A User Manual for the Fisher Information Matrix

Michele Vallisneri
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109
(Dated: Feb 9 2007)

The Fisher-matrix formalism is used routinely in the literature on gravitational-wave detection to characterize the parameter-estimation performance of gravitational-wave measurements, given parametrized models of the waveforms, and assuming detector noise of known colored Gaussian distribution. Unfortunately, the Fisher matrix can be a poor predictor of the amount of information obtained from typical observations, especially for waveforms with several parameters and relatively low expected signal-to-noise ratios, or for waveforms depending weakly on one or more parameters, when their priors are not taken into proper consideration. I discuss these pitfalls and describe practical recipes to recognize them and cope with them.

Specifically, I answer the following questions: (i) What is the significance of (quasi-)singular Fisher matrices, and how do we deal with them? (ii) When is it necessary to take into account prior probability distributions for the source parameters? (iii) When is the signal-to-noise ratio high enough to believe the Fisher-matrix result?

PACS numbers: 04.80.Nn, 95.55.Ym, 02.50.Tt

I. INTRODUCTION

Over the last two decades, the prevailing attitude in the gravitational-wave (GW) source-modeling community has been one of pre-data positioning: in the absence of confirmed detections, the emphasis has been on exploring which astrophysical systems, and which of their properties, would become accessible to GW observations with the sensitivities afforded by planned (or desired) future experiments, with the purpose of committing theoretical effort to the most promising sources, and of directing public advocacy to the most promising detectors. In this positioning and in this exploration, the expected accuracy of GW source parameters, as determined from the signals to be observed, is often employed as a proxy for the amount of physical information that can be gained from detection campaigns. However, predicting the parameter-estimation performance of future observations is a complex matter, even with the benefit of accurate theoretical descriptions of the expected waveforms and of faithful characterizations of the noise and response of detectors; and in practice, the typical source modeler has had much less to go with. The main problem is that there are few analytical tools that can be applied generally to the problem, before resorting to relatively cumbersome numerical simulations that involve multiple explicit realizations of signal-plus-noise datasets.

In the source-modeling community, the analytical tool of choice has been the Fisher information matrix $F_{ij}[h] = (h_i,h_j)$: here $h_i(t)$ is the derivative of the gravitational waveform $h(t)$ of interest with respect to the $i$-th source parameter $\theta_i$, and $\langle \cdot, \cdot \rangle$ is a signal product weighted by the expected power spectral density of detector noise, as described in Sec. II B. Now, it is usually claimed that the inverse Fisher matrix $F^{-1}_{ij}[h_0]$ represents the covariance matrix of parameter errors in the parameter-estimation problem for the true signal $h_0$. This statement can be interpreted in three slightly different ways (all correct), which we examine in detail in Sec. II, and preview here:

1. The inverse Fisher matrix $F^{-1}_{ij}[h_0]$ is a lower bound (generally known as the Cramér–Rao bound) for the error covariance of any unbiased estimator of the true source parameters. Thus, it is a frequentist error (see Sec. II A): for any experiment characterized by the true signal $h_0$ and a certain realization $n$ of detector noise, the parameter estimator $\hat{\theta}$ is a single-valued function of the total detector output $s = n + h_0$, and $F^{-1}_{ij}[h_0]$ is a lower bound on the covariance (i.e., the fluctuations) of $\hat{\theta}$ in an imaginary infinite sequence of experiments with different realizations of noise. The Cramér–Rao bound is discussed in Sec. II C.

2. The inverse Fisher matrix $F^{-1}_{ij}[h_0]$ is the frequentist error covariance for the maximum-likelihood (ML) parameter estimator $\hat{\theta}^{\text{ML}}$, assuming Gaussian noise, in the limit of strong signals (i.e., high signal-to-noise ratio SNR) or, equivalently, in the limit in which the waveforms can be considered as linear functions of their parameters (we shall refer to this limit as the linearized-signal approximation, or LSA). This well-known result is rederived in Sec. II D.

3. The inverse Fisher matrix $F^{-1}_{ij}[h_0]$ represents the covariance (i.e., the multidimensional spread around the mode) of the posterior probability distribution $p(\theta_0|s)$ for the true source parameters $\theta_0$, as it can be inferred (in Bayesian fashion) from a single experiment with true signal $h_0$, assuming Gaussian noise, in the high-SNR limit (or in the LSA), and in the case where any prior probabilities for the parameters are constant over the
parameter range of interest. Properly speaking, the inverse Fisher matrix is a measure of uncertainty rather than error, since in any experiment the mode will be displaced from the true parameters by an unknown amount due to noise.\footnote{In the LSA/high-SNR limit with negligible priors, the posterior probability mode, seen as a frequentist statistic, coincides with the ML estimator; thus its fluctuations are again described by the inverse Fisher matrix.} See Sec. II E for a rederivation of this result.

As pointed out by Jaynes \cite{1}, while the numerical identity of these three different error-like quantities has given rise to much confusion, it arises almost trivially from the fact that in a neighborhood of its maximum, the signal likelihood $p(s|\theta_0)$ is approximated by a normal probability distribution with covariance $F^{-1}_{ij}$. In this paper, I argue that the Cramér–Rao bound is seldom useful in the work of GW analysts (Sec. II C), and while the high-SNR/LSA frequentist and Bayesian results are legitimate, they raise the question of whether the signals of interest are strong (or linear) enough to warrant the limit, and of what happens if they are not. In addition, if we possess significant information about the prior distribution (or even the allowed range) of source parameters, it is really only in the Bayesian framework that we can fold this information reliably into the Fisher result (Sec. II D).

Thus, I recommend the Bayesian viewpoint as the most fruitful way of thinking about the Fisher-matrix result (although I will also derive parallel results from the frequentist viewpoint). Of course, the study of Bayesian inference for GW parameter-estimation problems need not stop at the leading-order (Fisher-matrix) expression for the posterior likelihood: Markov Chain Monte Carlo (MCMC) algorithms \cite{2} can provide very reliable results, immune from any considerations about signal strength, but they require a significant investment of time to implement them, and of computational resources to run them. By contrast, the Fisher-matrix formalism is singularly economical (especially since it does not involve explicit realizations of the noise, as MCMC methods must necessarily do), and it is clear that it will always be the first recourse of the GW analyst. To use it reliably, however, we must understand the limits of its applicability. The purpose of this paper is to explore these limits. I do so by addressing three questions:

1. What is the significance of singular or ill-conditioned Fisher matrix that often appear in estimation problems with several source parameters, and how do we deal with them? Can we still believe the Fisher result? (See Sec. IV.)

2. When is it necessary to take into account the prior probability distributions for the parameters, perhaps specified trivially by giving their allowed ranges? (See Sec. V.)

3. When is the high-SNR/LSA approximation warranted? (As we shall see, the high-SNR limit is equivalent to the LSA.) That is, how strong a signal will we need to measure if we are to believe the Fisher-matrix result as its uncertainty? (See Sec. VI.)

In addition, I discuss the extension of the LSA beyond the leading order, in both the frequentist and Bayesian parameter-estimation frameworks (Sec. VII), in a form that the adventurous GW analyst can use to test the reliability of the Fisher result (but higher-order derivatives and many-indexed expressions start to mount rapidly, even at the next-to-leading order).

To be sure, these questions were posed already in the seminal treatments of GW detection by Finn \cite{3} and Cutler and Flanagan \cite{4}, but it appears that the cautions suggested by these authors have gone largely unheeded. My treatment follows closely theirs, as well as the classic texts on the statistical analysis of noisy data (e.g., Refs. \cite{5, 6}). I am indebted to Jaynes and Bretthorst \cite{1, 7} for their enlightening (if occasionally blunt) perspective on frequentist and Bayesian parameter estimation. While most of this paper could be described fairly as a critical review of well-established knowledge, as far as I know the procedures and viewpoints described in Secs. V and VI are original, and Sec. IV is offered to help clarify an often-encountered but little-understood situation; last, although higher-order expressions for the variance appear in Ref. \cite{4} and elsewhere, I consider it useful to restate them in a notation compatible with the rest of this paper.

Whenever my discussion requires a practical example, I consider signals from inspiraling binaries of two black holes, both of mass $10M_\odot$, as described by the restricted post-Newtonian approximation for adiabatic, circular inspirals (see Sec. III); in my examples, I assume detection and parameter estimation are performed on Initial-LIGO \cite{8} data, and I adopt the LIGO noise curve of Ref. \cite[Table IV]{9}. Throughout, I use geometric units; I assume the Einstein summation convention for repeated indices; and I do not distinguish between covariant and contravariant indices, except in Sec. VII.
II. THREE ROADS TO THE FISHER MATRIX

In this section I discuss the “three roads” to the inverse Fisher matrix as a measure of uncertainty for GW observations: the Cramér-Rao bound (Sec. II C), the high-SNR/LSA limit for the frequentist covariance of the ML estimator (Sec. II D), and the high-SNR/LSA limit for the single-experiment covariance of the Bayesian posterior distribution (Sec. II E). Sections II A and II B are refreshers about frequentist and Bayesian parameter estimation, and about the analytical expression of Gaussian-noise likelihood for GW signals.

A. A refresher on the frequentist and Bayesian frameworks

The frequentist (or orthodox) approach to parameter estimation for GW signals can be summed up as follows:

1. We are given the detector data \( s \), consisting of the true signal \( h_0 = h(\theta_0) \) (where \( \theta_0 \) is the vector of the true system parameters) plus additive noise \( n \).

2. We select a point estimator \( \hat{\theta}(s) \): that is, a function of detector data that (it is hoped) approximates the true values of source parameters, except for the statistical error due to the presence of noise. One important example of point estimator is the ML estimator \( \hat{\theta}_{\text{ML}} \), which maximizes the likelihood \( p(s|\theta) \) of observing the measured data \( s \) given a value \( \theta \) of the true parameters. For additive noise, this likelihood coincides with the probability of a noise realization \( n = s - h(\theta) \).

3. We characterize statistical error as the fluctuations of \( \hat{\theta}(s) \), computed over a very long series of independent experiments where the source parameters are kept fixed, while detector noise \( n \) is sampled from its assumed probability distribution (often called sampling distribution).

The estimator \( \hat{\theta} \) is usually chosen according to one or more criteria of optimality: for instance, unbiasedness requires that \( \langle \hat{\theta}(s) \rangle_n \) (the average of the estimator over the noise probability distribution) be equal to \( \theta_0 \).

A rather different approach is that of Bayesian inference:

1. We do not assume the true value of the system parameters, but we posit their prior probability distribution \( p(\theta) \).

2. Given the data \( s \), we do not compute estimators, but rather the full posterior probability distribution \( p(\theta|s) \), using Bayes’ theorem \( p(\theta|s) = p(s|\theta) \times p(\theta)/p(s) \), where \( p(s) = \int p(s|\theta) p(\theta) d\theta \).

3. We characterize statistical error in a single experiment by the spread of the posterior distribution \( p(\theta|s) \).

The differences between the frequentist and Bayesian approaches are not only mathematical, but also epistemic: “frequentists” view probabilities essentially as the relative frequencies of outcomes in repeated experiments, while “Bayesians” view them as subjective indexes of certainty for alternative propositions. For an introduction to the contrasting views, I refer the reader to the excellent treatise (very partial to the Bayesian worldview) by Jaynes [1], and to Ref. [4] for a more GW-detection-oriented discussion.

Once actual detections are made, the Bayesian approach of computing posterior probability distributions for the signal parameters given the observed data seems more powerful than the frequentist usage of somewhat arbitrary point estimators; the latter will always result in throwing away useful information, unless the chosen estimators are sufficient statistics (i.e., unless the likelihood depends on the data only through the estimators). As for statistical error, it seems preferable to characterize it from the data we have (actually, from the posterior distributions that we infer from the data), rather than from the data we could have obtained (i.e., from the sampling distribution of estimators in a hypothetical ensemble of experiments).

As Cutler and Flanagan [4] point out, however, it is in the current pre-data regime that we seek to compute expected parameter accuracies; in the absence of actual confirmed-detection datasets, it seems acceptable to consider ensembles of possible parameter-estimation experiments, and to use frequentist statistical error as an inverse measure of potential physical insight. The best solution, bridging the two approaches, would undoubtedly be to examine the frequentist distribution of some definite measure of Bayesian statistical error; unfortunately, such a hybrid study is generally unfeasible, given the considerable computational requirements of even single-dataset Bayesian analyses.

---

2 Only in the sense that subjects with different prior assumptions could come to different conclusions after seeing the same data; indeed, Bayesian statistics describes how prior assumptions become deterministically modified by the observation of data.
B. Likelihood for GW signals in Gaussian noise

Under the assumption of stationary and Gaussian detector noise, the likelihood \( \log p(s|\theta) \) can be obtained very simply from a noise-weighted inner product of the detector output and of the signal \( h(\theta) \) (see for instance Ref. [4, Eq. (2.3)]):

\[
p(s|\theta) \propto e^{-(s-h(\theta),s-h(\theta))/2};
\]

(1)

the weighting is performed with respect to the expected power spectral density of detector noise by defining the noise-weighted inner product of two real-valued signals as

\[
(h, g) = 4 \text{Re} \int_0^{+\infty} \frac{\hat{h}(f)^* \hat{g}(f)}{S_n(f)} df,
\]

(2)

where \( \hat{h}(f) \) and \( \hat{g}(f) \) are the Fourier transforms of \( h(t) \) and \( g(t) \), \( \cdot^* \) denotes complex conjugation, and \( S_n(f) \) is the one-sided power spectral density of the noise. From the definition of \( S_n(f) \) as \( \langle \hat{n}(f)^* \hat{n}(f') \rangle_n = \frac{1}{2} S_n(|f|) \delta(f-f') \), we get the useful property

\[
\langle (h, n)(n, g) \rangle_n = (h, g),
\]

(3)

where again \( \langle \cdot \rangle_n \) denotes averaging over the probability distribution of the noise.

