively, so-called nonparametric methods are emerging, which use flexible representations capable of expressing a large family of probability distributions.

Figure 8 shows the posterior functional relation between y and x. We see that, the uncertainty on the predictions tends to spread away from the observed data, showing that the Bayesian recovery of the parameters takes automatically into account the lack of observations in those regions of the parameter space.

References

[1] M. Betoule et al. "Improved cosmological constraints from a joint analysis of the SDSS-II and SNLS supernova samples". In: *Astron. Astrophys.* 568 (2014), A22. DOI: 10.1051/0004-6361/201423413. arXiv: 1401.4064 [astro-ph.CO].

[2] Željko Ivezić et al. *Statistics, data mining, and machine learning in astronomy: a practical Python guide for the analysis of survey data*. Vol. 8. Princeton University Press, 2020.

[3] Samaya Nissanke et al. "Exploring short gamma-ray bursts as gravitational-wave standard sirens". In: *Astrophys. J.* 725 (2010), pp. 496–514. DOI: 10.1088/0004-637X/725/1/496. arXiv: 0904.1017 [astro-ph.CO].

[4] Carl L. Rodriguez et al. "Basic Parameter Estimation of Binary Neutron Star Systems by the Advanced LIGO/Virgo Network". In: *Astrophys. J.* 784 (2014), p. 119. DOI: 10.1088/0004-637X/784/2/119. arXiv: 1309.3273 [astro-ph.HE].

[5] Leslie Wade et al. "Systematic and statistical errors in a bayesian approach to the estimation of the neutron-star equation of state using advanced gravitational wave detectors". In: *Phys. Rev. D* 89.10 (2014), p. 103012. DOI: 10.1103/PhysRevD.89.103012. arXiv: 1402.5156 [gr-qc].