Comparing Bayes factors and hierarchical inference for testing general relativity with gravitational waves

Maximiliano Isi, Will M. Farr, and Katerina Chatziioannou

Phys. Rev. D **106**, 024048 — Published 26 July 2022

DOI: 10.1103/PhysRevD.106.024048

Comparing Bayes factors and hierarchical inference for testing general relativity with gravitational waves

Maximiliano Isi,[1,*] Will M. Farr,[1,2,†] and Katerina Chatziioannou[3,4,‡]

[1] *Center for Computational Astrophysics, Flatiron Institute, 162 5th Ave, New York, NY 10010*
[2] *Department of Physics and Astronomy, Stony Brook University, Stony Brook NY 11794, USA*
[3] *Department of Physics, California Institute of Technology, Pasadena, California 91125, USA*
[4] *LIGO Laboratory, California Institute of Technology, Pasadena, California 91125, USA*

(Dated: July 14, 2022)

In the context of testing general relativity with gravitational waves, constraints obtained with multiple events are typically combined either through a hierarchical formalism or though a combined multiplicative Bayes factor. We show that the well-known dependence of Bayes factors on the analysis priors in regions of the parameter space without likelihood support can lead to strong confidence in favor of incorrect conclusions when one employs the multiplicative Bayes factor. Bayes factors $\mathcal{O}(1)$ are ambivalent as they depend sensitively on the analysis priors, which are rarely set in a principled way; additionally, combined Bayes factors $> \mathcal{O}(10^3)$ can be obtained in favor of the incorrect conclusion depending on the analysis priors when many $\mathcal{O}(1)$ Bayes factors are multiplied, and specifically when the priors are much wider than the underlying population. The hierarchical analysis that instead infers the ensemble distribution of the individual beyond-general-relativity constraints does not suffer from this problem, and generically converges to favor the correct conclusion. Rather than a naive multiplication, a more reliable Bayes factor can be computed from the hierarchical analysis. We present a number of toy models showing that the practice of multiplying Bayes Factors can lead to incorrect conclusions.

I. INTRODUCTION

With an increasing number of LIGO-Virgo [1, 2] gravitational wave (GW) observations, we can leverage the collective set of measurements to study the properties of the astrophysical objects that generate GWs [3, 4] and the validity of general relativity (GR) [5–7]. Two broad and complementary approaches exist for drawing inferences from sets of detections. The first relies on the posterior distribution for some model and its continuous parameters whose range of possible values encapsulates the different physics we would like to study. The second phrases a question of interest in the language of model selection between two discrete hypotheses to compute Bayes factors (BFs). The latter approach is common in the context of testing GR where one introduces a parametrized deviation of the signal as predicted by GR [8–13], but further examples relate to black hole mimickers [14, 15], higher-order modes of the radiation [16], GW memory [17], the neutron star equation of state [18–23], gravitational lensing [24, 25], the association between GWs and potential electromagnetic counterparts [26], GW ringdowns [27], and the signal detection problem in general [28–34].

Although posteriors and BFs are mathematically related, in practice approaches focusing on one or the other can come to seemingly different, and even contradictory, conclusions. For example, Ref. [35] considers the case of testing the no-hair theorem with GW ringdowns and shows that BFs can favor the incorrect conclusion even in cases where the posterior has minimal support for it. The alternative of working directly with the posterior and hierachical inference has been introduced in the context of tests of GR [5–7] after the limitations of using BFs to combine information from multiple events were pointed out in [36]. Specifically, multipyling BFs corresponds to assuming that a GR deviation will manifest independently and distributed according to the underlying prior in each observation [36]. In this paper, we further this line of argument to highlight issues with the use of BFs in the context of nested models and show that the hierarchical modeling of population distributions offers a more flexible and reliable alternative.

BFs, or marginalized-likelihood ratios, provide a succinct way to compare the likelihood of two models in light of some data. The BF comparing a hypothesis $\mathcal{H}_0$ to another $\mathcal{H}_1$ for some data d is given by

$$\mathcal{B}^0_1 \equiv \frac{P(d \mid \mathcal{H}_0)}{P(d \mid \mathcal{H}_1)} = \frac{\int p(d \mid \theta_0, \mathcal{H}_0)\, p(\theta_0|\mathcal{H}_0)\, \mathrm{d}\theta_0}{\int p(d \mid \theta_1, \mathcal{H}_1)\, p(\theta_1|\mathcal{H}_1)\, \mathrm{d}\theta_1}\,, \tag{1}$$

where the likelihoods are marginalized over some (potentially different) sets of parameters $\theta_{0/1}$ for the $\mathcal{H}_{0/1}$ hypotheses respectively. The definition of the hypotheses encompasses the choice of parameter priors $p(\theta \mid \mathcal{H})$, as well as any other assumptions built into the functional form of the likelihoods $p(d \mid \theta, \mathcal{H})$. A larger value of $\mathcal{B}^0_1$ (or, equivalently, its natural logarithm, $\ln \mathcal{B}^0_1$) indicates a preference for $\mathcal{H}_0$ over $\mathcal{H}_1$. When enhanced with prior weights for each hypothesis, $P(\mathcal{H}_{0/1})$, this returns the betting odds in favor of one model over the other, conditional on the observed data—namely,

$$\mathcal{O}^0_1 \equiv \frac{P(\mathcal{H}_0)P(d \mid \mathcal{H}_0)}{P(\mathcal{H}_1)P(d \mid \mathcal{H}_1)} = \frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)}\mathcal{B}^0_1\,. \tag{2}$$

* misi@flatironinstitute.org
† will.farr@stonybrook.edu
‡ kchatziioannou@caltech.edu

To avoid expressing an *a priori* preference for either model, it is common to set $P(\mathcal{H}_0) = P(\mathcal{H}_1)$, and so $\mathcal{O}^0_1 = \mathcal{B}^0_1$.

BFs reduce the complicated problem of selecting between two models of reality to a single number—a feature which lies at the core of its appeal, but also of its shortcomings. Like other scalar statistics, interpreting this one number is usually far from straightforward, leading to somewhat *ad hoc* scales such as [37]. This difficulty is worsened by the fact that BFs necessarily are affected by *all* aspects of a model, including those corresponding to less interesting regions of the parameter space like regions of the prior space for which the likelihood offers no support. Consequently, BFs can vary wildly with different choices of prior bounds, which are often set arbitrarily. Since priors can rarely be set from first principles, calibrating BFs tends to require large scale injection campaigns [31]—although this is only possible when the injections themselves can be designed in a principled way (i.e., when we actually know how to simulate expected astrophysical distributions). Even when the model is specified correctly and we can take BFs at face value, the result offers no insight as to why exactly one model is to be preferred.

All these drawbacks compound when one attempts to combine multiple observations by multiplying BFs computed with a fixed prior, which enhances the sensitivity to prior choices. Moreover, naive BF computations from collections of events impose strong, generally unrealistic assumptions [36]. Multiplying BFs obtained from individual events results in a collective BF that assumes the targeted effect (say, a deviation from GR) manifests independently for each observation. On the other hand, generating a collective BF from the product of likelihoods from multiple measurements presumes that the effect manifests identically for all observations. Neither of these are valid assumptions in general [36]. In general the degree to which a targeted effect appears independently versus identically in each observation is something that should be leared from the data; this insight leads directly to hierarchical modeling [38–41].

Rather than assuming a fixed and known distribution (e.g., all events are the same, or all the events are different), a hierarchical analysis works by inferring the underlying distribution of the parameter whose values encode the targeted hypotheses. For example, in the context of tests of GR, a parameter x may represent the magnitude of a GR violation, so that the GR (non-GR) hypothesis implies $x = 0$ ($x \neq 0$); likewise, when searching for GW memory, this could be the amplitude of the effect in question. Once this parameter is identified, we can use hierarchical inference to characterize its values across the observed events—using the collection of measurements holistically to infer the distribution of x. The challenge here lies in choosing a suitable parametrization for this underlying distribution.

We can circumvent this through a moment expansion: as a first approximation, we will only be interested in recovering the mean and standard deviation of the unknown distribution, so we can parametrize it as a Gaussian, whose mean and standard deviation (μ, $\sigma_{\rm pop}$) we are to measure from the data [5]; the null hypothesis will often be constructed so that it is recovered for $\mu = \sigma_{\rm pop} = 0$. This approach allows us to study the population of measurements without assuming the targeted effect manifests either identically or distinctly for all events, learning more about the population distribution with each observation. In this way, the population model now plays the role of a prior in the analysis of any individual event.

In this paper, we compare the BF and hierarchical approaches directly in the context of multiple GW observations, and argue that BFs are unreliable in any context in which the prior does not adapt to the observations at hand. We show that in such context, BFs do not have the right scaling with the number of events, even in simple situations, due to their inherently strong dependence on the (fixed) priors in regions of no likelihood support. We show that hierarchical posteriors do not suffer from such limitations as they rely on "priors" that are inferred from the data, and have a weaker dependence on hyperparameter assumptions.

We begin in Sec. II by reviewing some basic properties of BFs obtained from single observations and their interplay with Occam penalties. Then, in Sec. III, we study the scaling of BFs when combining multiple observations under two example distributions for the true parameters under consideration; we show that this approach can often lead to the incorrect conclusion. We also consider the same scenarios under the hierarchical approach, showing that it does not suffer from the same drawbacks. In Sec. IV, we consider a final model that cannot be obtained as a special case of the distribution assumed by the hierarchical analysis, showing that the hierarchical approach yields the correct result even then. We conclude in Sec. V.

II. SINGLE EVENT: OCCAM PENALTY VS GOODNESS-OF-FIT

Before tackling the problem of multiple observations, we first review the behavior of BFs computed from a single event. A typical situation that is simple to understand analytically is that of nested models, i.e., two models constructed so that one can be recovered as a special case of the other. Parametrized tests of GR, for example, typically involve nested hypotheses wherein the non-GR model is characterized by all the usual GR signal parameters plus one or more additional variables $x_{\rm nGR}$ that quantify the deviation from GR [42–50]. These parametrizations are usually constructed so that GR is recovered when the deviation parameters vanish. Then, to determine whether GR is favored, the data are analyzed with some broad (typically flat) prior on the deviations in order to compute BFs comparing $x_{\rm nGR} = 0$ to $x_{\rm nGR} \neq 0$.