C. First road: Derivation and critique of the Cramér–Rao bound

The derivation in this section is inspired by the treatment of Ref. [1, p. 518], and it is given for simplicity in the case of one source parameter. We wish to pose a bound on the frequentist estimator variance

\[
\text{var} \theta = \left\langle \left( \hat{\theta}(s) - \langle \hat{\theta}(s) \rangle_n \right)^2 \right\rangle_n;
\]

(4)

to do this, we consider the ensemble product

\[
\langle u(s), v(s) \rangle_n = \int u(s) v(s) p(s|\theta_0) ds,
\]

(5)

where \( p(s|\theta_0) \) is the likelihood of observing the detector output \( s \) given the true source parameter \( \theta_0 \), or equivalently the likelihood of observing the noise realization \( n = s - h_0 \). Setting \( v(s) = \hat{\theta}(s) - \langle \hat{\theta}(s) \rangle_n \), we obtain a bound on \( \langle v, v \rangle_n \equiv \text{var} \hat{\theta} \) from the Schwarz inequality:

\[
\text{var} \hat{\theta} = \langle v, v \rangle_n \geq \frac{\langle u, v \rangle_n^2}{\langle u, u \rangle_n}.
\]

(6)

This inequality is true for any function \( u(s) \) of the data, and it becomes an equality when \( u(s) \propto v(s) \). Since we wish to derive a bound that applies generally to all estimators, we should not have (or try) to provide too much detail about \( \hat{\theta} \) (and therefore \( v(s) \)). A simple assumption to make on \( \hat{\theta} \) is that it is an unbiased estimator:

\[
\langle \hat{\theta}(s) \rangle_n = \theta_0 \Rightarrow \partial_{\theta_0} \langle \hat{\theta}(s) \rangle_n = 1.
\]

(7)

How does this help us? It turns out that we can write a function \( u(s) \) whose ensemble product with any other function \( w(s) \) yields the derivative \( \partial_{\theta_0} \langle w(s) \rangle_n \); this function is just \( u(s) = \partial_{\theta_0} \log p(s|\theta_0) \), because

\[
\int w(s)[\partial_{\theta_0} \log p(s|\theta_0)] p(s|\theta_0) ds = \int w(s) \partial_{\theta_0} p(s|\theta_0) ds = \partial_{\theta_0} \int w(s) p(s|\theta_0) ds = \partial_{\theta_0} \langle w(s) \rangle_n,
\]

(8)

assuming of course\(^3\) that we can exchange integration and differentiation with respect to \( \theta_0 \). For any \( s \), \( u(s) \) encodes the local relative change in the likelihood function as \( \theta \) is changed. It follows that \( \langle u(s), v(s) \rangle_n = \partial_{\theta_0} \langle \hat{\theta}(s) \rangle_n = 1 \), so

\(^3\) This assumption fails for some (mildly) pathological likelihood functions, which can provide counterexamples to the Cramér–Rao bound.
from Eq. (6) we get

$$\text{var} \hat{\theta} \geq \frac{1}{\langle \hat{u}(s), \hat{u}(s) \rangle} = \frac{1}{\langle \partial \log p(s|\theta_0), \partial \log p(s|\theta_0) \rangle_n},$$

(9)

which is the unbiased-estimator version of the Cramér–Rao bound. If the estimator is biased, we can still use the Schwarz inequality by providing the derivative of the bias $b(\theta_0)$ with respect to $\theta_0$:

$$\langle \hat{\theta}(s) \rangle_n = \theta_0 + b(\theta_0) \quad \Rightarrow \quad \partial \theta_n \langle \hat{\theta}(s) \rangle_n = 1 + \partial \theta_n b(\theta_0),$$

(10)

and therefore

$$\text{var} \hat{\theta} \geq \frac{(1 + \partial \theta_n b)^2}{\langle \partial \log p(s|\theta_0), \partial \log p(s|\theta_0) \rangle_n}. \quad \text{(11)}$$

Generalizing to a multidimensional expression is straightforward, if verbose:

$$\text{cov}_{ij}(\hat{\theta}_i, \hat{\theta}_j) \geq (\delta_{im} + \partial_m b_i(\theta_0)) F^{-1}_{ij} (\delta_{jl} + \partial_j b_l(\theta_0)),$$

(12)

where the Fisher information matrix is defined by

$$F_{ij} = \left\langle \left( \partial_i \log p(s|\theta_0) \right) \left( \partial_j \log p(s|\theta_0) \right) \right\rangle_n = -\left\langle \partial_i \partial_j \log p(s|\theta) \right\rangle_n.$$

The second equality is established by taking the gradient of $\int (\partial_i \log p(s|\theta_0)) p(s|\theta_0) \, ds$, and remembering that $\partial_i \int p(s|\theta_0) \, ds = \partial_i 1 = 0$. With the help of Eqs. (1) and (3), we can compute the Fisher matrix for GW signals in additive Gaussian noise, which is the familiar expression $F_{ij} = (h_i, h_j)$.

The full expression (12) for the Cramér–Rao bound, which includes the effects of bias, has interesting consequences, for it implies that biased estimators can actually outperform unbiased estimators, since the $\partial_n b_i(\theta_0)$ can be negative. Unfortunately, we have no handle on these derivatives without explicitly choosing a particular estimator (which goes against the idea of having a generic bound), so the Cramér–Rao bound can only give us a definite result for the subclass of unbiased estimators.

As pointed out by Cutler and Flanagan [4, App. A 5], it follows that the bound cannot be used to place absolute limits on the accuracy of estimators (i.e., lower bounds on frequentist error), which would exclude or severely limit the possibility of inferring the physical properties of sources from their emitted GWs. Even if the lower bound for unbiased estimators is very discouraging, there is always a chance that a biased estimator could do much better, so we cannot use the bound to prove “no go” theorems.

Going back to Eq. (6), we note that the bound is satisfied as an equality when

$$u(s) \propto v(s) \quad \Rightarrow \quad \partial \theta_n \log p(s|\theta_0) = q(\theta_0) [\hat{\theta}(s) - \langle \hat{\theta}(s) \rangle_n].$$

(14)

By integrating, we obtain a relation between the likelihood and the estimator:

$$p(s|\theta_0) = \frac{m(s)}{Z(\theta_0)} e^{-l(\theta_0)\hat{\theta}(s)}; \quad \text{(15)}$$

the estimation problems (i.e., the pairings of given likelihoods and chosen estimators) for which this relation holds true are said to belong to the exponential family, and they are the only ones for which the Cramér–Rao bound is satisfied exactly as an equality. Equation (15) generalizes trivially to multidimensional problems by replacing the exponential with $\exp \{-l_k(\theta_0)\hat{\theta}(s)\}$. Unfortunately, for a given $p(s|\theta_0)$ there is no guarantee that any unbiased estimator exists that satisfies Eq. (15) and that therefore can actually achieve the bound; all we can say in general

---

4 To obtain Eq. (9), we need to notice also that for any $w(s)$, $\langle u(s), v(s) \rangle_n = 0$, since $\langle w(s) \rangle_n$ does not depend on $s$ (but only on $\theta_0$), and the integral of Eq. (8) reduces to $\langle u(s) \rangle_n \int \partial \theta_n p(s|\theta_0) \, ds = \langle w(s) \rangle_n \theta_0 = 1 = 0$.

5 This is true even if we evaluate the performance of estimators on the basis of their quadratic error

$$\langle (\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0) \rangle_n \geq b_i(\theta_0) b_j(\theta_0) + (\delta_{im} + \partial_m b_i(\theta_0)) F^{-1}_{ij} (\delta_{jl} + \partial_j b_l(\theta_0))$$

rather than on the basis of their variance.
about the performance of unbiased estimators is that they will underperform the Cramér–Rao bias, but we do not know how badly. As discussed above, the bound tells us nothing in general about biased estimators.

It follows that the bound cannot be used to establish guaranteed levels of accuracy (i.e., upper bounds on frequentist error), which would prove the possibility of inferring the physical properties of sources from their GWs. We can only do so if we can identify a specific estimator that achieves the bound. In the next section we shall see that the ML estimator does so in the high-SNR limit, where waveforms can be approximated accurately as linear functions of their parameters within the region of parameter space where \( p(s|\theta) \) is not negligible (so the high-SNR limit coincides with the limit in which the LSA is accurate).

We conclude that the Cramér–Rao bound is never useful to the GW analyst as a proper bound, whether to make positive or negative expected-accuracy statements; where it is useful, it reduces to the high-SNR/LSA result for the ML estimator.

### D. Second road: Derivation and critique of the frequentist high-SNR/LSA result

We denote the true signal as \( h_0 \) (so \( s = h_0 + n \)), and expand the generic waveform \( h(\theta) \) around \( h_0 \), normalizing signals by the optimal signal-to-noise ratio of the true signal, \( A = \sqrt{(h_0, h_0)} \) (also known in this context as signal strength):

\[
h(\theta) = h_0 + \theta_k h_k + \theta_j \theta_k h_{jk} / 2 + \cdots = A(h_0 + \theta_k \bar{h}_k + \theta_j \theta_k \bar{h}_{jk} / 2 + \cdots);
\]

(16)

here we are translating source parameters as needed to have \( h(0) = h_0 \), defining \( h_i = \partial_i h|_{\theta=0} \), \( h_{ij} = \partial_{ij} h|_{\theta=0} \) (and so on), and \( h_0 = h_0/A, \ h_k = h_k/A \) (and so on).\(^7\) The likelihood is then given by

\[
p(s|\theta) \propto e^{-(s-h(\theta), s-h(\theta))/2} = \exp \left\{ - (n, n)/2 - A^2 \left[ \theta_j \theta_k (\bar{h}_j, \bar{h}_k) + \theta_j \theta_k \theta_l (\bar{h}_j, \bar{h}_{kl}) + \cdots \right] / 2 \right. \\
+ \left. A (\theta_j n, \bar{h}_j) + \theta_j \theta_k (n, \bar{h}_j + \bar{h}_k) / 2 + \theta_j \theta_k \theta_l (n, \bar{h}_{kl}) / 3! + \cdots \right\}.
\]

(17)

The ML equations \( \partial_j p(s|\theta^{\text{ML}}) = 0 \) are given by

\[
0 = \frac{1}{A} \left[ (n, \bar{h}_j) + \hat{\theta}_k (n, \bar{h}_{jk}) + \hat{\theta}_l \hat{\theta}_k (n, \bar{h}_{kl}) / 2 + \cdots \right] - \left[ \hat{\theta}_k (\bar{h}_j, \bar{h}_k) + \hat{\theta}_l \hat{\theta}_k (\bar{h}_j, \bar{h}_{kl}) / 2 + (\bar{h}_k, \bar{h}_{jl}) + \cdots \right],
\]

(18)

where we have divided everything by \( A^2 \), and we omit the \(^{\text{ML}}\) superscript for conciseness. A careful study of Eq. (18) shows that it can be solved in perturbative fashion by writing \( \hat{\theta}_j^{\text{ML}} \) as a series in \( 1/A \),

\[
\hat{\theta}_j^{\text{ML}} = \hat{\theta}_j^{(1)} / A + \hat{\theta}_j^{(2)} / A^2 + \hat{\theta}_j^{(3)} / A^3 + \cdots,
\]

(19)

and by collecting the terms of the same order in Eq. (18),

\[
O(1/A) : \ (n, \bar{h}_j) - \hat{\theta}_k^{(1)} (\bar{h}_j, \bar{h}_k) = 0,
\]

\[
O(1/A^2) : \ \hat{\theta}_k^{(1)} (n, \bar{h}_{jk}) - \hat{\theta}_k^{(1)} \hat{\theta}_l^{(1)} (\bar{h}_j, \bar{h}_{kl}) / 2 + (\bar{h}_k, \bar{h}_{jl}) - \hat{\theta}_k^{(2)} (\bar{h}_j, \bar{h}_k) = 0,
\]

(20)

\[
O(1/A^3) : \ 
\]

thus the ML solution \( \hat{\theta}_j^{\text{ML}} \) is given by

\[
\hat{\theta}_j^{\text{ML}} = \frac{1}{A} (\bar{h}_j, \bar{h}_k)^{-1} (\bar{h}_k, n) + \\
\frac{1}{A^2} (\bar{h}_j, \bar{h}_k)^{-1} \left\{ \left( n, \bar{h}_{ik} (\bar{h}_k, \bar{h}_l)^{-1} (\bar{h}_l, n) - (\bar{h}_i, \bar{h}_{kl}) / 2 + (\bar{h}_k, \bar{h}_{il}) \right) (\bar{h}_k, \bar{h}_m)^{-1} (\bar{h}_m, n) (\bar{h}_i, \bar{h}_n)^{-1} (\bar{h}_n, n) \right\} + \cdots
\]

(21)

\(^6\) Indeed, Eq. (15) implies that both an efficient (i.e., bound-achieving) unbiased estimator and the ML estimator exist, they must coincide. To show this, we notice that if the ML estimator exists, the log-derivative \( \partial \log p(s|\theta) = -\partial_l h_k (\theta) \partial \bar{\theta}_k - \partial_k \) must be zero at \( \theta = \hat{\theta}^{\text{ML}} \), from which it follows that \( \hat{\theta}_k = \hat{\theta}_j^{\text{ML}} \).

\(^7\) The statistical uncertainty in the estimated signal strength can still be handled in this notation by taking one of the \( \bar{h}_k \) to lie along \( h_0 \); the corresponding \( \hat{\theta}_k \) represents a fractional correction to the true \( A \).

\(^8\) Formally, it is troubling to truncate the series expression for the exponent at any order beyond quadratic, since the integral of the truncated likelihood may become infinite; the important thing to keep in mind, however, is that the series need converge only within a limited parameter range determined self-consistently by the truncated-likelihood estimator, by compact parameter ranges, or (in the Bayesian case) by parameter priors. Similar considerations apply to the derivation of the higher-order corrections given in Sec. VII.
Thus we see that the limit of large $A$ (i.e., high SNR) coincides with the linearized-signal approximation (LSA) where only the first derivatives of the signals are included. In the LSA, the likelihood is just

$$p(s|\theta) \propto \exp\left\{-\frac{(n,n)}{2} - \theta_j \theta_k (h_j,h_k)/2 + \theta_j (h_j,n)\right\} = \exp\left\{-\frac{(n,n)}{2} - A^2 \theta_j \theta_k (\bar{h}_j,\bar{h}_k)/2 + A \theta_j (h_j,n)\right\} \quad \text{(LSA)},$$

and the ML estimator is given by

$$\hat{\theta}_{j}^{\text{ML}} = \left(1/A\right)(\bar{h}_j,\bar{h}_k)^{-1}(\bar{h}_i,n) \quad \text{(LSA)}.$$  \hspace{1cm} (23)

Since $(\langle \bar{h}_k,n \rangle)_n = 0$, we see also that the ML estimator is unbiased. The variance of $\hat{\theta}_{j}^{\text{ML}}$ is then obtained by averaging $\hat{\theta}_{j}^{\text{ML}} \hat{\theta}_{k}^{\text{ML}}$ over noise realizations,

$$\langle \hat{\theta}_{j}^{\text{ML}} \hat{\theta}_{k}^{\text{ML}} \rangle_n = \frac{1}{A^2} (\bar{h}_j,\bar{h}_i)^{-1} \langle (\bar{h}_i,n)(\bar{h}_m,n) \rangle_n (\bar{h}_m,\bar{h}_k)^{-1} = \frac{1}{A^2} (\bar{h}_j,\bar{h}_i)^{-1} (\bar{h}_i,\bar{h}_m)(\bar{h}_m,\bar{h}_k)^{-1} = \frac{1}{A^2} (\bar{h}_j,\bar{h}_i)^{-1} \quad \text{(LSA)},$$

and it coincides with the mean quadratic error in the frequentist sense. In Eq. (24), the second equality follows from Eq. (3). The interpretation of the limit is that, for strong signals, the typical $\hat{\theta}_{j}^{\text{ML}} - \theta_0 j$ becomes small enough that the log-likelihood is accurately described by the product of detector data and a linearized signal.

Equation (24) is the standard Fisher-information–matrix result, and it implies that in the high-SNR/LSA limit the Cramér–Rao bound cannot be achieved. It is possible (but perhaps not desirable, as we shall see shortly) to modify the ML estimator to take into account prior information, known or not. Equation (24) is the standard Fisher-information–matrix result, and it implies that in the high-SNR/LSA limit the Cramér–Rao bound cannot be achieved. In that limit $\hat{\theta}_{j}^{\text{ML}}$ coincides with the mode of the posterior probability $p(\theta|s)$. Instead, we happen to be using an estimator that takes into account prior information, which enters into the Cramér–Rao bound via the derivative of the bias.

$$\langle \hat{\theta}_{j}^{\text{ML}} \hat{\theta}_{k}^{\text{ML}} \rangle_n = \langle \hat{\theta}_{j}^{\text{ML}} \rangle_n \langle \hat{\theta}_{k}^{\text{ML}} \rangle_n = \langle \hat{\theta}_{j}^{\text{ML}} \rangle_n \langle \hat{\theta}_{k}^{\text{ML}} \rangle_n \quad \text{(LSA)}.$$  \hspace{1cm} (28)

which coincides with the generalized Cramér–Rao bound of Eq. (12), proving that the estimation problem defined by the LSA likelihood and $\hat{\theta}_{j}^{\text{ML}}$ belongs to the exponential family.

\footnote{Note that the Fisher matrix that must be substituted into Eq. (12) is still $\langle \hat{\theta}_{i} \hat{\theta}_{k} p(\theta|s) \rangle_n = (h_j,h_k)$, and not $\langle \hat{\theta}_{i} \hat{\theta}_{k} p(\theta|s) p(\theta) \rangle_n = (h_j,h_k) + P_{jk}$. The prior distribution does not concern the Cramér–Rao bound, which is computed from the likelihood alone for a fixed known value of the true source parameters. Instead, we happen to be using an estimator that takes into account prior information, which enters into the Cramér–Rao bound via the derivative of the bias.}
The reason why $\hat{\theta}^{MP}$ is not too useful to characterize future parameter-estimation performance is that we expect a reasonable measure of error to converge to the effective width of the prior in the limit of vanishing signal strength. Instead, in the absence of any information from the experiment, $\hat{\theta}^{MP}$ becomes stuck at the mode of the prior, and its variance [in Eq. (28)] tends to zero. This behavior occurs for any nonuniform prior.10

E. Third-road Derivation of the Bayesian high-SNR/LSA result

We now wish to show that in any single experiment, if the high-SNR/LSA limit is warranted (and if the parameter priors are uniform over the parameter region of interest), the inverse Fisher-information matrix yields the variance of the Bayesian posterior probability distribution. To do so, we rewrite Eq. (18) in terms of normalized parameters $\bar{A}\theta_i = A\theta_i$.

$$p(s|\theta) \propto \exp\left\{-(n, n)/2 + \left[n, \bar{h}_j\right] \bar{A} \theta_j \bar{A} \theta_k / 2 + \frac{1}{A^2} n, \bar{h}_{jkl} \bar{A} \theta_j \bar{A} \theta_k / 3! + O(1/A^3)\right\}$$

$$- \left[(\bar{h}_j, \bar{h}_k) \bar{A} \theta_j \bar{A} \theta_k + \frac{1}{A} (\bar{h}_{jkl}, \bar{h}_{kim}) \bar{A} \theta_j \bar{A} \theta_k \bar{A} \theta_m / 4 + \frac{2}{A^2} (\bar{h}_{jklm}) \bar{A} \theta_j \bar{A} \theta_k \bar{A} \theta_m / 3! + O(1/A^3)\right]/2. \tag{29}$$

We can build the variance from the posterior mean

$$\langle \hat{\theta}_i \rangle_p = \int \bar{A} \theta_i p(s|\theta)\, d\theta / \int p(s|\theta)\, d\theta \tag{30}$$

and the quadratic moment

$$\langle \bar{A} \theta_i \bar{A} \theta_j \rangle_p = \int \bar{A} \theta_i \bar{A} \theta_j p(s|\theta)\, d\theta / \int p(s|\theta)\, d\theta \tag{31}$$

where “$\langle \cdot \rangle_p$” denotes integration over $p(s|\theta)$. The idea is to proceed in perturbative fashion, writing the moments as series in $\epsilon = 1/A$: taking $\langle \hat{\theta}_i \rangle_p$ as an example,

$$\langle \hat{\theta}_i \rangle_p = \sum_{n=0}^\infty \frac{\epsilon^n}{n!} \langle \hat{\theta}_i \rangle_p^{(n)} \quad \Rightarrow \quad \langle \hat{\theta}_i \rangle_p^{(n)} = \frac{\partial^n \langle \hat{\theta}_i \rangle_p}{\partial \epsilon^n} \bigg|_{\epsilon=0}. \tag{32}$$

Since $\epsilon$ appears at both the numerator and denominator of Eq. (30), we write

$$\langle \hat{\theta}_i \rangle_p = \int \bar{A} \theta_i p(0)\, d\theta / \int p(0)\, d\theta + \epsilon \int \bar{A} \theta_i \frac{\partial p(0)}{\partial \epsilon}\, d\theta + \frac{\epsilon^2}{2} \int \bar{A} \theta_i \frac{\partial^2 p(0)}{\partial \epsilon^2}\, d\theta + \cdots \tag{33}$$

(where the argument of $p$ implies that the $(n)$-th derivative is evaluated at $\epsilon = 0$), and therefore

$$\langle \hat{\theta}_i \rangle_p^{(0)} = \int \bar{A} \theta_i p(0)\, d\theta / \int p(0)\, d\theta,$n
$$\langle \hat{\theta}_i \rangle_p^{(1)} = \int \bar{A} \theta_i \frac{\partial p(0)}{\partial \epsilon} d\theta - \int \bar{A} \theta_i p(0)\, d\theta \times \int \frac{\partial p(0)}{\partial \epsilon} d\theta / \int p(\epsilon = 0)\, d\theta; \tag{34}$$

similar expressions hold for $\langle \bar{A} \theta_i \bar{A} \theta_j \rangle_p$, and a general expression for the $(n)$-th-order contribution is given in Sec. VII B.

The $\epsilon \rightarrow 0$ limit coincides with the limit of large signal strengths, or of vanishing derivatives higher than the first, since in that case Eq. (29) truncates to Eq. (22). In this limit,

$$\langle \hat{\theta}_i \rangle_p = \langle \hat{\theta}_i \rangle_p^{(0)} = (\bar{h}_i, \bar{h}_j)^{-1} (n, \bar{h}_j) \quad \text{(LSA)} \tag{35}$$

10 For uniform priors (e.g., rectangular distributions corresponding to the allowed parameter ranges), $\hat{\theta}^{MP}$ actually becomes undefined in the $A \rightarrow 0$ limit.
and
\[ \langle \Delta \tilde{\theta}_i \Delta \tilde{\theta}_j \rangle_p = \langle (\tilde{\theta}_i - \langle \tilde{\theta}_i \rangle_p)(\tilde{\theta}_j - \langle \tilde{\theta}_j \rangle_p) \rangle_p = (\tilde{h}_i, \tilde{h}_j)^{-1} \quad \text{(LSA)}, \]
(36)
and therefore
\[ \langle \tilde{\theta}_i \tilde{\theta}_j \rangle_p = \frac{1}{A^2}(\tilde{h}_i, \tilde{h}_j)^{-1} = (\tilde{h}_i, \tilde{h}_j)^{-1} \quad \text{(LSA)}, \]
(37)
as it can be seen by rewriting the exponential of Eq. (22) as
\[ p(s|\bar{\theta}) \propto \exp\left\{ -\theta_i \bar{h}_i \right\} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(\bar{\theta}_i - \tilde{\theta}_i)_p^2}{2} \right\}, \]
(38)
where we have omitted factors independent from \( \bar{\theta} \) that cancel out in the normalization of \( p(s|\bar{\theta}) \).

Reinstating \( A \) in Eq. (35) we see that in the high-SNR/LSA limit the mean of the posterior distribution coincides with the ML estimator, as it is reasonable, since the average of a normal distribution coincides with its mode. The two however differ when higher-order terms are included, as we shall see in Sec. VII. From Eq. (36) we see also that, to leading order, the variance of the posterior distribution is experiment-independent, and it coincides with the variance of the ML estimator (remember however that the two have very different interpretations\(^{11}\)).

With the addition of a Gaussian prior \( p(\theta) \propto e^{-P_i \theta_i^2}/2 \) centered at \( \theta = 0 \), Eqs. (35) and (36) change only slightly:\(^{12}\)
\[ \langle \tilde{\theta}_i \rangle_p = [(\tilde{h}_i, \tilde{h}_j) + P_{ij}/A^2]^{-1}(n, \tilde{h}_j) \]
\[ \langle \Delta \tilde{\theta}_i \Delta \tilde{\theta}_j \rangle_p = [(\tilde{h}_i, \tilde{h}_j) + P_{ij}/A^2]^{-1} \quad \text{(LSA/Gaussian prior)}. \]
(39)
Note that \( p(\theta) \propto e^{-1/(A^2)}P_{ij} \bar{\theta}_i \bar{\theta}_j/2 \) is formally a \( O(1/A^2) \) contribution to the likelihood exponential that would enter the \( 1/A \) expansion beginning at that order. However, if \( P_{ij} \) is large enough to matter at the signal strengths of interest, it probably makes sense to bundle it with the zeroth-order terms as we did here. In contrast with Eq. (28) for the frequentist variance of \( \bar{\theta}^{\text{ML}} \), we see that in the limit of vanishing signal strength the variance of the posterior goes to the variance \( P_{ij} \) of the prior.

III. STANDARD COMPACT-BINARY SIGNAL MODEL

Throughout the rest of this paper, our fiducial model for compact-binary signals will be simple stationary-phase–approximated (SPA) waveforms including phasing terms from the spin–orbit and spin–spin interactions of parallel or antiparallel component spins. Parameter estimation with these waveforms was studied by Poisson and Will\(^ {10} \). In these paper we adopt second-order post-Newtonian\(^ {13} \) (2PN) Fourier-domain waveforms as written by Arun and colleagues\(^ {11} \):
\[ \tilde{h}(M_c, \eta, \beta, \sigma, \phi_0, t_0; f) \propto f^{-7/6} \exp i\left\{ \psi(M_c, \eta, \beta, \sigma; f) + \phi_0 + 2\pi ft_0 \right\}, \]
(40)

---

11 If we define the quadratic error of the posterior distribution as \( \langle \tilde{\theta}_i \rangle_p \) (which is appropriate given that the true signal is at \( \theta = 0 \)), we must increment \( (\tilde{h}_i, \tilde{h}_j)^{-1} \) by the experiment-dependent quantity \( \langle \tilde{\theta}_i \rangle_p \rangle = (\tilde{h}_i, \tilde{h}_j)^{-1}(n, \tilde{h}_j)(\tilde{h}_m, \tilde{h}_j) \). Interestingly, the frequentist average of the Bayesian error \( \langle \tilde{\theta}_i \rangle_p \rangle \) is \( 2(\tilde{h}_i, \tilde{h}_j)^{-1} \), twice the frequentist variance of \( \tilde{\theta}^{\text{ML}} \).