In that spirit, consider some real-valued parameter x and two related hypotheses $\mathcal{H}_{x=0}$ and $\mathcal{H}_{x\neq 0}$, respectively defined to imply $x = 0$ and $x \neq 0$, with some prior over

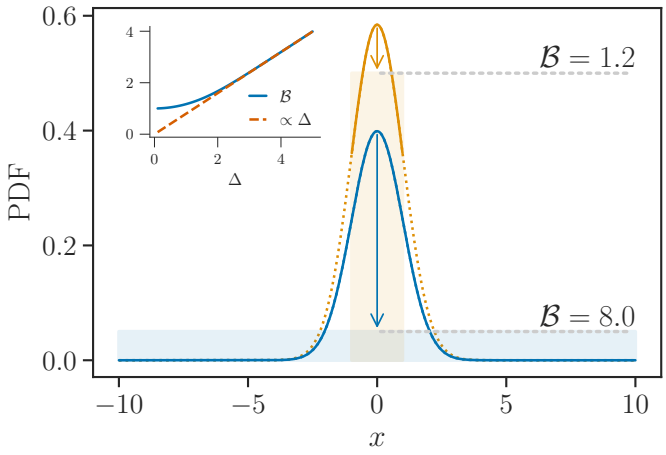

FIG. 1. Bayes factor computation as a Savage-Dickey density ratio. The BF ($\mathcal{B}$) comparing a model in which some parameter x takes the value $x = 0$ to one in which $x \neq 0$ is given by the ratio of the posterior to the prior evaluated at $x = 0$. The main panel shows the case of a normal posterior truncated by a broad prior (half-width $\Delta = 10$, in blue) and a narrow prior ($\Delta = 1$, in orange). The inset shows the scaling of $\mathcal{B}$ as a function of Δ (solid curve); for broad priors, $\mathcal{B} \propto \Delta$ (dashed line). In this toy example, we assume the posterior peaks always at $x = 0$.

x and any other relevant parameters. Since $x = 0$ is a special case of the model in which x is allowed to vary over a broad range including the origin, we say the hypotheses are nested.[1] In that case, the BF comparing the two is given exactly by the Savage-Dickey ratio [51, 52],

$$\mathcal{B}^{x=0}_{x\neq 0} = \frac{p(x=0 \mid d)}{p(x=0)}\,, \tag{3}$$

where $p(x = 0 \mid d)$ is the marginal posterior and $p(x = 0)$ is the prior, both evaluated at $x = 0$. In other words, the BF in favor of $x = 0$ is simply the ratio of the posterior to the prior evaluated at the origin.

We can elucidate the role of the prior in Eq. (3) by considering a specific functional form. For simplicity, assume the marginal posterior is given by a standard normal distribution, truncated symmetrically around the origin by some uniform prior of half-width Δ (i.e., flat in $-\Delta < x < \Delta$). Then, the BF can be computed analytically to yield

$$\mathcal{B}^{x=0}_{x\neq 0} = \frac{1}{\sqrt{2\pi}}\frac{2\Delta}{\Phi(\Delta)-\Phi(-\Delta)} = \sqrt{\frac{2}{\pi}}\frac{\Delta}{\mathrm{erf}(\Delta/\sqrt{2})}\,, \tag{4}$$

in terms of the standard cumulative distribution function $\Phi(x) \equiv \left(1+\mathrm{erf}(x/\sqrt{2})\right)/2$ and the error function $\mathrm{erf}(x) \equiv (2/\sqrt{\pi})\int_0^x \exp(-y^2)\,\mathrm{d}y$. Since $\mathrm{erf}(\Delta/\sqrt{2}) \to 1$ for large Δ, the BF can be made to favor $x = 0$ with arbitrarily-high confidence by sufficiently broadening the prior—in fact, $\mathcal{B}^{x=0}_{x\neq 0} \propto \Delta$ in the large Δ limit. We illustrate this in Fig. 1.

The dependence on the prior range is a general feature not specific to our example: the same data can produce arbitrary odds in favor of a specific value of a parameter ($x = 0$ here) relative to a model with increasing prior volume (proxied by Δ). This is related to the concept of the Occam penalty in Bayesian inference: BFs do not only favor the model that fits the data best, but also the one that is simplest—where simplicity is defined as a model's ability to fit the data without having to significantly constrain its parameters relative to their a priori allowed values. The interplay between goodness of fit and Occam penalty creates the possibility of a BF that strongly favors the incorrect conclusion. In the context of testing GR, if the theory is indeed violated, then some observation can give $p(x = 0 \mid d) \ll 1$ in Eq. (3); yet, this can always be countered with a wide enough prior that makes the Occam penalty $1/p(x = 0)$ so large that goodness of fit cannot overcome it. This is the expected manifestation of the Occam penalty in BFs; the failure is symptomatic of a mismatch between the implemented prior and the observer's expectation.

Since we generally have little a priori information about the nature of possible beyond-GR effects, and these effects are not strongly constrained by the likelihood of an individual measurement, a natural inclination is to make the prior much wider than the likelihood. However, the argument above shows that a broad prior prevents us from detecting small deviations from GR, which is an important regime for tests of GR [53]—either because GR is close to correct or because of selection effects that disfavor the detection of signals with morphology far from GR. The same tension arises in other contexts where BFs are used without a principled prior. In the next section, we show that this behavior is not unique to single-event analyses but carries over to combined constraints.

III. COMBINING EVENTS UNDER A BROAD PRIOR

We now turn to collections of measurements and show that the "combined" multiplicative BF does not have the correct scaling in the regime of interest, with support accumulating in favor of the null hypothesis even when this conclusion is incorrect. Combining BFs from multiple events will only lead to the right conclusion when the deviation (e.g., the deviation from GR) is large enough to be apparent in individual posteriors, negating the need for combining observations in the first place. We then show that the hierarchical approach is not susceptible to this issue.

Again, consider a single parameter, x, that stands in for the magnitude of the GR deviation or any other effect

[1] Here we are identifying $\mathcal{H}_{x\neq 0}$ with the model in which x is allowed to vary freely; this is legitimate because the point $x = 0$ is a set of measure zero, so it does not need to be explicitly excised from the arbitrary-x model.

of interest, and let $x = 0$ be our null hypothesis (e.g., GR is correct). We conduct experiments to measure x, and assume additive Gaussian noise so that the observed value $x_{\rm obs}$ is normally distributed about the true value x with standard deviation (measurement noise) $\sigma_{\rm obs}$:

$$p\left(x_{\rm obs} \mid x\right) = \frac{1}{\sqrt{2\pi\sigma_{\rm obs}^2}} \exp\left(-\frac{\left(x_{\rm obs}-x\right)^2}{2\sigma_{\rm obs}^2}\right). \quad (5)$$

In the previous section, we considered a simplified case of this model, in which we had a single measurement with $x_{\rm obs} = 0$ and $\sigma_{\rm obs} = 1$.

We consider two "populations" for the true values of x under repeated measurements, i.e., the true magnitude of the GR deviation for each observed GW event. In the first the true value of x is fixed to some value x_0 for each measurement so that the population distribution is a Dirac delta,

$$p_1\left(x\right) = \delta\left(x - x_0\right). \quad (6)$$

In the second, the value of x is randomly distributed with mean zero and standard deviation σ_0 for each measurement, so that the population distribution is

$$p_2\left(x\right) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{x^2}{2\sigma_0^2}\right). \quad (7)$$

In the language of tests of gravity, the two models recover GR (our null hypothesis) when $x_0 = 0$ or $\sigma_0 = 0$, respectively. In the second model, the *mean* deviation from GR vanishes, but the actual variation in a given measurement fluctuates. In both cases the deviation parameter x is assumed independent and identically distributed (iid) for each measurement, so there is some prior choice such that multiplying BFs for repeated measurements is optimal [36].[2] However, that prior is unknown in realistic situations because the true distribution of GR deviations is not known a priori. Rather, in all cases we assume a measurement is analyzed with a flat prior on $-\Delta < x < \Delta$, following common practice.

A. Bayes factors

With that prior, the BF between the null and generalized hypotheses (for concreteness, "GR" and "non-GR") for a given observation $x_{\rm obs}$ is a generalization of Eq. (4)

$$\mathcal{B}^{\rm GR}_{\rm nGR} = \frac{2\Delta}{\sqrt{2\pi\sigma_{\rm obs}^2}\left(\Phi\left(\frac{\Delta-x_{\rm obs}}{\sigma_{\rm obs}}\right) - \Phi\left(\frac{-\Delta-x_{\rm obs}}{\sigma_{\rm obs}}\right)\right)} \times \exp\left(-\frac{x_{\rm obs}^2}{2\sigma_{\rm obs}^2}\right). \quad (8)$$

When the prior is very broad, $\Delta \gg x_{\rm obs}$, and the Gaussian posterior on x is not meaningfully truncated by the prior, then the BF simplifies to

$$\mathcal{B}^{\rm GR}_{\rm nGR} \simeq \frac{2\Delta}{\sqrt{2\pi\sigma_{\rm obs}^2}} \exp\left(-\frac{x_{\rm obs}^2}{2\sigma_{\rm obs}^2}\right), \quad (9)$$

corresponding to the $\mathcal{B}^{x=0}_{x\neq 0} \propto \Delta$ limit discussed in the previous section. Ensuring $\Delta \gg x_{\rm obs}$ is a common analysis choice because this prior permits the true value of x to correspond to the observed value $x_{\rm obs}$ which is the value of x that maximizes the likelihood in each observation. A broad prior is also desired when combining multiple observations in order to accommodate the expected scatter in the individual likelihoods.

If we choose to combine observations by adding log BFs[3] (equivalently, multiplying BFs), it is sufficient to compute the expected value of an individual log BF under the true deviations; the expected total log BF will then be the expected individual log BF times the number of events. The expected value of the log of Eq. (8) is not expressible in closed form, but can be straightforwardly computed numerically. In the limit that $\Delta \gg x_{\rm obs}$, the expected log BFs under our two populations of GR deviations become

$$\begin{aligned}\left\langle \ln \mathcal{B}^{\rm GR}_{\rm nGR}\right\rangle\big|_{x_0} &\simeq \ln\frac{2\Delta}{\sqrt{2\pi\sigma_{\rm obs}^2}} - \frac{\sigma_{\rm obs}^2+x_0^2}{2\sigma_{\rm obs}^2} \\ &\simeq \ln\frac{\Delta}{\sigma_{\rm obs}} - 0.23 - \frac{\sigma_{\rm obs}^2+x_0^2}{2\sigma_{\rm obs}^2}, \end{aligned} \quad (10)$$

and

$$\begin{aligned}\left\langle \ln \mathcal{B}^{\rm GR}_{\rm nGR}\right\rangle\big|_{\sigma_0} &\simeq \ln\frac{2\Delta}{\sqrt{2\pi\sigma_{\rm obs}^2}} - \frac{\sigma_{\rm obs}^2+\sigma_0^2}{2\sigma_{\rm obs}^2} \\ &\simeq \ln\frac{\Delta}{\sigma_{\rm obs}} - 0.23 - \frac{\sigma_{\rm obs}^2+\sigma_0^2}{2\sigma_{\rm obs}^2}. \end{aligned} \quad (11)$$

From the above, we can see that, whenever the GR deviation is nonzero but small enough to be undetectable in a single observation ($0 < x_0, \sigma_0 \ll \sigma_{\rm obs}$), then for $\Delta \gtrsim 2\sigma_{\rm obs}$ the expected log BF is positive, and evidence accumulates *in favor* of GR even though there is a deviation. For deviations that are marginally detectable in a single observation $x_0, \sigma_0 \simeq \sigma_{\rm obs}$, choosing a wide, uninformative prior $\Delta \gtrsim 3.5\sigma_{\rm obs}$ (which, recall, is necessary to ensure that all observations $x_{\rm obs}$ in a modest-sized catalog of observations are within the prior range) will result in evidence that accumulates *against* modifications to GR.

An exact calculation of the expected log BF without the assumption that $\Delta \gg x_{\rm obs}$ appears in Fig. 2. Interestingly, any value of Δ will accumulate evidence for GR when $x_0 = 0$ or $\sigma_0 = 0$; but if $x_0 > 0$ or $\sigma_0 > 0$, so that the GR model is incorrect, prior choices that encompass deviation parameter values comparable to those observed,

[2] The optimal prior is the actual population from which the true parameters controlling the measurements are iid draws.

[3] In this paper, all logarithms are natural logarithms.

i.e., $\Delta \simeq \text{few} \times \sigma_{\rm obs}$, accumulate evidence for the incorrect GR model unless the deviation parameter is comparable to or larger than the observational uncertainty. In this regime, either the BF fails to select the correct theory (non-GR) on average (with more and more certainty as the number of observations grows); or the deviation is so large that it is marginally detectable (i.e., "$\sim 1\sigma$") with a single observation. Multipliying BFs in this situation is counter-productive, and the best constraint is achieved with a single measurement.

Even when the log BF is expected to take the correct sign on average, the result will vary for any given set of detections. In the regime where $\Delta \gg x_{\rm obs}$, we can quantify this through the variance associated with the means in Eqs. (10) and (11). For the former, this is just

$$\text{var}\left(\ln \mathcal{B}^{\rm GR}_{\rm nGR}\right)\big|_{x_0} \simeq \left(\frac{x_0}{\sigma_{\rm obs}}\right)^2 + \frac{1}{2}\,, \tag{12}$$

while for the latter we have

$$\text{var}\left(\ln \mathcal{B}^{\rm GR}_{\rm nGR}\right)\big|_{\sigma_0} \simeq \frac{\left(\sigma_0^2 + \sigma_{\rm obs}^2\right)^2}{2\sigma_{\rm obs}^4}\,. \tag{13}$$

These scalings are illustrated in Fig. 3 for $\Delta = 5\sigma_{\rm obs}$. On the left hand side, the variance asymptotes to 1/2 for decreasing $\sigma_{\rm obs}$ and fixed intrinsic scatter σ_0; on the right hand side, the BF variance keeps increasing as we make the individual measurements more precise (decreasing $\sigma_{\rm obs}$) because for narrower posteriors the value at $x = 0$ varies more drastically from event to event. In the case where GR is correct ($x_0 = \sigma_0 = 0$), the scatter in the single-event log BF is large enough that it is not extremely uncommon for moderately large catalogs to yield evidence for the wrong hypothesis, even when the null hypothesis is correct (Fig. 4).

The qualitative behavior here can be understood in the context of [36]. The flat prior on x implicitly assumes that each observation has a true x parameter that is independent of other observations and uniformly distributed on $(-\Delta, \Delta)$. If one chooses Δ large enough to include all event likelihoods and not truncate them, then the flat prior for the true parameter implicitly demands that most of the true deviation parameters are comparable to Δ. If, instead, the deviation parameters are considerably smaller than the observational uncertainty $\sigma_{\rm obs}$—which is the regime where stacking multiple events *should* be the most beneficial—then the assumption is so badly violated that pooling observations prefers the incorrect model where the deviation parameters are zero to the even-less-correct model where the deviation parameters are iid uniform on $(-\Delta, \Delta)$.

The solution to this problem is to allow the assumed population to adapt its properties to the stacked set of observations, for example by allowing its location and scale to fit the set, as suggested in [5]. This effectively constructs a model that better represents our beliefs about the combined data set.

B. Hierarchical treatment

We now revisit the two experiments above using a hierarchical model for the distribution of GR deviations, instead of computing BFs with a uniform prior. Assuming we are only interested in the first two moments of the distribution, we parametrize the true deviations as drawn from a Gaussian such that $x \sim \mathcal{N}(\mu, \sigma_{\rm pop})$ [5]. The goal will be to infer the μ and $\sigma_{\rm pop}$ hyperparameters from the collection of measurements, and to quantify agreement with the null hypothesis $\mu = \sigma_{\rm pop} = 0$ based on the corresponding 2D posterior. The two non-GR models we considered above are encompassed within this parametrization when ($\mu = x_0$, $\sigma_{\rm pop} = 0$) and ($\mu = 0$, $\sigma_{\rm pop} = \sigma_0$).[4]

As before, we assume that the observed value $x_{\rm obs}$ in an individual event is normally distributed around the true value per Eq. (5), and we take the true values themselves to be distributed normally given μ and $\sigma_{\rm pop}$. In that case, the likelihood for a given observation becomes (see Appendix A)

$$p(x_{\rm obs} \mid \mu, \sigma_{\rm pop}, \sigma_{\rm obs}) = \frac{1}{\sqrt{2\pi\sigma_{\rm tot}^2}} \exp\left(-\frac{(x_{\rm obs} - \mu)^2}{2\sigma_{\rm tot}^2}\right), \tag{14}$$

where $\sigma_{\rm tot}^2 \equiv \sigma_{\rm obs}^2 + \sigma_{\rm pop}^2$ is the total variance arising from the combination of statistical uncertainty and the intrinsic population scatter. Unlike in Eq. (5), the true value x for the individual measurement does not appear in this likelihood because we have marginalized over it and replaced it with the hyperparameters μ and $\sigma_{\rm pop}$.

Consider the same two true distributions for x above, which depart from GR as given by Eqs. (6) and (7). We simulate catalogs of detections following these distributions and analyze them hierarchically with the likelihood of Eq. (14), and Gaussian priors on μ and $\sigma_{\rm pop}$ (with zero mean and standard deviation equal to $\sigma_{\rm obs}$, restricting to positive values for $\sigma_{\rm pop}$). Although other choices are possible, it is natural to tie the scale of the hyperprior to the measurement uncertainty because this is the only scale built into the problem a priori. Furthermore, this choice is guaranteed to be sufficiently broad in the small-deviation regime ($x_0, \sigma_0 < \sigma_{\rm obs}$) in which we are interested, and smooth enough to accommodate larger deviations if needed.

We quantify agreement with the null hypothesis through the marginal posteriors for μ and $\sigma_{\rm pop}$, as well as the credible level at which GR is recovered in the 2D posterior for those two quantities (the 2D quantile, $\mathcal{Q}_{\rm GR}$ in [6]), defined as

$$\mathcal{Q}_{\rm GR} \equiv \int_{p<p(0,0)} p(\mu, \sigma_{\rm pop} \mid x_{\rm obs}, \sigma_{\rm obs})\, \mathrm{d}\mu\, \mathrm{d}\sigma_{\rm pop}\,, \tag{15}$$

[4] The next section examines a non-GR model that is not fully encompassed in the Gaussian hierarchical model.