12 With the Gaussian prior, the quadratic error \( \langle \tilde{\theta}_i \rangle_p \rangle \) becomes \( (\tilde{h}_i, \tilde{h}_j) + P_{ij}/A^2 \). Note that \( \langle \tilde{\theta}_i \rangle_p \rangle \) is large enough to matter at the signal strengths of interest, it probably makes sense to bundle it with the zeroth-order terms as we did here. In contrast with Eq. (28) for the frequentist variance of \( \tilde{\theta}^{\text{ML}} \), we see that in the limit of vanishing signal strength the variance of the posterior goes to the variance \( P_{ij} \) of the prior.

13 Waveform phasing expressions accurate to 3.5PN order are also provided in Ref. [11]. We do not use these in this article for the sake of simplicity, since they would not change the qualitative picture of parameter estimation presented here. For the reader’s reference, however, the higher-than-2PN corrections to Eq. (41) are

\[
\pi \left( \frac{38645}{252} - \frac{65}{9} \eta^2 \right) v^3 \log v + \left( \frac{1158323136051}{4694215680} - \frac{640}{3} \right) \eta^2 + \frac{6848}{21} \eta \log(4) + \left( -\frac{1533597827}{3048192} + \frac{2255}{12} \pi^2 \right) \eta^2 - \frac{1760}{3} \eta + \frac{12320}{9} \lambda \eta + \frac{70055}{1728} \eta^3 - \frac{127825}{1296} \eta^3 \right) v^5 - \frac{6848}{21} v^5 \log v + \pi \left( \frac{77096675}{254016} + \frac{378515}{1512} \eta - \frac{74045}{756} \eta^2 \right) v^7,
\]
where \( \gamma = 0.57721 \cdots \) is Euler’s constant, and \( \lambda = -1987/3080 \) and \( \theta = -11831/9240 \) are recently determined constants in the PN expansion\(^ {12} \).
with
\[
\psi(M_c, \eta, \beta, \sigma; f) = \frac{3}{128 \eta v^5} \left( 1 + \frac{20}{9} \left( \frac{743}{336} + \frac{11}{4} \eta \right) v^2 + (4\beta - 16\pi) v^3 + 10 \left( \frac{3058673}{1016064} + \frac{5429}{1008} \eta + \frac{617}{144} \eta^2 - \sigma \right) v^4 \right),
\]
(41)

where \( v = (\pi M f)^{1/3} \), \( M = m_1 + m_2 \) is the total mass of the binary, \( \eta = m_1 m_2 / M^2 \) is the symmetric mass ratio, \( M_c = M^3/5 \) is the chirp mass. The spin–orbit parameter \( \beta \) and the spin–spin parameter \( \sigma \) [10, 13] are given by
\[
\beta = \sum_{i=1}^{2} \frac{\hat{L} \cdot S_i}{12 m_i^2} \left[ 113 \left( \frac{m_i}{M} \right)^2 + 75 \eta \right] = \frac{113 \hat{L} \cdot (S_1 + S_2) + 75 \hat{L} \cdot [(m_2/m_1)S_1 + (m_1/m_2)S_2]}{12 M^2},
\]
(42)
and
\[
\sigma = \frac{721(\hat{L} \cdot S_1)(\hat{L} \cdot S_2) - 247 S_1 \cdot S_2}{48 m_1 m_2 M^2},
\]
(43)

with \( S_1 \) and \( S_2 \) the spins of the binary components. We truncate waveforms at the (Keplerian) last stable circular orbit \( (v = 1/\sqrt{6}) \).

For simplicity, in this article we do not discuss the estimation of the amplitude parameter \( A \) that would multiply the right-hand side of Eq. (40). [From Eqs. (2) and (40) it follows that \( \partial A / \partial h_i = 0 \) for \( i \neq A \), so the amplitude \( A \) effectively decouples from all other parameters in the Fisher matrix.] However, all discussions to follow can accommodate the addition of \( A \) with trivial modifications.

### IV. THE SINGULAR CASE OF THE DISAPPEARING PARAMETER

In Sec. II we have examined the interpretation of the inverse Fisher matrix as a frequentist or Bayesian measure of error or uncertainty. In this section, we discuss what happens when the Fisher is singular, or almost so, so that the attempts to invert it numerically yield warnings that it is badly conditioned. It is pedagogical to begin this discussion by considering the case where the matrix is exactly singular (Sec. IV A), and then to widen our scope to approximate singularity (Sec. IV B). The conclusion is that a singular Fisher matrix is almost always a symptom that the high-SNR/LSA limit is not to be trusted, that prior probabilities play an important role, or both.

#### A. Singular Fisher matrix

A singular Fisher matrix implies that the corresponding LSA likelihood (22) is a singular normal distribution. There is nothing wrong with that, except that the results of Secs. II D and II E cease to apply:

- (In the frequentist ML framework.) A singular normal distribution has no maximum, so the ML equation has no solutions, and Eqs. (23) and (24) are invalid. The LSA likelihood is constant along the directions of the Fisher-matrix eigenvectors with null eigenvalues\(^\dagger\) (henceforth, somewhat improperly, we shall call these null eigenvectors). This is generally interpreted as saying that the data do not favor any value of the linear combinations of parameters along those directions, since parameter changes in those directions do not alter the signal model.\(^\dagger\)

- (In the Bayesian framework.) A singular normal distribution has infinite even moments, even for variables that do not appear in the Fisher-matrix null eigenvectors (just consider that they being integrated against a constant function). We are going to largely ignore this issue, treating the parameters as pure numbers resulting from adopting a God-given system of units.

\(^\dagger\) A reasonable objection to computing the eigensystem of the Fisher matrix is that it leads to taking linear combinations of parameters that may have different units. It is possible to avoid this problem by looking at the Fisher matrix more abstractly as a linear operator, and talking of its range and null space [14]; or more pragmatically, by dividing all parameters by their typical range; or perhaps by taking their logarithm (since we are working with errors, units can be forgotten as additive constants), which in the linearized theory is equivalent to dividing by the true parameters. We are going to largely ignore this issue, treating the parameters as pure numbers resulting from adopting a God-given system of units.

\(^\dagger\) This is true at least with respect to the noise-weighted inner product: if \( \langle \vec{h}_i, \vec{h}_j \rangle = 0 \), then also \( \langle \hat{h}(\theta) - \hat{h}_0, \hat{h}(\theta) - \hat{h}_0 \rangle = \langle \hat{h}_i \theta_i, \hat{h}_j \theta_j \rangle = 0 \).
Thus, for a singular Fisher matrix the frequentist variance of the ML estimator and the Bayesian variance of the posterior distribution are (formally) infinite for all parameters. As we just discussed, this infinity is just a disagreeable symptom that the ML equation has no solution, and that the LSA likelihood is not a normalizable probability density. What is the solution? Well, if the signal is really linear, so that the LSA expressions are exact, it is possible to discard the combinations of parameters that correspond to the null eigenvectors, and characterize the variance of the remaining parameters. Let us see how, in the frequentist and Bayesian frameworks. We denote the total number of source parameters by \( N \), and the number of non-null eigenvectors by \( R \).

- (In the frequentist ML framework.) Writing \((\bar{h}_i, \bar{h}_j)\) in the singular-value (SV) decomposition\(^{16}\) \(^{[15]} \) as \(\sum_{\lambda(k)\neq 0} \theta_i^{(k)} \lambda^{(k)} \theta_j^{(k)} \) (with \( k = 1, \ldots, R \), or in matrix notation \( \Theta \Sigma \Theta^T \) (with \( \Theta \) a \( N \times R \) matrix with orthonormal columns, and \( \Sigma \) a diagonal matrix formed from the \( R \) non-zero eigenvalues), we can refactor the ML equation as

\[
(\Theta \Sigma \Theta^T) \hat{\theta} = A^{-1} \hat{n} 
\Rightarrow (\hat{\theta}^T \hat{\theta}) = A^{-1} (\Sigma^{-1} \Theta^T \hat{n}) 
\Rightarrow \hat{c}^{(k)} = A^{-1} (\lambda^{(k)})^{-1} n^{(k)},
\]

where \( \hat{c}^{(k)} \) and \( n^{(k)} \) denote the coefficients of the decompositions of \( \hat{\theta} \) and \( (\bar{h}_i, n) \) with respect to the normalized non-null eigenvectors of \((\bar{h}_i, \bar{h}_j)\). Since the ensemble average \( \langle n^{(k)} n^{(l)} \rangle_{\bar{n}} \) is just \( \lambda^{(k)} \delta^{(k,l)} \) (where \( \delta \) is Kronecker’s delta), the frequentist covariance of the ML estimators \( \hat{c}^{(k)} \) is the diagonal matrix \( (\lambda^{(k)})^{-1} \delta^{(k,l)} \).

From this we can work back to the components of the covariance matrix that involve involving any \( \hat{\theta}_i \) that do not appear in the null eigenvectors. All other \( \hat{\theta}_i \), however, are completely indeterminate.\(^{17}\) It may be possible to work back to interval estimates of their values by combining a ML estimate of the \( \hat{c}^{(k)} \) with finite allowed ranges for some of the \( \hat{\theta}_i \); however, this would constitute a form of prior distribution for the \( \hat{\theta}_i \), which is not entirely compatible with the ML estimator (what happens if the solution of the ML equation falls outside the allowed range?).

- (In the Bayesian framework.) Here we may use the following argument to drop the (non)dependence of the likelihood on the null eigenvectors. The quantities of interest are the moments of the \( c^{(k)} \) (with \( k = 1, \ldots, R \)) over infinite ranges of the \( c^{(k)} \) and of the coefficients \( C^{(K)} \) (with \( K = 1, \ldots, N - R \)) corresponding to the null eigenvectors, which are not included in the SV decomposition. Formally, these moments are ratios of two infinites, such as

\[
\langle \Delta c^{(k)} \rangle_p = \frac{\int c^{(k)} p(s|c) dc dC}{\int p(s|c) dc dC},
\]

because by hypothesis the LSA likelihood is not a function of the \( C \), not even through the \( n^{(K)} \equiv \theta^{(K)}(\bar{h}_i, n) \) terms, which are zero since \( \langle \theta_i^{(K)} \bar{h}_i, \theta_j^{(K)} \bar{h}_j \rangle = 0 \). However, these moments may be evaluated trivially as improper integrals, in the limit of the ranges for the \( c^{(K)} \) extending to infinity:

\[
\langle c^{(K)} \rangle_p = \lim_{C^{(K)} \rightarrow \infty} \frac{\int c^{(k)} p(s|c) dc}{\int p(s|c) dc} = A^{-1} (\lambda^{(k)})^{-1} n^{(k)}.
\]

From these integrals, we can work back to all the posterior moments involving the \( \hat{\theta}_i \) that do not appear in the null eigenvectors. But again, the other \( \hat{\theta}_i \) are completely indeterminate, and the moments involving them are formally infinite (for even moments) or indeterminate (for odd moments, whose value depends on the centering of the range in the improper-integral limits).

If salvation is to come within the Bayesian framework, it is from the prior probability distributions that make the posterior integrable.\(^{18}\) Unless the priors are also normal, though, the resulting moments cannot be expressed simply as analytical expressions of the Fisher matrix.

\(^{16}\) For square matrices, the SV decomposition is essentially equivalent to an eigenvector–diagonal-matrix decomposition where we drop the rows and columns corresponding to the null eigenvalues and eigenvectors.

\(^{17}\) In a truly linear system, this is true no matter how small the eigenvector component in that parameter direction; clearly, we this raises a problem of accuracy in the numerical computation of eigenvectors.

\(^{18}\) Even a single prior in the form of a rectangle function will regularize the integration over all the null-eigenvector coordinates that include that parameter. For normal priors, whether the posterior becomes integrable depends on the eigenstructure of \( A^2(\bar{h}_i, \bar{h}_j) + P_{ij}, \)
The most benign outcome occurs when the null eigenvectors correspond individually to one or more of the original parameters, or when the subspace spanned by null eigenvectors corresponds to a subset of the original parameters. The null-eigenvector combinations of parameters may also have clear physical interpretations: for instance, for a monochromatic, continuous sinusoid of frequency $f$, the absolute time offset $t_0$ and the initial phase $\phi_0$ are essentially the same parameter, so the Fisher matrix has a null eigenvector along the parameter combination $f t_0 + \phi_0$, which can be discarded, while $ft_0 + \phi_0$ remains well determined. A similar case is the degeneracy between luminosity distance and a certain function of the sky-position angles in the analysis of short GW chirps with a single ground-based detector.\footnote{Although neither of these examples is a linear model described exactly by the LSA, the degeneracy persists in the exact likelihood, so its Fisher-matrix diagnosis is correct. For such “perfect” degeneracies to occur, the two parameters must appear in all waveform expressions only as a sum or product; this would imply that their units can be sensibly summed, or that their combination has direct physical meaning.}

Other combinations of parameters can be more ambiguous and troubling—what is the meaning of estimating a parameter equal to a mass plus a spin? In those cases, our best hope is again that the degeneracy will be cured by prior probabilities, or by higher-order corrections in the $1/A$ expansion, in which cases the Fisher-matrix formalism is certainly insufficient.\footnote{Unless the cure lies in Gaussian priors centered on the true source parameters, which can be handled by an extension the Fisher formalism, as discussed in Sec. II. However, these priors are rarely appropriate for realistic signals.}

## B. Ill-conditioned Fisher matrix

All nonsingular matrices have well-defined inverses, although these may be difficult to compute. Indeed, the notion of ill conditioning was developed to quantify this difficulty, helping us to judge whether the accuracy to which a system of linear equations is known and the numerical precision available for our calculations are sufficient to trust our results. To wit, given the system

$$ (M + \delta M)(x + \delta x) = f + \delta f, \quad (47) $$

we ask how the perturbation $\delta x$ in the vector of unknowns scales with $\delta M$ and $\delta f$. This question is best tackled using vector and matrix norms [15], such as the vector 2-norm $||x||_2 = (\sum x_i^2)^{1/2}$ and the corresponding matrix 2-norm $||M||_2 = \text{sup}_x ||Mx||_2/||x||_2$. It is easy to see that the 2-norm of a matrix is equal to the largest of its eigenvalues, while the reciprocal of its smallest-modulus eigenvalue is the 2-norm of the inverse matrix. Under reasonable conditions it can be shown that [15]

$$ ||\delta x||/||x|| \leq \kappa(M) \left( ||\delta M||/||M|| + ||\delta f||/||f|| \right) + O(||\delta f||^2, ||\delta M||^2), \quad (48) $$

where the condition number $\kappa(M) = ||M||/||M^{-1}||$ measures the sensitivity of the linear system to perturbations in its inputs. From the definition of 2-norm given above, it follows that the 2-norm condition number $\kappa_2(M)$ is given by the ratio of the largest- to smallest-modulus eigenvalues of $M$. A similar bound holds for the norm of the perturbation of the inverse of a perturbed matrix [15],

$$ ||(M + \delta M)^{-1} - M^{-1}||/||M^{-1}|| \leq \kappa(M)||\delta M||/||M|| + O(||\delta M||^2). \quad (49) $$

From a numerical-analysis perspective, as Finn [3] points out, the gist of Eqs. (48) and (49) is that, roughly speaking, matrix inversion can amplify roundoff error by a factor $\kappa$, leading to the loss of up to $\log_{10} \kappa$ digits of precision. The same amplification will apply to any inaccuracies in our knowledge of $M$ and $f$. Taken at face value, this means that the Fisher-matrix results for the variances [Eqs. (24) and (36)] may be inaccurate at a 100% level if the components of the Fisher matrix are not known to a fractional accuracy better than $\kappa^{-1}(F)$.

Of course, Eqs. (48) and (49) are only upper bounds, and this doomsday scenario needs not be realized in practice. One way to check whether the matrix-inversion sensitivity is a concern is to add small random perturbations, Monte Carlo-style, to the Fisher matrix elements, and then verify the change in the covariance matrix. Such an experiment for our standard SPA model is shown in Fig. 1, which tells a rather cautionary tale about numerical sensitivity in the inversion of that particular Fisher matrix. These problems can be cured, somewhat trivially, by adopting higher-precision arithmetics, and by computing the Fisher matrix to better accuracy. It may also be possible to improve
the condition number by changing the units of source parameters, which may reduce the magnitude gap between the largest- and smallest-modulus eigenvalues (however, see note 14).

More to the point, in the light of the discussion of the previous section, it is the consequences of the Fisher-matrix condition number on the substance (rather than the numerical accuracy) of Eqs. (23) and (35) that should attract our attention. A large condition number implies one or more small Fisher-matrix eigenvalues, and consequently large statistical fluctuations for the combinations of source parameters corresponding to the small-eigenvalue eigenvectors, at least according to the LSA. The interpretation is that large parameter changes in the direction of the small-eigenvalue eigenvectors are needed to produce changes in the waveform comparable to typical noise fluctuations.

Under this condition, we have to worry whether the LSA can really describe the likelihood over the entire parameter ranges of interest: of course, these depend on the SNR available at detection (at leading order, their extent is inversely proportional to signal strength). In Sec. VI we describe a numerical criterion to decide when the SNR is high enough to believe the LSA. We also have to worry whether prior probability distributions for the parameters (perhaps in the simple form of allowed ranges) already restrict the estimated (for frequentists) or probable (for Bayesians) values of parameters beyond what is predicted by the Fisher-matrix variance. In Sec. V we discuss a simple test to decide whether priors should be included.

V. THE BURDEN OF PRIOR COMMITMENTS

As Cutler and Flanagan point out [4, p. 2691], “it is not necessary for a priori information to be very detailed or restrictive in order that it have a significant effect on parameter-extraction accuracy. All that is necessary is that it be more restrictive than the information contained in the waveform, for some of the parameters [...] what is more surprising is that due to the effects of correlations, the rms errors obtained for the other parameters may also be overestimated by large factors.” In this section we seek a practical recipe to determine, in the context of a parameter-estimation problem specified by a family of waveforms and a fiducial SNR, whether it is necessary to take priors into consideration when evaluating projected parameter accuracies.

Since prior probabilities can only be discussed consistently in the framework of Bayesian parameter estimation, in this section we will restrict ourselves to that context. The Gaussian priors examined at the end of Sec. II E are rarely appropriate in actual practice, but they do provide a quick test to see if the prior-less Fisher result can be taken as it stands, or whether a more careful analysis is need that includes the effects of priors. In Sec. VA we try out this quick test on the SPA model of Sec. III. For simplicity, we shall consider the effects of priors as logically independent from the sufficiency of the LSA, although the two problems clearly come into play together in real situations.
\[ m_1 = m_2 = 10M_\odot, \quad \beta = \sigma = 0, \quad \phi_0 = t_0 = 0, \quad \text{SNR} = 10 \]

\[
\begin{array}{ccccccc}
\Delta M_c/M_c & \Delta \eta & \Delta \beta & \Delta \sigma & \Delta \phi_0 & \Delta t_0 (\text{ms}) \\
4\text{pp}, \text{ no priors} & 2.9 \times 10^{-2} & 8.3 \times 10^{-2} & 7.3 & 3.0 \\
4\text{pp}, \text{ NTC prior on } \phi_0 & 1.3 \times 10^{-2} & 3.4 \times 10^{-2} & 2.9 & 1.2 \\
5\text{pp}, \text{ NTC prior on } \phi_0 & 1.3 \times 10^{-2} & \text{1.3} & 28.5 & 3.1 & 2.8 \\
5\text{pp}, \text{ NTC priors on } \phi_0, \beta & 1.3 \times 10^{-2} & 0.38 & 8.1 & 2.9 & 1.4 \\
5\text{pp}, \text{ NTC priors on } \phi_0, \eta, \beta & 1.3 \times 10^{-2} & 0.21 & 4.5 & 2.9 & 1.3 \\
6\text{pp}, \text{ NTC priors on } \phi_0, \eta, \beta & 3.5 \times 10^{-2} & 0.21 & 5.6 & 4.8 & 2.2 \\
4\text{pp}, \text{ exact priors on } \phi_0, \eta & 6.9 \times 10^{-3} & 1.2 \times 10^{-2} & 1.0 & 0.67 \\
5\text{pp}, \text{ exact priors on } \phi_0, \eta, \beta & 9.8 \times 10^{-3} & 7.2 \times 10^{-2} & 1.6 & 1.8 & 0.81 \\
6\text{pp}, \text{ exact priors on } \phi_0, \eta, \beta, \sigma & 2.1 \times 10^{-2} & 7.1 \times 10^{-2} & 2.5 & 2.5 & 1.8 & 1.3 \\
\end{array}
\]

TABLE I: Fisher-matrix rms errors in the 4-, 5-, and 6-parameter estimation problems for a \((10 + 10)M_\odot\) binary with no spins, evaluated under different combinations of normal-true-parameter-centered (NTC) priors (upper section of table) and of the exact priors of Sec. VB (lower section). The underlined errors are larger than the physical range for the parameter.

A. Testing for the influence of priors (normal true-parameter-centered priors)

We shall discuss our quick test by way of an example. The standard SPA model of Sec. III has six parameters: \(M_c, \eta, \beta, \sigma, \phi_0, \) and \(t_0\) (plus \(A\), which we disregard). We work at 2PN with \(\text{SNR} = A = 10\), with true parameters \(m_1 = m_2 = 10M_\odot\) (corresponding to \(M_c = 8.71M_\odot, \eta = 0.25\)), and \(\beta = \sigma = \phi_0 = t_0 = 0\). We wish to examine the effect of priors for three related parameter-estimation problems involving different subsets of parameters: we initially disregard spin parameters (which is equivalent to working under the assumption that we know a priori that the true binary has no spin); we then consider systems where the spin–orbit coupling [as represented by the terms proportional to \(\beta\) in Eq. (41)] can be important, but spin–spin interactions can be neglected; last, we include spin–spin interactions [as represented by the terms proportional to \(\sigma\)]. We refer to these estimation problems as the 4-, 5-, and 6-parameter problems (4pp, 5pp, and 6pp). As we shall see, priors become increasingly important as the number of parameters increases.

In each problem, we compute the expected covariance matrix of the posterior distribution as the inverse of (a submatrix of) the Fisher matrix, neglecting any non-LSA effects. We represent priors as normal distributions centered around null parameter displacements (i.e., the true parameter value), with standard deviations of 0.25 for \(\eta, \pi\) for \(\phi_0\), and (following Poisson and Will [10]) 8.5 for \(\beta\) and 5 for \(\sigma\). This representation is very crude, but it is the only one that leads to a simple analytical result [Eq. (39)] for the resulting posterior covariance, and it should give at least a qualitative idea of the effect of imposing rectangular priors covering the allowed parameter range ([0, 0.25]) for \(\eta, [-\pi, \pi]\) for \(\phi_0\), [−8.5, 8.5] for \(\beta\), and [−5, 5] for \(\sigma\). Results are shown in the upper section of Table I, and are as follows.

The first line of Table I shows the 4pp no-prior 1σ values for the single-parameter rms errors (i.e., the square roots of the diagonal elements in the covariance matrix). Among these, \(\Delta M_c, \Delta \eta\) seem reasonable, but we get immediately hung up on the value of \(\Delta \phi_0\). How can the 1σ error region be larger than the physically meaningful range for this angle? The answer is that the LSA cannot know that waveforms are exactly periodic (and therefore nonlinear) in \(\phi_0\) with a period of 2\(\pi\); instead, it extrapolates the small-\(\phi_0\) effects over an infinite range of \(\phi_0\). The Fisher-matrix result could still be essentially correct, in that the likelihood is essentially constant across the physical range for \(\phi_0\). Can we then conclude that \(\phi_0\) is undetermined, but continue to trust the errors for \(\Delta M_c, \Delta \eta\), and \(\Delta t_0\)?

It turns out that we cannot (see Table I, line 2: 4pp, prior on \(\phi_0\)). With the addition of an NTC prior for \(\phi_0\), which roughly restricts its range to physically sensible values, we find that the expected errors for \(M_c\) and \(\eta\) improve by factors of \(\geq 2\). Thus, omitting the \(\phi_0\) prior leads to seriously overestimating the errors in other parameters. We can understand how this happens by considering the simpler case of a two-by-two Fisher matrix and its inverse,

\[
\begin{pmatrix}
F_{11} & F_{12} \\
F_{12} & F_{22}
\end{pmatrix}^{-1} = \begin{pmatrix}
F_{11}^{-1}(1 - C_{12}^{-1})^{-1} & -C_{12}(F_{11}F_{22})^{-1/2}(1 - C_{12}^{-1})^{-1} \\
-C_{12}(F_{11}F_{22})^{-1/2}(1 - C_{12}^{-1})^{-1} & F_{22}^{-1}(1 - C_{12}^{-1})^{-1}
\end{pmatrix},
\]

where \(C_{12} = F_{12}/\sqrt{F_{11}F_{22}} = (h_1, h_2)/\sqrt{(h_1, h_3)(h_2, h_3)}\) (with \(|C_{12}| < 1\) is the correlation between the waveform derivatives along parameters 1 and 2. The expected error variances \(\sigma_1\) and \(\sigma_2\) appear on the diagonal. If we add a Gaussian prior for parameter 2, replacing \(F_{22}\) by \(F_{22} + P_{22}\) (with the parameter \(\gamma = (F_{22} + P_{22})/F_{22}\) measuring the relative strength of the prior), we find new variances \(\sigma_1^p = F_{11}^{-1}/(1 - \gamma^{-1}C_{12}^{-1})\) and \(\sigma_2^p = \gamma^{-1}F_{22}^{-1}/(1 - \gamma^{-1}C_{12}^{-1})\), and therefore variance improvement factors \(\sigma_1/\sigma_1^p = (1 - \gamma^{-1}C_{12}^{-1})/(1 - C_{12}^{-1})\) and \(\sigma_2/\sigma_2^p = \gamma(\sigma_1/\sigma_1^p)\). While the
improvement is stronger for parameter 2 (with the infinitely-strong–prior limit $\gamma \to \infty$ removing all variance), the parameter-1 variance is also considerably reduced if the correlation $C_{12}$ is appreciable.

What about the spin parameters? Adding $\beta$ to the mix (Table I, line 3: 5pp, prior on $\phi_0$) yields unphysically large errors for both $\beta$ and $\eta$; imposing the NTC prior on $\beta$ (5pp, priors on $\phi_0$, $\beta$) fixes the problem for $\beta$, but not for $\eta$. Adding the prior on $\eta$ (5pp, priors on $\phi_0$, $\eta$, $\beta$) contains its error reasonably, and improves the error for $\beta$. Last, going to the 6-parameter problem with priors on $\phi_0$, $\eta$, and $\beta$ yields a physical $\Delta \sigma$, while somewhat increasing $\Delta M_c$ and $\Delta \beta$. Thus we stop here, noting that in all of the 4pp, 5pp, and 6pp, the expected parameter-estimation accuracies are seriously distorted if priors are not included.

As a rule of thumb, we should expect such distortions whenever the signal derivatives show significant correlations, and when the magnitudes of the priors, measured crudely as the squared inverses $(\theta^\text{max} - \theta^\text{min})^2$ of the effective parameter ranges induced by the priors, are comparable to the corresponding diagonal Fisher-matrix elements $F_{ii}$. Less crudely, the quick test exemplified in this section can signal problems where priors are important.

**B. Testing for the influence of priors (exact priors)**

We can perform an even better test by evaluating the effects of exact priors (and not just of their NTC mock-ups), while still working in the LSA. Doing this requires some numerics, which are however very manageable on a workstation-class system. The idea is to integrate $\langle \Delta \theta_i \Delta \theta_j \rangle_p$ as a Monte Carlo sum, which can be accomplished as follows.