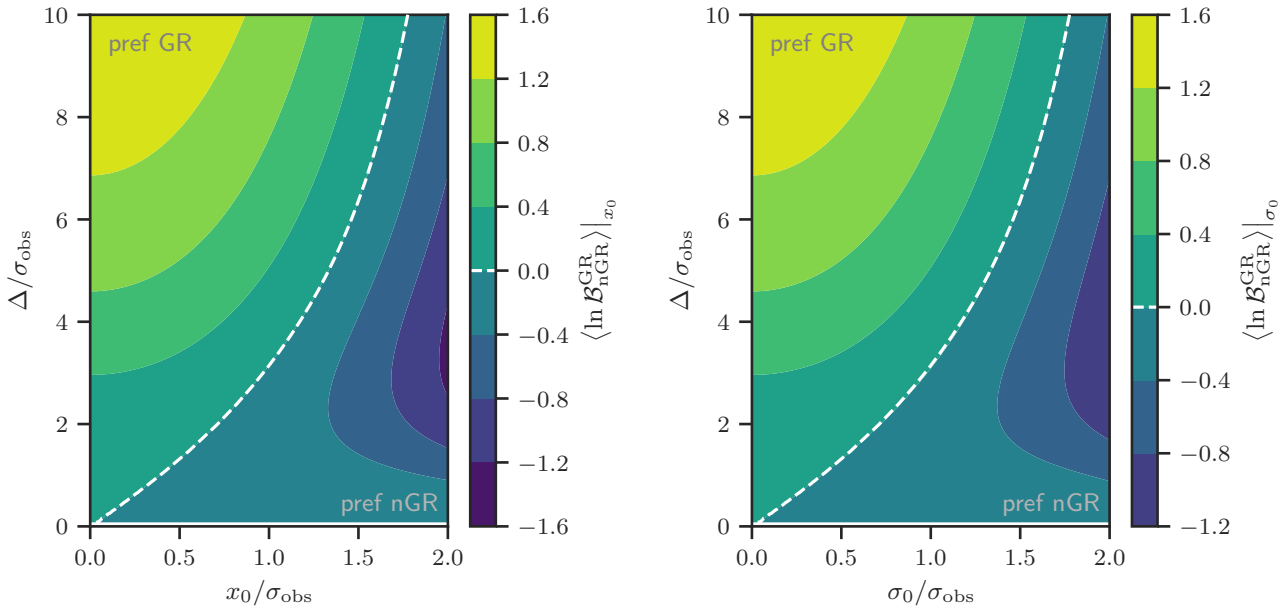

FIG. 2. The expected log BF in favor of GR from the toy model discussed in Sec. III (color) as a function of the prior width Δ (ordinate) and the true deviation parameter x_0 (abscissa left) or the scatter in the zero-mean deviation parameter σ_0 (abscissa right). Both quantities are shown relative to the measurement uncertainty on the deviation parameter $\sigma_{\rm obs}$. For any non-zero value of the deviation parameter or its scatter, there is a prior width that results in incorrectly accumulating evidence *for* GR (above dashed line); moreover, for reasonable prior widths of a few times the measurement uncertainty (so that observed values of the deviation parameter lie within the prior range and the prior is therefore broad and uninformative) the deviation parameter must be large enough to be marginally detectable in a single measurement ($x_0, \sigma_0 \sim \sigma_{\rm obs}$) before evidence against the incorrect GR model accumulates on average over many stacked measurements. Fig. 3 demonstrates this effect for $\Delta = 5\sigma_{\rm obs}$.

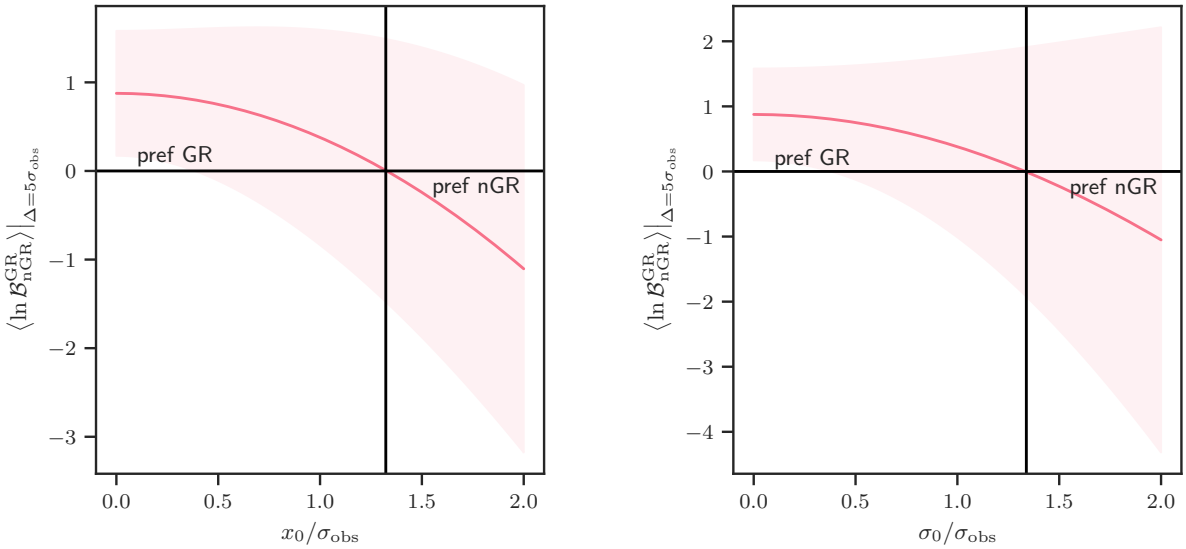

FIG. 3. The expected log BF for the GR model when $\Delta = 5\sigma_{\rm obs}$ (a broad, uninformative prior) versus the fixed value of the deviation parameter (x_0, left) or scatter in the zero-mean deviation parameters (σ_0, right). For this choice of prior width, the average log BF favors the incorrect GR model until the deviation parameter is $x_0, \sigma_0 \gtrsim 1.3\sigma_{\rm obs}$, marginally detectable in a single measurement. Shading shows the expected variation in the log BF ($\pm 1\sigma$).

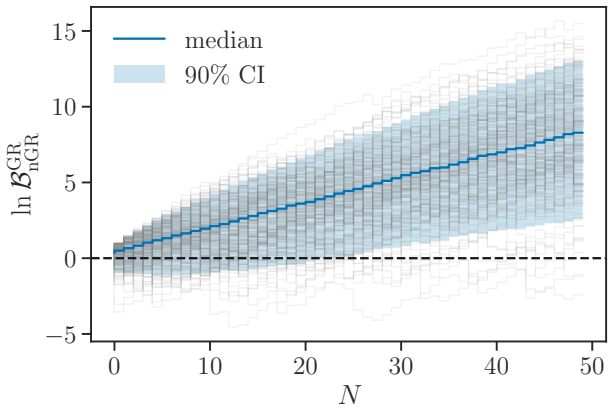

FIG. 4. Log BF for a growing catalog of N observations in the case where the null hypothesis (GR) is correct, and where $\Delta = 2\sigma_{\rm obs}$, for 200 simulations (represented by each thin gray trace). On average, the log BF favors the right hypothesis (blue line), but there is enough variance in this quantity that some of the individual catalogs favor the wrong hypothesis, even for a moderately high number of detections (traces below the dashed line).

where the shorthand "$p < p(0,0)$" stands in for values of μ and $\sigma_{\rm pop}$ such that $p(\mu, \sigma_{\rm pop} \mid x_{\rm obs}, \sigma_{\rm obs}) < p(\mu = 0, \sigma_{\rm pop} = 0 \mid x_{\rm obs}, \sigma_{\rm obs})$. Given this definition, a value of $\mathcal{Q}_{\rm GR} = 1$ means that the posterior peaks at the origin $\mu = \sigma_{\rm pop} = 0$, while $\mathcal{Q}_{\rm GR} = 0$ means that the posterior offers no support for that point (i.e., a higher value of $\mathcal{Q}_{\rm GR}$ implies better agreement with GR). In all cases, we simulate each catalog of N observations 50 times and report medians over the ensemble.

Unlike with BFs, there are no values of x_0 or σ_0 for which the hierarchical analysis converges to the wrong answer given enough observations. This is apparent from Fig. 5, which shows the value of $\mathcal{Q}_{\rm GR}$ as a function of the deviation magnitude ($x_0/\sigma_{\rm obs}$ or $\sigma_0/\sigma_{\rm obs}$) and the number of observations: even for deviations small relative to the individual-measurement uncertainty, $\mathcal{Q}_{\rm GR}$ approaches zero for large N—indicating that the posterior offers little support for $\mu = \sigma_{\rm pop} = 0$, in tension with GR. For small catalogs ($N \leq 10$), the value of $\mathcal{Q}_{\rm GR}$ is more strongly influenced by the prior, to a greater or lesser extent depending on the magnitude of the deviation.

Besides indicating that the data are inconsistent with the null hypothesis, the hierarchical analysis provides descriptive information about the nature of the deviation. With enough observations, the measurements of μ and $\sigma_{\rm pop}$ converge to the x_0 and σ_0 values respectively for the two models with increasing precision for larger N. In Fig. 6, we show this behavior explicitly for two example magnitudes of the GR deviation. As expected, we can detect larger deviations with fewer detections, and need more observations to notice a nonzero scatter than a nonzero mean.

Prior choices play a lesser role in the hierarchical approach than in BF computations. The hierarchical analysis takes as input the likelihoods for individual events, so the prior used to initially analyze the data is largely irrelevant as long as it offers the likelihood ample support (e.g., $\Delta \gg \sigma_{\rm obs}$ in the notation of the previous section). The choice of prior is, instead, transferred to the μ and $\sigma_{\rm pop}$ hyperparameters; however, any reasonably smooth choice will work assuming we have a enough events for the hierarchical measurement to be informative. If observations are not sufficiently numerous, the result will be influenced by the μ and $\sigma_{\rm pop}$ prior.

In the case of our idealized examples, we can analytically predict the number of detections needed for the hierarchical measurement to be informative. As we show in Appendix A, the variance of the marginalized hierarchical likelihood is expected to scale as

$$\mathrm{var}\,(\mu) = \frac{\sigma^2_{\rm obs} + \sigma^2_0}{N} \tag{16}$$

for the inferred population mean and

$$\mathrm{var}\left(\sigma^2_{\rm pop}\right) = \frac{2}{N}\left(\sigma^2_{\rm obs} + \sigma^2_0\right)^2 \tag{17}$$

for the inferred population variance, assuming the true population variance is σ^2_0 and irrespective of the true mean. As a rule of thumb, the hierarchical measurement will become informative once the characteristic width of the likelihood of Eq. (14) becomes smaller than the scale imposed by the prior. With hyperpriors of scale $\sigma_{\rm prior}$, this implies that the μ measurement should start becoming informative (in the sense that we obtain a hyperposterior narrower than the prior) once we accumulate

$$N \gtrsim \frac{\sigma^2_{\rm obs} + \sigma^2_{\rm pop}}{\sigma^2_{\rm prior}} = \left(\frac{\sigma_{\rm tot}}{\sigma_{\rm prior}}\right)^2 \tag{18}$$

measurements; meanwhile, for $\sigma_{\rm pop}$, the equivalent requirement is

$$N \gtrsim \frac{\left(\sigma^2_{\rm obs} + \sigma^2_{\rm pop}\right)^2}{\sigma^4_{\rm prior}} = \left(\frac{\sigma_{\rm tot}}{\sigma_{\rm prior}}\right)^4, \tag{19}$$

as we show in Appendix A.