1. First, we need to fix a reference experiment by drawing the random variable $n_j \equiv (n, h_j)$ from its ensemble distribution, which in Gaussian noise is a normal distribution with mean zero and covariance matrix $F_{ij}$ (see Eq. (3)). To do so, we generate a zero-mean, unit-variance, normal $N$-tuple, and multiply it by $\sqrt{F^{-1}_{ij}}$ (where the square root is taken in the linear-operator sense and exists for nonsingular Fisher matrices). Note that we cannot work with SNR-invariant expressions (e.g., normalized parameter errors $\bar{\theta}_i$), since the priors set a scale for the strength of the signal.

2. Now we can draw samples distributed according to the LSA likelihood

$$p(s|\theta) \propto \exp\{-(h_j, h_k)\theta_j\theta_k/2 + n_j\theta_j\} \quad \text{(LSA)}.$$  

To do so, we generate zero-mean, unit-variance, normal $N$-tuples, multiply them by $\sqrt{F^{-1}_{ij}}$, and offset them by $F^{-1}_{ij} n^2$.

3. Last, we include the effects of priors (therefore obtaining a population $\{\theta^{(i)}\}$ distributed according to the LSA posterior probability) by going through the samples, and discarding each of them with a probability $1 - p(\theta)/\max_\theta p$ (for rectangular priors the discard probability is always 0 or 1). 

4. The covariance matrix of the posterior distribution can then be computed from the surviving samples.

5. We repeat this procedure for many different $n_i$s, and take a frequentist average of the covariance-matrix components (or study their frequentist distribution). (For the sake of clarity, let me restate the Bayesian interpretation of this entire procedure: we place the true signal at $\theta = 0$; we draw from the possible noise realizations according to their ensemble probability; and we compute the variance of the posterior distribution for each noise realization.)

If the priors are very restrictive compared to Fisher-matrix-only errors, we may find that we are discarding a very large percentage of the samples. To avoid this, we can incorporate a normal approximation to the priors in the probability distribution used to generate the samples (i.e., by multiplying normal variates by $\sqrt{[\langle h_i, h_j \rangle + P_{ij}]^{-1}}$, and offsetting them by $[\langle h_i, h_j \rangle + P_{ij}]^{-1} n^2$), and then sieve the resulting samples with respect to $\propto p(\theta)e^{P_{ij}\theta_i\theta_j/2}$ instead of $p(\theta)$. (Thus we implement an ad hoc correction to the initial distribution, to avoid sampling unrealistically large errors that would be overwhelmingly culled when we apply the priors, and then we correct the resampling probability to offset the ad hoc correction.) It is also possible to use the Metropolis algorithm [16] with the likelihood

\[21\] The nominal 1σ interval for $\eta$ is $[0.04, 0.46]$, which covers most of the acceptable range $[0, 0.25]$ and extends beyond it into “illegal” $\eta$ values; this happens because we can only easily impose normal priors centered on the true parameter value.
or likelihood-plus-NTC-prior as proposal distribution, and the full posterior as the target distribution. A simpler alternative is rejection sampling [17], which we employed for the results reported in this section: if there exists an $M$ for which $p_1(x) \leq M p_2(x)$, then a population of samples distributed according to $p_2$ may be culled to a population distributed according to $p_1$ by accepting each sample with a probability $p_1(x)/[M p_2(x)]$. In the case of rectangular priors extending through the physical ranges of the parameters, and of diagonal $P_{ij}$, we find a possible $M$ is $\exp \sum_i P_{ii}(\theta_i^B)^2$, with $\theta_i^B$ is the largest-modulus boundary value for each parameter.

Applying the procedure outlined above to our $(10 + 10)M_\odot$ system yields the results listed in the lower section of Table I. We adopt exact priors given by rectangular probability distributions covering the intervals $[0, \infty]$ for $M$, $[0, 0.25]$ for $\eta$, $[-8.5, 8.5]$ for $\beta$, $[-5, 5]$ for $\sigma$, and $[-\pi, \pi]$ for $\phi_0$. Each quoted error is a frequentist average of 200 independent Monte Carlo estimates, each computed for a different realization of noise from an initial sampling of $10^6$ parameter sets, reduced to $5 \times 10^4$ to $2 \times 10^5$ samples after rejection sampling, depending on the estimation problem.

The expected errors are significantly reduced compared to the NTC-prior estimates. These reductions stem mainly from the greater tightness of the rectangular priors, and are especially significant for for $\eta$, for which the symmetric NTC prior is indeed very crude. The lesson is that we can use $[(h_i, h_j) + P_{ij}]^{-1}$ (i.e., the quick test) to decide whether priors are important, but we need something more sophisticated, such as the procedure described in this section, to gauge their effects accurately. Of course, this gain in accuracy may be only virtual if the LSA is not warranted for our problem. Deciding that question is the object of the next section.

VI. THE UNBEARABLE LIGHTNESS OF SIGNAL TO NOISE

As we have seen in Sec. II, the high-SNR and LSA limits coincide because larger signal strengths correspond to smaller statistical errors, which in turn imply that the linearized-signal expression (22) for the likelihood is more accurate. The equivalence of the two limits is manifest in the $1/A$ smallness of a term in a series does not guarantee that the series is converging.

As we have seen in Sec. II, the high-SNR and LSA limits coincide because larger signal strengths correspond to smaller statistical errors, which in turn imply that the linearized-signal expression (22) for the likelihood is more accurate. The equivalence of the two limits is manifest in the $1/A$ smallness of a term in a series does not guarantee that the series is converging.

As in Secs. II and V, we assume that the detector output is $s = A h_0 + n$. In the LSA, the ML estimator $\hat{\theta}^{ML}$ is a normal variable with mean zero and covariance matrix $(1/A^2)(h_j, h_k)^{-1}$ [Eq. (3)]. For a given signal amplitude $A$, let $\hat{\theta}^{ML}$ take on the specific value $\theta^{1}\sigma$ on its $1\sigma$ surface. From Eqs. (1), (22), and (52), the mismatch ratio $r$ is given by

$$r(\theta^{1}\sigma) = \exp \left\{ -\left( s - A(\bar{h}_0 + \theta^{1}\sigma \bar{h}_j), s - A(\bar{h}_0 + \theta^{1}\sigma \bar{h}_k) \right)/2 \right\} / \exp \left\{ -\left( s - A \bar{h}(\theta^{1}\sigma), s - A \bar{h}(\theta^{1}\sigma) \right)/2 \right\};$$

where $\Delta h(\theta) = h(\theta) - h(0)$, $A$ is the signal strength, and $\theta$ is the error (in a sense to be made precise shortly). Since $h(\theta) = h(0) + \theta_{1}\bar{h}_1 + \cdots$, the product in Eq. (52) represents the noise-weighted norm of the higher-than-linear contributions to $h(\theta)$, expanded around the true source parameters. The idea of the criterion is to choose an iso-probability surface (say, the $1\sigma$ surface), as predicted by the Fisher matrix, and then explore it to verify that the mismatch between the LSA and exact likelihoods is smaller than a fiducial value (say, $|\log r| < 1/10$), so that we can actually believe the LSA in predicting the $1\sigma$ surface to begin with.

We stress that this is just a criterion of consistency. Even if the Fisher-matrix result is internally consistent, it may still be inaccurate; conversely, the structure of the ambiguity function across parameter space could conspire in such a way as to make the LSA results correct, although we have no reason to expect that in general. In the rest of this section, we explain how the criterion comes about in the frequentist (Sec. VI A) and Bayesian (Sec. VI B) frameworks, and we show a concrete example of the criterion in use (Sec. VI C).

A. Frequentist justification of the maximum-mismatch criterion

As in Secs. II and V, we assume that the detector output is $s = A \bar{h}_0 + n$. In the LSA, the ML estimator $\hat{\theta}^{ML}$ is a normal variable with mean zero and covariance matrix $(1/A^2)(h_j, h_k)^{-1}$ [Eq. (3)]. For a given signal amplitude $A$, let $\hat{\theta}^{ML}$ take on the specific value $\theta^{1}\sigma$ on its $1\sigma$ surface. From Eqs. (1), (22), and (52), the mismatch ratio $r$ is given by

$$r(\theta^{1}\sigma) = \exp \left\{ -(s - A(\bar{h}_0 + \theta^{1}\sigma \bar{h}_j), s - A(\bar{h}_0 + \theta^{1}\sigma \bar{h}_k))/2 \right\} / \exp \left\{ -(s - A \bar{h}(\theta^{1}\sigma), s - A \bar{h}(\theta^{1}\sigma))/2 \right\};$$

where $\Delta h(\theta) = h(\theta) - h(0)$, $A$ is the signal strength, and $\theta$ is the error (in a sense to be made precise shortly). Since $h(\theta) = h(0) + \theta_{1}\bar{h}_1 + \cdots$, the product in Eq. (52) represents the noise-weighted norm of the higher-than-linear contributions to $h(\theta)$, expanded around the true source parameters. The idea of the criterion is to choose an iso-probability surface (say, the $1\sigma$ surface), as predicted by the Fisher matrix, and then explore it to verify that the mismatch between the LSA and exact likelihoods is smaller than a fiducial value (say, $|\log r| < 1/10$), so that we can actually believe the LSA in predicting the $1\sigma$ surface to begin with.

We stress that this is just a criterion of consistency. Even if the Fisher-matrix result is internally consistent, it may still be inaccurate; conversely, the structure of the ambiguity function across parameter space could conspire in such a way as to make the LSA results correct, although we have no reason to expect that in general. In the rest of this section, we explain how the criterion comes about in the frequentist (Sec. VI A) and Bayesian (Sec. VI B) frameworks, and we show a concrete example of the criterion in use (Sec. VI C).
writing $s$ out, we eliminate all instances of $\bar{h}_0$: 
\[
\log r(\theta^{1\sigma}) = -A^2\theta_j^{1\sigma}\theta_k^{1\sigma}(\bar{h}_j, \bar{h}_k)/2 + A^2(\Delta \bar{h}(\theta^{1\sigma}), \theta^{1\sigma})/2 + A\theta_j^{1\sigma}(\bar{h}_j, n) - A(\Delta \bar{h}(\theta^{1\sigma}), n). \quad (54)
\]

The first two terms in the exponent can be computed given $\theta^{1\sigma}$; not so the two products involving $n$. To obtain the first, we note that if $\hat{\theta}^{\text{ML}} = \theta^{1\sigma}$, then the noise must be such that $\theta^{1\sigma} = (1/A)(\bar{h}_j, \bar{h}_k)^{-1}(\bar{h}_j, n)$ [Eq. (23)], so $(\bar{h}_j, n) = A(\bar{h}_j, \bar{h}_k)\theta^{1\sigma}$. To obtain the second, we change our perspective slightly, and average $\log r(\theta^{1\sigma})$ over all noise realizations $n$ compatible with $\theta^{1\sigma}$. This is how. Let $x_j = (\bar{h}_j, n)$, $y = (\Delta \bar{h}(\theta^{1\sigma}), n)$: separately, $x_j$ and $y$ are random normal variables with mean zero and covariances equal to $(\bar{h}_j, \bar{h}_k)$ and $(\Delta \bar{h}(\theta^{1\sigma}), \Delta \bar{h}(\theta^{1\sigma}))$, respectively; taken together, they are jointly normal variables with covariance $(\bar{h}_j, \Delta \bar{h}(\theta^{1\sigma}))$ [Eq. (3) again]. We now know enough to build $p(x, y)$, from which we can derive the conditional probability $p(y|x)$ and compute the conditional mean of $y$, which (after the algebra of App. A) turns out to be $A\theta_j^{1\sigma}(\bar{h}_j, \Delta \bar{h}(\theta^{1\sigma}))$. Altogether, we find 
\[
\langle \log r(\theta^{1\sigma}) \rangle_{n(\theta^{1\sigma})} = A^2\theta_j^{1\sigma}\theta_k^{1\sigma}(\bar{h}_j, \bar{h}_k)/2 + A^2(\Delta \bar{h}(\theta^{1\sigma}), \theta^{1\sigma})/2 - A^2\theta_j^{1\sigma}(\bar{h}_j, \Delta \bar{h}(\theta^{1\sigma}))
\]
\[
= (\theta_j^{1\sigma}h_j - \Delta h(\theta^{1\sigma}), \theta_k^{1\sigma}h_k - \Delta h(\theta^{1\sigma}))/2,
\]
just as anticipated in Eq. (52). The signal strength $A$ enters Eq. (55) explicitly, but also implicitly through the parameter width of the Fisher-matrix $1\sigma$ surface. Thus Eq. (52) can be solved for the $A$ that corresponds to $\theta^{1\sigma}$ small enough to yield $r$ close to unity as desired. Since to leading order $\theta_j^{1\sigma} - \Delta h(\theta^{1\sigma}) = A(h_j, \theta^{1\sigma})^\prime\theta_{k}^{1\sigma}$, and since to leading order $\theta^{1\sigma}$ scales as $1/A$, we expect $\log r$ to scale as $1/A^2$ for large enough $A$.

In summary, the maximum-mismatch criterion is justified from a frequentist viewpoint as a constraint on the ratio $r$ at points on a constant–LSA-probability surface, averaged over all realizations of noise compatible with finding the ML estimator at those points.

**B. Bayesian justification of the maximum-mismatch criterion**

The justification of the maximum-mismatch criterion from a Bayesian viewpoint requires another slight change of perspective. Again we assume $s = A\bar{h}_0 + n$; this time, however, we expand the waveform not with respect to the true parameters ($\theta = 0$), but to the observed location $\hat{\theta}^{\text{ML}}(n) = \theta_0^n$ of maximum LSA likelihood for a given experiment. In the absence of priors, it is with respect to this location that the uncertainty of the posterior would be judged in a single experiment. Thus we write $h(\theta) = h(\theta_0^n + \theta^n) \simeq h_0^n + \theta^n h_0^n$, where $h_0^n = h(\theta_0^n)$: the “$n$” superscripts serve to remind us that the parameter displacements $\theta^n$ (and the waveform derivatives $h_0^n$) are evaluated from (at) $\theta_0^n$. We also write $\Delta h_0^n$ for $h_0^n - h_0^n$.

The $1\sigma$ surface over which we are going to evaluate the ratio $r$ will be a surface of equiprobable true-signal locations, given the observed location $\theta_0^n$ of maximum LSA likelihood. In the LSA, the distribution over experiments of the true-signal location with respect to $\theta_0^n$ is again normal with covariance matrix $(h_0^n, h_0^n)^{-1}$. Thus we have $\theta^n = -\theta_0^n$, and the mismatch ratio $r$ is given by
\[
r(\theta_0^n) = \exp \left\{ -(s - A(h_0^n + \theta_j^n h_j))/2 \right\} / \exp \left\{ -(s - A\bar{h}_0, s - A\bar{h}_0) \right\}; \quad (56)
\]
writing $s$ out, the denominator reduces to $\exp(-n/2)$, and $\bar{h}_0$ enters the numerator only through $\Delta h_0^n$:
\[
\log r(\theta_0^n) = -A^2\theta_j^n\theta_k^n(h_j^n, h_k^n)/2 - A^2(\Delta h_0^n, \Delta h_0^n)/2 + A^2\theta_j^n(h_j^n, \Delta h_0^n) + A\theta_j^n(h_j^n, n) - A(\Delta h_0^n, n). \quad (57)
\]

Now, since the LSA likelihood can be written as $p(s|\theta^n) \propto \exp(-n + \Delta h_0^n - \theta^n h_0^n, n + \Delta h_0^n - \theta^n h_0^n)/2$, the ML equation $\partial p/\partial\theta_j^n = 0$ at $\theta_j^n = 0$ implies $(h_0^n, n) = -(\Delta h^n, \Delta h_0^n)$. Thus, the third and fourth terms in (57) cancel out. We handle the last term of the equation by evaluating the conditional mean of $\theta^n \equiv -A(\Delta h_0^n, n)$ given $x^n = (\bar{h}_j^n, n) = -A(\bar{h}_j^n, \Delta h_0^n)$, producing $-A(\bar{h}_j^n, n)(\bar{h}_j^n, \Delta h_0^n)^{-1}(\bar{h}_k^n, \Delta h_0^n)$ (again, see App. A). Before we can use this expression in Eq. (57), we use Eq. (23) to replace $(\bar{h}_j^n, n)$ with $(A(h_j^n, \bar{h}_k^n))\theta_0^n \simeq -A(\bar{h}_j^n, \bar{h}_k^n)\theta_0^n$ (to leading order), yielding $A^2\theta_j^n(h_j^n, \Delta h_0^n)$, just as needed to complete the square in Eq. (57):
\[
\log r(\theta_0^n) = -A^2\theta_j^n\theta_k^n(h_j^n, \Delta h_0^n)/2 + A^2\theta_j^n(h_j^n, \Delta h_0^n) - A^2(\Delta h_0^n, \Delta h_0^n)/2
\]
\[
= -(\theta_j^n - \Delta h(\theta^n), \theta_k^n - \Delta h(\theta^n))/2. \quad (58)
\]
Again, this equation can be solved for the $A$ that corresponds to $1\sigma$ true-signal locations $\theta^n$ small enough to yield $r$ close to unity. Interestingly, the signs of the frequentist and Bayesian expressions (55) and (58) are opposite,
Too see why this is the right thing to do, consider the integration of a function against the LSA distribution, and make a change of maximum LSA likelihood, and then exploring a surface of equiprobable true-signal locations, evaluating for each the only their absolute value as rough indicators of the appropriateness of the high-SN/LSA limit.

Bayesian case. Given the conditions under which we have obtained Eqs. (55) and (58), it is perhaps best to consider prima facie | that the likelihood is overestimated in the frequentist case, underestimated in the Bayesian case. Given the conditions under which we have obtained Eqs. (55) and (58), it is perhaps best to consider only their absolute value as rough indicators of the appropriateness of the high-SN/LSA limit.

In summary, the maximum-mismatch criterion is justified from a Bayesian viewpoint by fixing the location of the false, to unrealistic values) as we move from the 4pp to 5pp and 6pp.

FIG. 2: Cumulative distribution function for | log r | on the 1σ surface at various SNRs for our reference SPA model with \( m_1 = m_2 = 10 M_\odot \). The SNR required to have 90% of the 1σ points at | log r | = 0.1 (dashed lines) increases considerably (in fact, to unrealistic values) as we move from the 4pp to 5pp and 6pp.

indicating (at least prima facie) that the likelihood is overestimated in the frequentist case, underestimated in the Bayesian case. Given the conditions under which we have obtained Eqs. (55) and (58), it is perhaps best to consider only their absolute value as rough indicators of the appropriateness of the high-SN/LSA limit.

In summary, the maximum-mismatch criterion is justified from a Bayesian viewpoint by fixing the location of maximum LSA likelihood, and then exploring a surface of equiprobable true-signal locations, evaluating for each the average of log r over all experiments (i.e., realizations of noise) compatible with having the true signal there.

C. Practical usage of the maximum-mismatch criterion

In both the frequentist and the Bayesian pictures, Eq. (52) yields | log r | (the logarithm of the likelihood mismatch) as a function of the signal strength \( A \) and of a direction in parameter space that identifies a point on the 1σ surface given by the solutions of the LSA equation \( A^2(\bar{h}, \bar{h})\theta_j\theta_k = 1 \), and interpreted as equiprobable locations for the ML estimator given the true signal \( \theta = 0 \) (in the frequentist picture), or for the true signal given the mode of the likelihood at \( \theta = 0 \) (in the Bayesian picture). We use Eq. (52) by fixing the signal strength to what is reasonably expected in observations, perhaps close to the minimum detection SNR, although the astronomical distribution and intrinsic strengths of sources may prompt other choices (e.g., the the supermassive–black-hole binaries to be observed by LISA have typical SNRs in the hundreds); and then by evaluating | log r | as a function of direction in parameter space, sampling the entire 1σ surface in a manner consistent with the LSA distribution at 1σ.

One way to do so is to obtain the eigenvalues \( \lambda^{(i)} \) and eigenvectors \( \theta_j^{(i)} \) of \((\bar{h}, \bar{h})\), and then sample the parameter values \( \theta = \sum_{i=1}^N \bar{c}^{(i)} \theta_j^{(i)}/(A\bar{\lambda}^{(i)}) \), with \( \bar{c}^{(i)} \) distributed uniformly on the \( N \)-dimensional unit sphere. We then obtain the cumulative distribution function for the values of | log r |, which we plot in Fig. 2 for our reference model. If we consider the high-SN/LSA limit to be sufficiently realized when (say) | log r | < 0.1 over 90% of the 1σ surface, we conclude that the Fisher-matrix formalism (with no priors) is self-consistent for SNRs between 10 and 20 in the 4-parameter problem, between 100 and 200 in the 5pp, and between 4000 and 10000 in the 6pp.

The eigenvalue directions that push the required SNR toward higher values are usually those associated with the smallest-magnitude eigenvalues. To confirm this, and to get some clues about the beyond-LSA structure of the likelihood, we can fix the maximum acceptable value of | log r | (say, again to 0.1) and then solve Eq. (52) for \( A \) as a function of direction in parameter space. We do so for the 4pp in Fig. 3, where we show all two-dimensional parameter

---

22 Too see why this is the right thing to do, consider the integration of a function against the LSA distribution, and make a change of variables (with unit Jacobian) to eigenvalue components, and a second to rescaled components \( e^{(i)} = A\bar{\lambda}^{(i)}/\bar{c}^{(i)} \):

\[
\int A^{2}\sigma_{j}(\bar{h}, \bar{h})/2d\theta = \int \left(\ldots\right) e^{-A^{2}\sum_{i}\lambda^{(i)}e^{(i)}^{2}/2}dc \propto \int \left(\ldots\right) e^{-\sum_{i}(\bar{c}^{(i)})^{2}/2}dc
\]

we see that the source parameters that correspond to \( \bar{c} \) lying on a sphere of fixed radius must lie on an isoprobability surface. To reassemble \( \theta \) from the \( \bar{c} \), we need to divide the eigenvectors by \( A\bar{\lambda}^{(i)} \).
subspaces along pairs of eigenvectors (strictly speaking, were are not sampling a single $1\sigma$ surface, but considering the set of such surfaces for all SNRs, and determining on which of them $|\log r| = 0.1$, as a function of parameter angle).

In conclusion, I submit that graphs like those of Fig. 2 can be useful to assess the consistency of the “straight” Fisher-matrix formalism, are easy to produce with little additional machinery, and should be included in all articles that use the formalism to predict the future parameter-estimation performance of GW observations. If a single number must be quoted, it could be something like the SNR at which 90% of the $1\sigma$ surface yields $|\log r| < 0.1$.

VII. BEYOND THE LINEARIZED-SIGNAL APPROXIMATION

In this section we develop mathematical tools to derive higher-than-LSA expressions for the frequentist mean and variance of the ML estimator over an ensemble of noise realizations (Sec. VII A), and for the Bayesian mean and variance of the posterior distribution (without priors) in a single experiment (Sec. VII B). These expressions provide corrections to the Fisher-matrix result, and can therefore be used to check its accuracy, as suggested by Cutler and Flanagan [4] (who compute the $1/A^4$ correction to the frequentist variance). A formal treatment of the $1/A$ expansion for the frequentist moments can be found in Barndorff-Nielsen and Cox [18]; Naftali, Makris, and colleagues [19] use the expansion to determine conditions for the ML estimate to become unbiased and attain the Cramér–Rao bound. However, computing higher-order corrections involves a considerable amount of tensorial algebra that calls for the use of specialized software, such as MathTensor [20]; they also involve higher-than-first derivatives of the waveforms and products of several inverse Fisher matrices, which may raise concerns about the numerical accuracy of the computations. Throughout this section, we distinguish between covariant and contravariant indices (as in $n_i = (\bar{h}_i, n)$ and $\theta^i$, respectively); in fact, we find it convenient to use the inverse normalized Fisher matrix $(\bar{h}^i, \bar{h}^j)^{-1}$ to raise indices, therefore hiding its repeated appearance in tensor expressions.
A. In the frequentist framework

Using the $1/A$ expansion of Eqs. (18) and (19), the perturbative ML equations can be written in general as

\[ H_{ij} \theta^i_{(1)} = N_i, \]
\[ H_{ij} \theta^i_{(2)} = N_{ij} \theta^j_{(1)} + N_{ijk} \theta^j_{(2)} - H_{ijk} \theta^j_{(2)}, \]
\[ H_{ij} \theta^i_{(3)} = N_{ij} \theta^j_{(2)} + N_{ijk} \theta^j_{(2)} - H_{ijk} \theta^j_{(3)} - H_{ijkl} \theta^j_{(3)}, \]
\[ H_{ij} \theta^i_{(4)} = N_{ij} \theta^j_{(3)} + N_{ijkl} \theta^j_{(3)} - H_{ijkl} \theta^j_{(4)} - H_{ijklm} \theta^j_{(4)}, \]

(59)

with $N_i = (n, \bar{h}_i)/0!$, $N_{ij} = (\bar{h}_{ij}, n)/1!$, $N_{ijk} = (\bar{h}_{ijk}, n)/2!$ (and so on), and

\[ H_{ij} = \frac{\bar{h}_{i,j}}{0! 1!}, \]
\[ H_{ijk} = \frac{\bar{h}_{i,jk}}{0! 2!} + \frac{\bar{h}_{ij,k}}{1! 1!}, \]
\[ H_{ijkl} = \frac{\bar{h}_{i,jkl}}{0! 3!} + \frac{\bar{h}_{ij,kl}}{1! 2!} + \frac{\bar{h}_{ijkl}}{2! 1!}, \]

(60)

where $\bar{h}_{i,j} = (\bar{h}_i, \bar{h}_j)$, $\bar{h}_{i,j,k} = (\bar{h}_i, \bar{h}_{ij}, \bar{h}_{ik})$ (and so on), and where the two factorials at each denominator are (respectively) of the number of indices before the comma minus one, and of the number of indices after the comma. Also, the $\theta^i_{(n)}$ of Eq. (59) are the unknown $1/A^n$ contributions to $\hat{\theta}^{\text{ML}}$ (as in Eq. (19), dropping hats for simplicity), while the multi-index parameter objects such as $\theta^{jk}_{(2)}$ are given by

\[ \theta^{jk}_{(2)} = \theta^{j(1)}_{(1)} \theta^k_{(1)}, \quad \theta^{jk}_{(3)} = \theta^{j(1)}_{(1)} \theta^k_{(2)} + \theta^{j(2)}_{(2)} \theta^k_{(1)}, \quad \theta^{jk}_{(4)} = \theta^{j(1)}_{(1)} \theta^k_{(3)} + \theta^{j(2)}_{(2)} \theta^k_{(2)} + \theta^{j(3)}_{(3)} \theta^k_{(1)}, \]
\[ \theta^{jk}_{(4)} = \theta^{j(1)}_{(1)} \theta^k_{(4)} + \theta^{j(2)}_{(2)} \theta^k_{(3)} + \theta^{j(3)}_{(3)} \theta^k_{(2)} + \theta^{j(4)}_{(4)} \theta^k_{(1)} , \]
\[ \theta^{ijkl}_{(4)} = \theta^{i(1)}_{(1)} \theta^{j(1)}_{(1)} \theta^k_{(1)} \theta^l_{(1)}, \]

(61)

and so on. In general, the object $\theta^{i_1 \ldots i_m}_{(n)}$ will consist of as many addends as there are partitions of $n$ into $m$ integers, including all permutations of each partition. For instance, the $n = 5$, $m = 3$ object $\theta^{ijkl}_{(5)}$ would have terms for each of the partitions $1 + 1 + 3, 1 + 3 + 1, 3 + 1 + 1, 1 + 2 + 2, 2 + 1 + 2, 2 + 2 + 1$.