We demonstrate these scalings in Fig. 7, where we show the variance in the inferred population mean and variance from simulated populations of measurements with $\sigma_0 = \sigma_{\rm obs}/2$ and no mean, and setting $\sigma_{\rm prior} = \sigma_{\rm obs}$ for concreteness. For increasing number of measurements N, the posterior converges to the true values ($\mu = 0, \sigma_{\rm pop} = \sigma_{\rm obs}/2$). For small catalogs, i.e., for N comparable or smaller than the thresholds above, the average uncertainty in these measurements is broad but smaller than expected simply from Eqs. (16) and (17) because it is dominated by the prior. As the number of detections increases, the posterior variance becomes well described by Eqs. (16) and (17), meaning that the data become informative and the likelihood dominates over the prior.

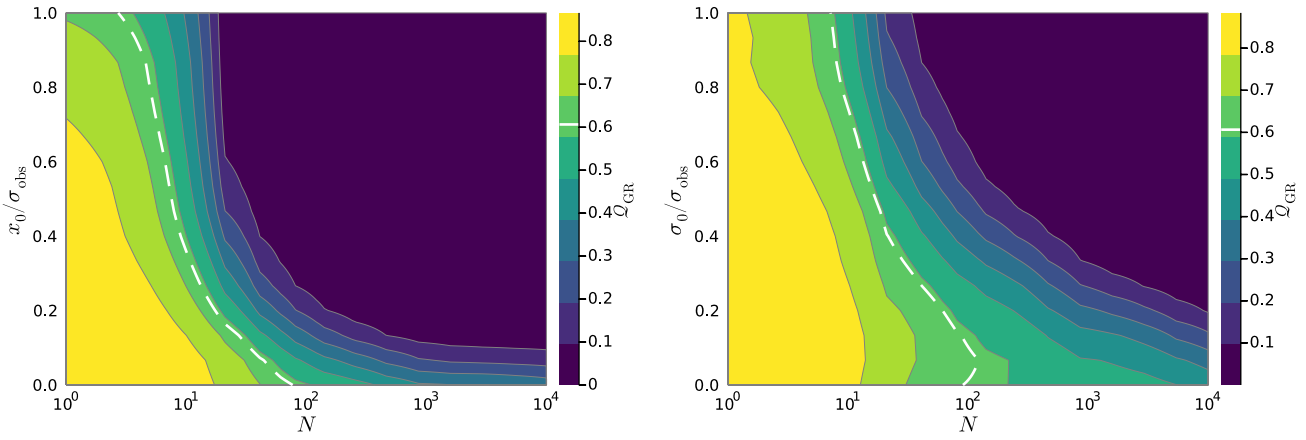

FIG. 5. The GR quantile $\mathcal{Q}_{\rm GR}$ (color) obtained in a hierarchical analysis of data in which there is a fixed deviation for all events (x_0, ordinate left) or a scatter across events (σ_0, ordinate right), as a function of the number of events (N, abscissa). The reported value of $\mathcal{Q}_{\rm GR}$ is the median over 50 simulated catalogs with N-events. Low values of $\mathcal{Q}_{\rm GR}$ indicate that the null hypothesis ($\mu = \sigma_{\rm pop} = 0$) is disfavored; for reference, the dashed line marks the point at which GR is disfavored at 1σ, i.e., $\mathcal{Q}_{\rm GR} = \exp(-1/2) \approx 0.61$. Unlike in Fig. 2, given enough N we always detect the deviation, even when $x_0, \sigma_0 < \sigma_{\rm obs}$.

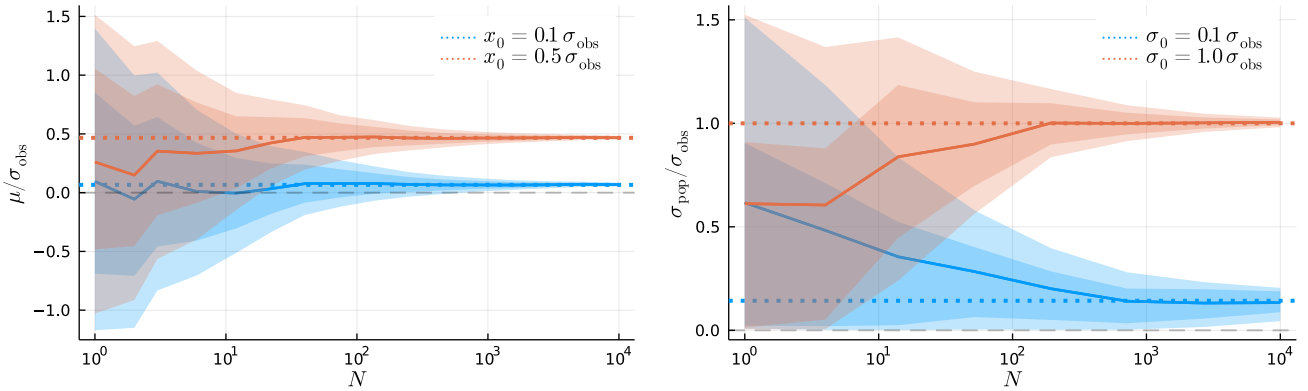

FIG. 6. Recovered population mean (μ, left) and standard deviation ($\sigma_{\rm pop}$, right) as a function of catalog size (N, abscissa), for our two toy models: a fixed deviation x_0 for all events (left) and a deviation scatter σ_0 across events (right). In each case, we show two values for the true deviation: $x_0, \sigma_0 = 0.1\,\sigma_{\rm obs}$ (blue) and $x_0, \sigma_0 = 0.8\,\sigma_{\rm obs}$ (red). The measurement is represented by posterior median (solid lines) surrounded by 68% and 90% highest-density credible bands (shading); we also show the true value (dotted lines) and the null expectation (dashed gray line). Smaller deviations require more observations to be detected—for example, we only need $N \gtrsim 3$ events to notice $\mu > 0$ at 1σ (68% credibility) if $x_0 = 0.8\,\sigma_{\rm obs}$, but $N \gtrsim 100$ if $x_0 = 0.1\,\sigma_{\rm obs}$.

Alternately, we may ask how many measurements would be required to establish a non-vanishing population mean or variance. This requires $\mathrm{var}\,(\mu) \lesssim x_0^2$ or $\mathrm{var}\left(\sigma_{\rm pop}^2\right) \lesssim \sigma_0^4$. The former would imply

$$N \gtrsim \left(\frac{\sigma_{\rm tot}}{x_0}\right)^2 \tag{20}$$

and the latter

$$N \gtrsim 2\left(\frac{\sigma_{\rm tot}}{\sigma_0}\right)^4. \tag{21}$$

IV. FENCEPOST MODEL

In the examples above, the simulated populations could be perfectly reproduced as special cases of the hierarchical population model—that is, there existed a choice of μ and $\sigma_{\rm pop}$ for which the hierarchical model reduced exactly to the true distribution we simulated. Of course, we do not necessarily expect this to be the case in reality: a deviation from GR could manifest as a nontrivial function of the source parameters and the coupling constants in the theory; similarly, for other effects such as memory, it is not realistic to expect the true population of parameters to be fully described by a simple Gaussian if the null hypothesis is incorrect. In light of that, one might worry that the hierarchical method only outperformed BFs in

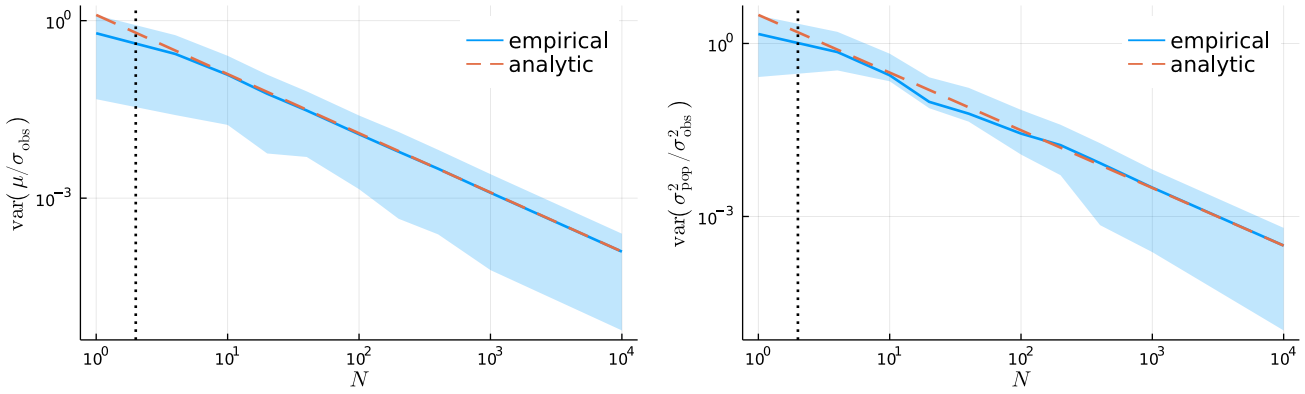

FIG. 7. Scaling of the uncertainty in the recovered hyperparameters μ (left) and σ^2_{pop} (right) as a function of catalog size N, computed for a case in which the true values are $\mu = 0$ and $\sigma_{\mathrm{pop}} = \sigma_{\mathrm{obs}}/2$. A solid line marks the median of the posterior variance computed over 50 simulated catalogs for any given N, while surrounding bands enclose the corresponding 68% highest-density interval. For low N, the variances $\mathrm{var}\,(\mu/\sigma_{\mathrm{obs}})$ and $\mathrm{var}\left(\sigma^2_{\mathrm{pop}}/\sigma^2_{\mathrm{obs}}\right)$ are dominated by the hyperprior; as N increases, the data become more informative and the variances approach the analytic prediction of Eqs. (16) and (17) (red dashed lines). Dotted vertical lines mark the expected number of observations needed for the likelihood of Eq. (14) to become narrower than the prior, as dictated by Eqs. (18) and (19) with $\sigma_{\mathrm{prior}} = \sigma_{\mathrm{obs}}$, but rounded up to the closest integer.

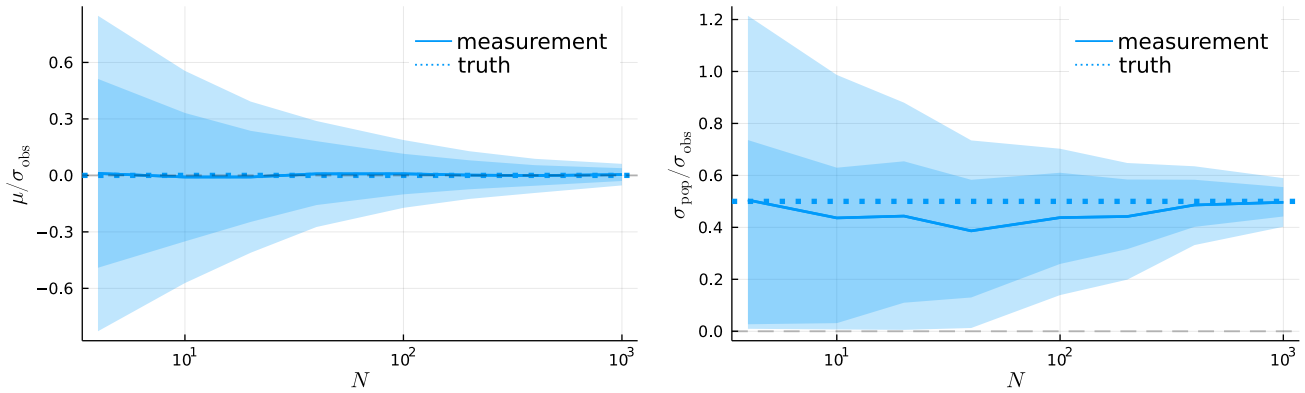

FIG. 8. As in Fig. 6, except the true population follows the fencepost model of Sec. IV, by which the true deviation is $x = \pm\sigma_{\mathrm{obs}}/2$ with equal probability for either sign. Even though this distribution cannot be expressed as the limit of a Gaussian, the hierarchical analysis infers the correct values of $\mu = 0$ and $\sigma_{\mathrm{pop}} = \sigma_{\mathrm{obs}}/2$.

the above examples because the hierarchical model was able to match the true population exactly, and that this gain would fail to materialize in realistic situations. Yet, as we show in this section, this is not the case: the hierarchical method is more robust than products of BFs even when the underlying population cannot be fit exactly by a Gaussian.

Consider a situation in which the true deviation parameter is either $x = \pm x_0$ for some x_0 and with equal probability for both signs. The true distribution in this "fencepost" model is simply the sum of two delta functions,

$$p_3\,(x) = \frac{1}{2}\left[\delta\,(x - x_0) + \delta\,(x + x_0)\right], \tag{22}$$

and, therefore, has zero mean and standard deviation $\sigma_0 = |x_0|$. This population cannot be described as a limiting case of a Gaussian distribution; nevertheless, we can always analyze it hierarchically with the Gaussian likelihood of Eq. (14), and expect to recover the correct values for the population mean and spread (namely, $\mu = 0$ and $\sigma_{\mathrm{pop}} = |x_0|$) given enough detections [5]. In Fig. 8, we show this explicitly for simulated measurements in which $x_0 = \sigma_{\mathrm{obs}}/2$.

On the other hand, when presented with the fencepost model, BF computations suffer from the same problems already identified above: with a broad prior relative to the measurement uncertainty ($\Delta > \sigma_{\mathrm{obs}}$), the combined BF for multiple observations will necessarily converge to the wrong answer (i.e., favor of the null hypothesis) unless $x_0 \gtrsim \sigma_{\mathrm{obs}}$, in which case the deviation is detectable in individual observations. Similarly to Fig. 2, in Fig. 9 we show the scaling of the expected BF accumulated from fencepost-model observations, as a function of x_0 and Δ.

In Fig. 10, we compare the hierarchical and Bayes-factor

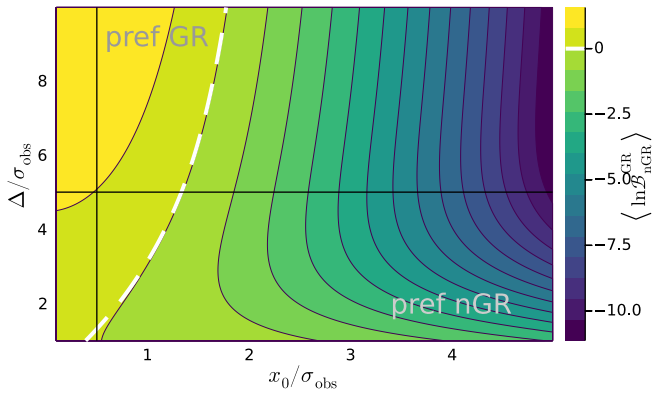

FIG. 9. Average log BF in favor of GR accumulated for each observation in the "fencepost" model described in Sec. IV as a function of the prior width Δ and the location of the true deviations at $\pm x_0$. The black lines indicate the parameter values chosen for the numerical experiment illustrated in Fig. 10.

results for a progressively-higher number of simulated measurements from a fencepost population with $x_0 = \sigma_{\rm obs}/2$. As the number of measurements increases, the hierarchical model disfavors the null hypothesis more strongly, with $\mathcal{Q}_{\rm GR}$ vanishing rapidly for increasing N. On the other hand, a BF computed with $\Delta = 5\sigma_{\rm obs}$ grows exponentially in favor of the null hypothesis, yielding a spectacularly incorrect result.

Of course, in the fencepost example as in any realistic situation, determining that the observations are inconsistent with a delta function at the origin would not be sufficient to fully characterize the population distribution. Were we to find inconsistency with $\mu = \sigma_{\rm pop} = 0$, then we would carry out a followup analysis to more comprehensively infer the properties of the population, e.g., by applying a population model with higher moments than a Gaussian. Regardless, the hierarchical test is always well suited as a first null-test, because a vanishing mean and variance is a necessary condition for the null hypothesis.

V. CONCLUSIONS

Although conceptually appealing in idealized situations, the use of BFs to aggregate information from multiple observations presents difficulties in practice. Their apparent simplicity in reducing a complex model selection problem to a single number hides an opaque dependence strict and unrealistic population assumptions. Unless priors (aka, the "population") adapt to the observations at hand, BFs are difficult to interpret—a problem that is compounded when multiplying such BFs from a catalog of observations. Even when priors are adequate, the result on its own provides no insight as to *why* a model is to be preferred over another. This and related problems have been widely discussed in the statistics literature [e.g., 54, and references therein], but not extensively in the context of GWs and testing GR.

In this paper, we have examined BFs and hierarchical posteriors as two commonly-used alternatives to derive information from collections of GW detections in order to decide between two models, e.g., the presence or absence of beyond-GR effects in the detected waveforms. We furthered arguments in [5, 36] to show that, without a principled way to set priors, BFs are an unreliable tool for this task. We demonstrated this with three examples in which the value of some parameter x encoding the effect in question (e.g., a deviation from GR) follows different distributions deviating from the null hypothesis.

We have found that, when the truth does not conform to the null model, the usual approach of multiplying single-event BF converges to the incorrect answer for an increasing number of observations, except in a regime where the targeted effect is discernible in individual observations (thus negating the need for combining events in the first place). On the other hand, hierarchical modeling of the underlying population leads to identification of the appropriate priors (aka, the "population model") and converges to the correct answer. We established this in the context of nested models for which GR can be recovered as a special case of the beyond-GR model (i.e., $x_0 = 0$); however, the issue of sensitivity to the prior width will still be present in non-nested models [e.g., 55] where it will require a different solution.

In principle, BFs could be computed after the hierarchical population inference [54] or between different population models [3], but we here show that they are unreliable without this step. Even then, it is not possible to evade the core problem of prior dependence when computing BFs, no matter how many levels of inference are applied: the BF computation based on the highest level of inference in a hierarchical model will still depend on the choice of priors on that level, reducing the problem once again to the choice of a prior distribution that is difficult to establish in a principled way. This issue is devistatingly acute for the approach that multiplies Bayes factors with a simple, fixed prior because each observation contributes an additional prior factor.

ACKNOWLEDGMENTS

We are grateful to Tyson Littenberg for insightful feedback on this manuscript. During part of this work, M.I. was supported by NASA through the NASA Hubble Fellowship grant No. HST-HF2-51410.001-A awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS5-26555. The Flatiron Institute is a division of the Simons Foundation, supported through the generosity of Marilyn and Jim Simons. We thank Gregorio Carullo for comments on non-nested models. This paper carries LIGO document number LIGO-P2200099. Software: `matplotlib` [56], `julia` [57], `Turing.jl` [58], `Plots.jl` [59].

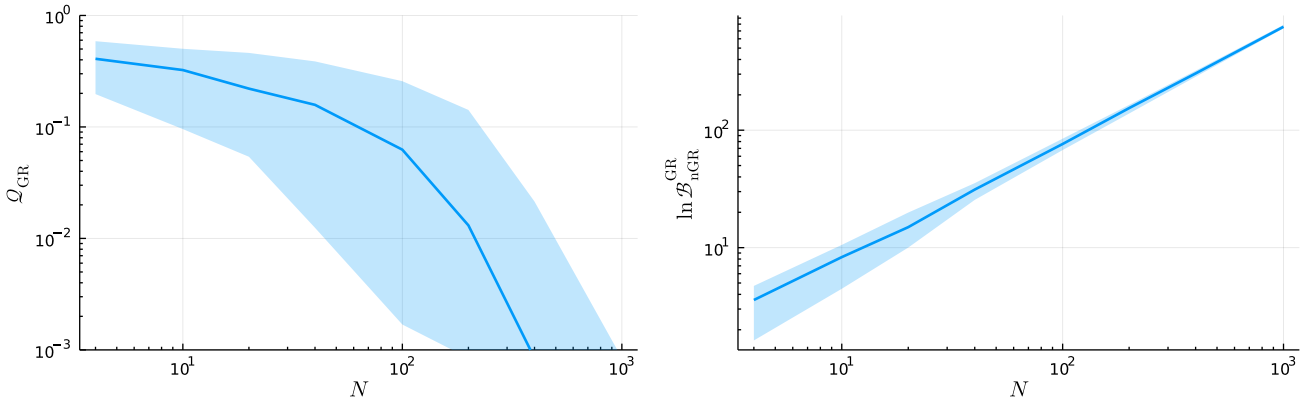

FIG. 10. Accumulation of credibility (left) or log BF (right) for GR versus catalog size N in numerical experiments corresponding to the parameter choices indicated by the black lines in Fig. 9. The solid line gives the median over 100 catalog realizations at each catalog size N while the band shows the range of the 16th to 84th percentile values. The credibility of GR (left) is the fraction of posterior mass for μ and σ in our hierarchical model that lies at a lower posterior density than the GR values $\mu = \sigma = 0$ when the model is fit to a catalog of observations whose true and observed deviations are drawn from the "fencepost" model described in Sec. IV with $x = \pm\sigma_{\rm obs}/2$. The log Bayes factor (right) is the sum of the log BFs for each observation in the catalog using a flat prior on the true deviation $-\Delta < x < \Delta$ with $\Delta = 5\sigma_{\rm obs}$. The hierarchical model correctly finds that there is little credibility for GR once the catalog size is a few hundred; the accumulated BF, on the other hand, becomes very confident in the incorrect GR model even at small catalog sizes.

Appendix A: Expectation value and variance of hierarchical parameters

Consider $i = 0 \ldots N-1$ measurements of parameters x_i whose true values, μ_i, are drawn from a normal distribution $\mu_i \sim \mathcal{N}(\mu, \sigma_{\rm pop})$; further assume each measurement is unbiased, i.e., $\langle x_i \rangle = \mu_i$, where the angle brackets denote a noise average, and that it is well represented by a Gaussian distribution such that $x_i \sim \mathcal{N}(\mu_i, \sigma_i)$, where σ_i is the measurement uncertainty.[5]

The joint likelihood for μ and $\sigma_{\rm pop}$ conditional on the N uncertainties σ_i can be obtained by marginalizing over the true values μ_i:

$$\begin{aligned} p(x_i \mid \mu, \sigma_{\rm pop}, \sigma_i) &= \int p(x_i \mid \mu_i, \sigma_i)\, p(\mu_i \mid \mu, \sigma_{\rm pop})\, {\rm d}\mu_i \\ &= \frac{1}{2\pi\sigma_i\sigma_{\rm pop}} \int e^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}} e^{-\frac{(\mu_i-\mu)^2}{2\sigma_{\rm pop}^2}}\, {\rm d}\mu_i \\ &= \frac{1}{\sqrt{2\pi\sigma_{{\rm tot},i}^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma_{{\rm tot},i}^2}}\,, \end{aligned} \tag{A1}$$

where $\sigma_{{\rm tot},i}^2 = \sigma_i^2 + \sigma_{\rm pop}^2$ is the total variance for the ith measurement. For the full set of N measurements $\{x_i\}$, then, the hierarchical likelihood is

$$p(\{x_i\} \mid \mu, \sigma_{\rm pop}, \{\sigma_i\}) = \prod_{i=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma_{{\rm tot},i}^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma_{{\rm tot},i}^2}}\,. \tag{A2}$$

The maximum-likelihood estimators for μ and $\sigma_{\rm pop}$, denoted $\hat{\mu}$ and $\hat{\sigma}_{\rm pop}$ respectively, can be found by enforcing

$$\partial_\mu \ln\left[p(\{x_i\} \mid \mu, \sigma_{\rm pop}, \{\sigma_i\})\right]\big|_{\mu=\hat{\mu}} = 0\,, \tag{A3}$$

where ∂_μ denotes the partial derivative $\frac{\partial}{\partial\mu}$, and similar for $\sigma_{\rm pop}$. We find that $\hat{\mu}$ and $\hat{\sigma}_{\rm pop}$ satisfy

$$\hat{\mu} = \frac{\sum x_i w_i}{\sum w_i}\,, \tag{A4}$$

and

$$\sum w_i - \sum w_i^2 \left(x_i - \hat{\mu}\right)^2 = 0 \tag{A5}$$

for $w_i \equiv \sigma_{{\rm tot},i}^{-2} = \left(\sigma_i^2 + \hat{\sigma}_{\rm pop}^2\right)^{-1}$.

We cannot solve this most general (heteroskedastic) case for $\hat{\mu}$ and $\hat{\sigma}_{\rm pop}$ in closed form, so we specialize to the (homoskedastic) case where all measurements have similar error, $\sigma_i = \sigma_{\rm obs}$ for all i. With this simplification, the above relations reduce to

$$\hat{\mu} = \frac{1}{N}\sum x_i\,, \tag{A6}$$

and

$$\hat{\sigma}_{\rm pop}^2 = \frac{1}{N}\sum \left(x_i - \hat{\mu}\right)^2 - \sigma_{\rm obs}^2\,. \tag{A7}$$

[5] In the main text, we used slightly different notation: instead of (x_i, μ_i, σ_i) we had $(x_{\rm obs}, x, \sigma_{\rm obs})$; the former is slightly more succinct, which will be helpful here given the increased number of mathematical expressions.

As might be expected, the maximum-likelihood estimate of the population mean is simply the sample mean, and the inferred population variance corresponds to the variance in the data that cannot be accounted for by the statistical uncertainty in each individual measurement.

We can go one step further and compute the uncertainty in these estimators, which we can take as a proxy for the width of the marginal likelihoods. The uncertainty in $\hat{\mu}$ is straightforward to compute, since this is just a linear combination of independent random variables x_i with known variance $\sigma^2_{\rm tot} \equiv \sigma^2_{\rm obs} + \sigma^2_{\rm pop}$, hence

$$\mathrm{var}\left(\hat{\mu}\right) = \frac{\sigma^2_{\rm obs} + \sigma^2_{\rm pop}}{N}\,. \tag{A8}$$

Obtaining $\mathrm{var}\left(\hat{\sigma}^2_{\rm pop}\right)$ is less straightforward, but we can do so by writing $\mathrm{var}\left(\hat{\sigma}^2_{\rm pop}\right) = F^{-1}$, in terms of the corresponding Fisher element[6]

$$F \equiv -\left\langle \partial^2_{\sigma^2_{\rm pop}} \ln\left[p(\{x_i\} \mid \mu, \sigma_{\rm pop}, \sigma_{\rm obs})\right]\Big|_{\sigma^2_{\rm pop}=\hat{\sigma}^2_{\rm pop}} \right\rangle . \tag{A9}$$

Doing the math, we find

$$\mathrm{var}\left(\hat{\sigma}^2_{\rm pop}\right) = 2\frac{\left(\sigma^2_{\rm obs} + \hat{\sigma}^2_{\rm pop}\right)^2}{N}\,. \tag{A10}$$

In the main text, we quoted these results for $\mathrm{var}\left(\hat{\mu}\right)$ and $\mathrm{var}\left(\hat{\sigma}^2_{\rm pop}\right)$ in Eqs. (16) and (17) respectively.

We can compare the width of the hierarchical likelihood, as proxied by the estimator variances above, to some typical scale of interest in the problem. Below we consider the scale imposed by the μ and $\sigma_{\rm pop}$ hyperpriors to estimate the number of events before the likelihood become informative with respect to the prior. We chose the hyperpriors to be Gaussians with scale $\sigma_{\rm prior}$, restricting to positive $\sigma_{\rm pop}$ values, i.e.,

$$p(\mu \mid \sigma_{\rm prior}) = \frac{1}{\sqrt{2\pi\sigma^2_{\rm prior}}} \exp\left(-\frac{\mu^2}{2\sigma^2_{\rm prior}}\right) \tag{A11}$$

for the mean, and

$$p(\sigma_{\rm pop} \mid \sigma_{\rm prior}) = \begin{cases} \sqrt{\frac{2}{\pi\sigma^2_{\rm prior}}} \exp\left(-\frac{\sigma^2_{\rm pop}}{2\sigma^2_{\rm prior}}\right) & (\sigma_{\rm pop} \geq 0) \\ 0 & (\sigma_{\rm pop} < 0) \end{cases} \tag{A12}$$

for the standard deviation, where the difference in normalization arises from the $\sigma_{\rm pop} \geq 0$ truncation. The prior variances in our example are, thus, $\mathrm{var}\left(\mu\right) = \sigma^2_{\rm prior}$ for the mean, and $\mathrm{var}\left(\sigma^2_{\rm pop}\right) = 2\sigma^4_{\rm prior}$ for the variance (obtained through direct computation). For concreteness, in the main text we set $\sigma_{\rm prior} = \sigma_{\rm obs}$, since that is the only scale intrinsic to the measurement.

We can now directly compute the number of observations required for the likelihood to achieve comparable widths by equating these variances to the likelihood variances from Eqs. (A8) and (A10) above. The result is

$$N \gtrsim \frac{\sigma^2_{\rm obs} + \sigma^2_{\rm pop}}{\sigma^2_{\rm prior}} \tag{A13}$$

for μ, and

$$N \gtrsim \left(\frac{\sigma^2_{\rm obs} + \sigma^2_{\rm pop}}{\sigma^2_{\rm prior}}\right)^2 = \left(\frac{\sigma_{\rm tot}}{\sigma_{\rm prior}}\right)^4 \tag{A14}$$

for $\sigma_{\rm pop}$. We quoted these results in Eqs. (18) and (19) in the main text. We require at least this many measurements before the uncertainty in the population variance can be smaller than the measurement uncertainty. Until that point, the $\sigma_{\rm pop}$ posterior will be dominated by the prior.

The results for the N thresholds quoted above hinge on the specific choice of prior for μ and $\sigma_{\rm pop}$. In the main text, we justified our decision to set the scale of those priors based on $\sigma_{\rm obs}$ by noting that this is the only intrinsic scale to the problem, and should always be sufficiently broad as long as the deviation from GR is not visible in a single detection—the regime in which we are interested in the first place. Had we chosen to increase the prior variance by some factor, then the N thresholds would decrease by the same factor, i.e., we need fewer detections to gain information relative to a broad (less informative) prior than a narrower prior. Either way, the result converges to the right answer as we accumulate more observations.

[6] We carry out this calculation for $\sigma^2_{\rm pop}$ instead of $\sigma_{\rm pop}$ because the latter is irregular at the origin.

[1] J. Aasi *et al.* (LIGO Scientific Collaboration), Classical Quantum Gravity **32**, 074001 (2015), arXiv:1411.4547 [gr-qc].

[2] F. Acernese *et al.* (Virgo Collaboration), Classical Quantum Gravity **32**, 024001 (2015), arXiv:1408.3978 [gr-qc].

[3] R. Abbott *et al.* (LIGO Scientific, Virgo), Astrophys. J. Lett. **913**, L7 (2021), arXiv:2010.14533 [astro-ph.HE].

[4] R. Abbott *et al.* (LIGO Scientific, VIRGO, KAGRA), (2021), arXiv:2111.03634 [astro-ph.HE].

[5] M. Isi, K. Chatziioannou, and W. M. Farr, Phys. Rev. Lett. **123**, 121101 (2019), arXiv:1904.08011 [gr-qc].

[6] R. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. D **103**, 122002 (2021), arXiv:2010.14529 [gr-qc].

[7] R. Abbott *et al.* (LIGO Scientific, VIRGO, KAGRA), (2021), arXiv:2112.06861 [gr-qc].

[8] T. G. F. Li, W. Del Pozzo, S. Vitale, C. Van Den Broeck, M. Agathos, J. Veitch, K. Grover, T. Sidery, R. Sturani, and A. Vecchio, Phys. Rev. D **85**, 082003 (2012), arXiv:1110.0530 [gr-qc].

[9] T. G. F. Li, W. Del Pozzo, S. Vitale, C. Van Den Broeck, M. Agathos, J. Veitch, K. Grover, T. Sidery, R. Sturani, and A. Vecchio, J. Phys. Conf. Ser. **363**, 012028 (2012), arXiv:1111.5274 [gr-qc].

[10] M. Agathos, W. Del Pozzo, T. G. F. Li, C. Van Den Broeck, J. Veitch, and S. Vitale, Phys. Rev. D **89**, 082001 (2014), arXiv:1311.0420 [gr-qc].

[11] L. Sampson, N. Cornish, and N. Yunes, Phys. Rev. D **87**, 102001 (2013), arXiv:1303.1185 [gr-qc].

[12] L. Sampson, N. Cornish, and N. Yunes, Phys. Rev. D **89**, 064037 (2014), arXiv:1311.4898 [gr-qc].

[13] J. Meidam, M. Agathos, C. Van Den Broeck, J. Veitch, and B. S. Sathyaprakash, Phys. Rev. D **90**, 064009 (2014).

[14] S. Datta, K. S. Phukon, and S. Bose, Phys. Rev. D **104**, 084006 (2021), arXiv:2004.05974 [gr-qc].

[15] M. Saleem, N. V. Krishnendu, A. Ghosh, A. Gupta, W. Del Pozzo, A. Ghosh, and K. G. Arun, Phys. Rev. D **105**, 104066 (2022), arXiv:2111.04135 [gr-qc].

[16] K. Chatziioannou *et al.*, Phys. Rev. D **100**, 104015 (2019), arXiv:1903.06742 [gr-qc].

[17] M. Hübner, C. Talbot, P. D. Lasky, and E. Thrane, Phys. Rev. D **101**, 023011 (2020), arXiv:1911.12496 [astro-ph.HE].

[18] W. Del Pozzo, T. G. F. Li, M. Agathos, C. Van Den Broeck, and S. Vitale, Phys. Rev. Lett. **111**, 071101 (2013), arXiv:1307.8338 [gr-qc].

[19] K. Chatziioannou, K. Yagi, A. Klein, N. Cornish, and N. Yunes, Phys. Rev. D **92**, 104008 (2015), arXiv:1508.02062 [gr-qc].

[20] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Class. Quant. Grav. **37**, 045006 (2020), arXiv:1908.01012 [gr-qc].

[21] S. Ghosh, X. Liu, J. Creighton, I. Magaña Hernandez, W. Kastaun, and G. Pratten, (2021), arXiv:2104.08681 [gr-qc].

[22] M. Isi, M. Pitkin, and A. J. Weinstein, Phys. Rev. D **96**, 042001 (2017), arXiv:1703.07530 [gr-qc].

[23] T. Callister, A. S. Biscoveanu, N. Christensen, M. Isi, A. Matas, O. Minazzoli, T. Regimbau, M. Sakellariadou, J. Tasson, and E. Thrane, Phys. Rev. X **7**, 041058 (2017), arXiv:1704.08373 [gr-qc].

[24] X. Liu, I. M. Hernandez, and J. Creighton, Astrophys. J. **908**, 97 (2021), arXiv:2009.06539 [astro-ph.HE].

[25] R. K. L. Lo and I. Magaña Hernandez, (2021), arXiv:2104.09339 [gr-qc].

[26] G. Ashton, K. Ackley, I. M. n. Hernandez, and B. Piotrzkowski, (2020), arXiv:2009.12346 [astro-ph.HE].

[27] I. Ota and C. Chirenti, (2021), arXiv:2108.01774 [gr-qc].

[28] J. Veitch and A. Vecchio, Class. Quantum Grav. **25**, 184010 (2008), arXiv:0807.4483 [gr-qc].

[29] J. Veitch and A. Vecchio, Phys.Rev. **D81**, 062003 (2010), arXiv:0911.3820 [astro-ph.CO].

[30] N. J. Cornish and T. B. Littenberg, Classical Quantum Gravity **32**, 135012 (2015), arXiv:1410.3835 [gr-qc].

[31] M. Isi, R. Smith, S. Vitale, T. J. Massinger, J. Kanner, and A. Vajpeyi, Phys. Rev. D **98**, 042007 (2018), arXiv:1803.09783 [gr-qc].

[32] G. Ashton, E. Thrane, and R. J. E. Smith, Phys. Rev. D **100**, 123018 (2019), arXiv:1909.11872 [gr-qc].

[33] G. Pratten and A. Vecchio, Phys. Rev. D **104**, 124039 (2021), arXiv:2008.00509 [gr-qc].

[34] N. J. Cornish, T. B. Littenberg, B. Bécsy, K. Chatziioannou, J. A. Clark, S. Ghonge, and M. Millhouse, Phys. Rev. D **103**, 044006 (2021), arXiv:2011.09494 [gr-qc].

[35] J. C. Bustillo, P. D. Lasky, and E. Thrane, Phys. Rev. D **103**, 024041 (2021), arXiv:2010.01857 [gr-qc].

[36] A. Zimmerman, C.-J. Haster, and K. Chatziioannou, Phys. Rev. D **99**, 124044 (2019), arXiv:1903.11008 [astro-ph.IM].

[37] R. E. Kass and A. E. Raftery, J. Am. Statist. Assoc. **90**, 773 (1995).

[38] W. James and C. Stein, in *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* (Univ. California Press, Berkeley, Calif., 1961) pp. 361–379.

[39] D. V. Lindley and A. F. M. Smith, Journal of the Royal Statistical Society. Series B (Methodological) **34**, 1 (1972).

[40] B. Efron and C. Morris, Scientific American **236**, 119 (1977).

[41] D. B. Rubin, Journal of Educational Statistics **6**, 377 (1981).

[42] L. Blanchet and B. Sathyaprakash, Phys. Rev. Lett. **74**, 1067 (1995).

[43] L. Blanchet and B. Sathyaprakash, Class. Quant. Grav. **11**, 2807 (1994).

[44] K. G. Arun, B. R. Iyer, M. S. S. Qusailah, and B. S. Sathyaprakash, Phys. Rev. D **74**, 024006 (2006), arXiv:gr-qc/0604067.

[45] K. G. Arun, B. R. Iyer, M. S. S. Qusailah, and B. S. Sathyaprakash, Classical Quantum Gravity **23**, L37 (2006), arXiv:gr-qc/0604018.

[46] N. Yunes and F. Pretorius, Phys. Rev. D **80**, 122003 (2009), arXiv:0909.3328 [gr-qc].

[47] C. K. Mishra, K. G. Arun, B. R. Iyer, and B. S. Sathyaprakash, Phys. Rev. D **82**, 064010 (2010), arXiv:1005.0304 [gr-qc].

[48] T. Li, W. Del Pozzo, S. Vitale, C. Van Den Broeck, M. Agathos, *et al.*, Phys.Rev. **D85**, 082003 (2012), arXiv:1110.0530 [gr-qc].

[49] T. G. F. Li, W. Del Pozzo, S. Vitale, C. Van Den Broeck, M. Agathos, J. Veitch, K. Grover, T. Sidery, R. Sturani, and A. Vecchio, *Gravitational waves. Numerical relativity - data analysis. Proceedings, 9th Edoardo Amaldi Conference, Amaldi 9, and meeting, NRDA 2011, Cardiff, UK, July 10-15, 2011*, J. Phys. Conf. Ser. **363**, 012028 (2012), arXiv:1111.5274 [gr-qc].

[50] M. Agathos, W. Del Pozzo, T. G. F. Li, C. Van Den Broeck, J. Veitch, and S. Vitale, Phys. Rev. D **89**, 082001 (2014), arXiv:1311.0420 [gr-qc].

[51] J. M. Dickey, The Annals of Mathematical Statistics **42**, 204 (1971).

[52] I. Verdinelli and L. Wasserman, Journal of the American Statistical Association **90**, 614 (1995).

[53] S. Perkins and N. Yunes, (2022), arXiv:2201.02542 [gr-qc].

[54] S. Lotfi, P. Izmailov, G. Benton, M. Goldblum, and A. G. Wilson, arXiv e-prints , arXiv:2202.11678 (2022), arXiv:2202.11678 [cs.LG].

[55] D. Laghi, G. Carullo, J. Veitch, and W. Del Pozzo, Classical and Quantum Gravity **38**, 095005 (2021), arXiv:2011.03816 [gr-qc].

[56] J. D. Hunter, Computing In Science & Engineering **9**, 90 (2007).

[57] J. Bezanson, S. Karpinski, V. B. Shah, and A. Edelman, arXiv e-prints (2012), 10.48550/arXiv.1209.5145, arXiv:1209.5145 [cs.PL].
[58] H. Ge, K. Xu, and Z. Ghahramani, in *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain* (2018) pp. 1682–1690.
[59] T. Breloff, "Plots.jl," (2022).