The solution of each equation in Eq. (59) is trivial given the solutions of all equations of lower order. Since the frequentist mean and covariance of the $\theta^i$ can be built from these solutions, remembering Wick-product rule [21] for the ensemble average of products of Gaussian variables:

\[ \langle (a, n) \rangle_n = 0, \]
\[ \langle (a, n)(b, n) \rangle_n = (a, b), \]
\[ \langle (a, n)(b, n)(c, n) \rangle_n = 0, \]
\[ \langle (a, n)(b, n)(c, d) \rangle_n = (a, b)(c, d) + (a, c)(b, d) + (a, d)(b, c) \]

(63)
(for any signals \(a, b, c,\) and \(d\)), where all the products with an odd number of factors vanish, while the products with an even number of factors are given by the sum of terms corresponding to all distinct pairings of signals into inner products. Thus we find that all the \(\langle \theta_i^\text{odd}(k) \rangle_n\) vanish, while the first non-zero correction to \(\langle \theta_i^i \rangle_n\) is

\[
\langle \theta_i^i \rangle_n = \langle N^i_j N^j_i \rangle_n - H^j_{jk} \langle N^j_i N^k_j \rangle_n = \bar{h}^j_{jk} - H^j_{jk} \bar{h}^i_{jk}.
\]

(64)

As for the covariance,

\[
\langle \theta^i \theta^j \rangle_n - \langle \theta^i \rangle_n \langle \theta^j \rangle_n = \left[ \frac{1}{A^2} \langle \theta^i(1) \theta^j(1) \rangle_n + \frac{1}{A^2} \langle \theta^i(2) \theta^j(2) \rangle_n + \frac{1}{A^2} \langle \theta^i(1) \theta^j(3) \rangle_n + \frac{1}{A^2} \langle \theta^i(2) \theta^j(2) \rangle_n + \frac{1}{A^2} \langle \theta^i(1) \theta^j(3) \rangle_n + \cdots \right]

- \left[ \frac{1}{A^2} \langle \theta^i(1) \rangle_n \langle \theta^j(1) \rangle_n + \frac{1}{A^2} \langle \theta^i(2) \rangle_n \langle \theta^j(2) \rangle_n + \frac{1}{A^2} \langle \theta^i(3) \rangle_n \langle \theta^j(1) \rangle_n + \cdots \right]

+ \frac{1}{A^4} \left( \langle \theta^i(3) \rangle_n \langle \theta^j(3) \rangle_n + \langle \theta^i(2) \rangle_n \langle \theta^j(2) \rangle_n + \langle \theta^i(3) \rangle_n \langle \theta^j(1) \rangle_n + \cdots \right),
\]

(65)

where all the stricken-through terms vanish because they are proportional to ensemble products of an odd number of \(n\) terms. The surviving contributions are given by

\[
\langle \theta^i(1) \theta^j(1) \rangle_n = \langle N^i N^j \rangle_n = \bar{h}^i_{jk},
\]

\[
\langle \theta^i(1) \theta^j(3) \rangle_n = \langle N^i N^j k N^k N^i \rangle_n - H^j_{km} \langle N^i N^j_k N^i N^m \rangle_n + \langle N^i N^j k N^k N^i \rangle_n - H^j_{km} \langle N^i N^k N^l N^m \rangle_n.
\]

\[
\langle \theta^i(2) \theta^j(2) \rangle_n = \langle N^i k N^j N^i m N^m \rangle_n - H^j_{kl} \langle N^i k N^1 N^j m N^m \rangle_n - H^j_{kl} \langle N^i N^q N^k N^1 \rangle_n + H^i_{kl} H^j_{mq} \langle N^k N^q N^i k N^1 \rangle_n.
\]

(66)

and of course \(\langle \theta^i(3) \rangle_n \). The first of these equations, of course, reproduces the standard Fisher-matrix result. The four-\(N\) products in Eq. (66) follow from Eq. (63). For instance, the last two products are given by

\[
\langle N^k N^l N^m N^q \rangle_n = \langle N^k N^l m N^q \rangle_n + \langle N^q N^m N^q \rangle_n + \langle N^k N^q \rangle_n \langle N^l N^m \rangle_n
\]

\[
= \bar{h}^k_{lm} \bar{h}^m_{q} + \bar{h}^k_{m} \bar{h}^l_{q} + \bar{h}^k_{q} \bar{h}^l_{m},
\]

(67)

\[
\langle N^m N^q N^i k N^k \rangle_n = \bar{h}^m_{q} \bar{h}^i_{k} + \bar{h}^m_{q} \bar{h}^i_{k} + \bar{h}^m_{q} \bar{h}^i_{k}.
\]

These expressions can be substituted into those of Eq. (66), and those into Eq. (65), yielding the frequentist variance to order \(1/A^4\). Unfortunately, this requires computing second- and third-order waveform derivatives (the latter for the latter for \(H^j_{km} \)).

**B. In the Bayesian framework**

To generalize Eq. (34), we write

\[
\mathcal{I}^{(n)} = \int \bar{\theta} \frac{\partial^n p(0)}{\partial e^n} d\theta, \quad \mathcal{N}^{(n)} = \int \frac{\partial^n p(0)}{\partial e^n} d\theta,
\]

(68)

and find the recurrence relation

\[
\langle \bar{\theta}^i \rangle_p^{(n)} = \left( \mathcal{I}^{(n)} - \sum_{j=1}^{n} \left( \binom{n}{j} \langle \bar{\theta}^i \rangle_p^{(n-j)} \times \mathcal{N}^{(j)} \right) \right) \mathcal{N}^{(0)},
\]

(69)

as we can prove by expanding the identity \(\mathcal{I}^{(0)} + \epsilon \mathcal{I}^{(1)} + \cdots = [\langle \bar{\theta}^i \rangle_p^{(0)} + \epsilon \langle \bar{\theta}^i \rangle_p^{(1)} + \cdots] \times [\mathcal{N}^{(0)} + \epsilon \mathcal{N}^{(1)} + \cdots] \) on both sides as a series of \(\epsilon\), leading to

\[
\mathcal{I}^{(n)} = \sum_{j=0}^{n} \mathcal{N}^{(j)} \langle \bar{\theta}^i \rangle_p^{(n-j)},
\]

(70)
whence Eq. (69). To obtain all needed derivatives with respect to \( \epsilon \), we rewrite Eq. (29) as

\[
p(s|\theta) \propto \exp\left\{ - (n, n)/2 + \left[ N_\theta \bar{\theta}^i + \epsilon N_{ij} \bar{\theta}^i \bar{\theta}^j + \epsilon^2 N_{ijk} \bar{\theta}^i \bar{\theta}^j \bar{\theta}^k + \ldots \right] \right.
\]

\[
- \left[ H_{jk} \bar{\theta}^j \bar{\theta}^k + \epsilon H'_{jk} \bar{\theta}^j \bar{\theta}^k \bar{\theta}^l + \epsilon^2 H'_{jkl} \bar{\theta}^j \bar{\theta}^k \bar{\theta}^l \bar{\theta}^m + \ldots \right]/2, \tag{71}
\]

where (as in Sec. VII A) \( N_i = (n, \tilde{h}_i)/n! = n_i \), \( N_{ij} = (\tilde{h}_{ij}, n)/n! \), \( N_{ijk} = (\tilde{h}_{ijk}, n)/2! \) (and so on), but where the \( H'_{j_1 \ldots j_n} \) have slightly different denominators than the \( H_{j_1 \ldots j_n} \) of Eq. (60):

\[
H'_{ijk} = \bar{h}_{ijk} \frac{1}{k!} + \bar{h}_{ij,k} \frac{1}{2!} + \bar{h}_{ijk,l} \frac{1}{3!} + \ldots \tag{72}
\]

namely the denominator is \( m! \) for the product \( h_{j_1 \ldots j_m, j_1 \ldots j_n} \equiv (\tilde{h}_{j_1 \ldots j_m}, \tilde{h}_{j_1 \ldots j_n}) \). Expanding as a series of \( \epsilon \) yields

\[
p(s|\theta) \propto e^{-H_{jk} \bar{\theta}^j \bar{\theta}^k + \epsilon \frac{1}{2} H'_{jk} \bar{\theta}^j \bar{\theta}^k \bar{\theta}^l} \left\{ 1 + \epsilon \left( N_{jk} \bar{\theta}^j \bar{\theta}^k - \frac{1}{2} H'_{jk} \bar{\theta}^j \bar{\theta}^k \bar{\theta}^l \right) + \frac{\epsilon^2}{2} \left( (N_{jk} \bar{\theta}^j \bar{\theta}^k)^2 + \epsilon H'_{jk} \bar{\theta}^j \bar{\theta}^k \bar{\theta}^l \bar{\theta}^m \right) + \ldots \right\}, \tag{73}
\]

so that the \( I^{(n)} \) and \( N^{(n)} \) are given by expressions akin to

\[
I^{(1)}/N^{(0)} = \int \bar{\theta}^i \left( N_{jk} \bar{\theta}^j \bar{\theta}^k - \frac{1}{2} H'_{jk} \bar{\theta}^j \bar{\theta}^k \bar{\theta}^l \right) e^{-H_{jk} \bar{\theta}^j \bar{\theta}^k /2 + n_j \bar{\theta}^j} d\theta / \int e^{-H_{jk} \bar{\theta}^j \bar{\theta}^k /2 + n_j \bar{\theta}^j} d\theta. \tag{74}
\]

Now, the integrals of the general form

\[
\langle \bar{\theta}^i \ldots \bar{\theta}^m \rangle_p^{(0)} = \int \bar{\theta}^i \ldots \bar{\theta}^m e^{-H_{jk} \bar{\theta}^j \bar{\theta}^k /2 + n_j \bar{\theta}^j} d\theta / \int e^{-H_{jk} \bar{\theta}^j \bar{\theta}^k /2 + n_j \bar{\theta}^j} d\theta, \tag{75}
\]

can be computed with the Wick identity \(^{23}[21]\)

\[
\langle F(\bar{\theta}) \rangle_p^{(0)} = F\left( \frac{\partial}{\partial n} \right) \exp\{ n_i (H^{ij})^{-1} n_j /2 \}; \tag{76}
\]

in particular (again using \( (H^{ij})^{-1} \) to raise indices),

\[
\langle \bar{\theta}^i \rangle_p^{(0)} = n_i, \quad \langle \bar{\theta}^i \bar{\theta}^j \rangle_p^{(0)} = (H^{ij})^{-1} + n_i n_j, \quad \langle \bar{\theta}^i \bar{\theta}^j \bar{\theta}^k \rangle_p^{(0)} = (H^{ij})^{-1} n_k + (H^{ik})^{-1} n_j + n_i (H^{jk})^{-1} n_k + n_i n_j n_k. \tag{77}
\]

\(^{23}\) Another way to organize this computation is to offset the integration variable \( \bar{\theta}^j \) to \( \bar{\theta}^j - (H^{jk})^{-1} n_k = \bar{\theta}^j - n_j \) in Eq. (75), obtaining

\[
\langle \bar{\theta}^i \ldots \bar{\theta}^m \rangle_p^{(0)} = \int (\bar{\theta}^1 + n_1)^{\ldots} (\bar{\theta}^m + n_m) e^{-H_{jk} \bar{\theta}^j \bar{\theta}^k /2} d\theta / \int e^{-H_{jk} \bar{\theta}^j \bar{\theta}^k /2} d\theta;
\]

we can then expand the product in the integrand, bring the \( n^k \) outside the integral, and apply Wick’s theorem [Eq. (63)] to obtain each addend of the form

\[
n^1 \ldots n^{m-1} \int \bar{\theta}^2 \ldots \bar{\theta}^1 e^{-H_{jk} \bar{\theta}^j \bar{\theta}^k /2} d\theta / \int e^{-H_{jk} \bar{\theta}^j \bar{\theta}^k /2} d\theta,
\]

all integrals with odd \( l \) are zero, while the integrals with even \( l \) are given by the sum of all possible pairings of indices into products of \( (H^{ij})^{-1} \).
Unfortunately, the $1/A^4$ (i.e., $\epsilon^2$) corrections to the variance turn out to be rather unwieldy, and belong in a symbolic-manipulation software package rather than on these pages. We content ourselves with the $1/A^2$ correction to the posterior mean (remember that the normalized parameters $\bar{\theta}$ carry an $A$):

$$\langle \bar{\theta} \rangle_p = n^i + \epsilon \left[ n^i n^k - \left( \frac{1}{2} \tilde{h}_{kl} + \bar{h}_{kl} \right) (n^k n^l + \bar{h}^k l) \right] + O(\epsilon^2),$$  \hspace{2cm} (78)

and the $1/A^3$ correction to the variance,

$$\langle \bar{\theta}^2 \rangle_p - \langle \bar{\theta} \rangle_p \langle \bar{\theta} \rangle_p = \tilde{h}^{ij} + \epsilon \left[ n^{ij} + \frac{1}{2} n^i n^j n_{kl} (n^k n^l - \bar{h}^{kl}) - n^k \left( \tilde{h}^{i} k + \bar{h}^{j} k \right) \right]$$

$$- n^i n^j n^k \left( \frac{1}{2} \tilde{h}_{kl} + \bar{h}_{kl} + \frac{1}{2} \tilde{h}_{kl} n^i n^j \right) + O(\epsilon^2).$$  \hspace{2cm} (79)

Thus the $1/A^3$ contribution to the variance does not vanish in any single experiment (unless $n^i = 0$). It does vanish, however, under frequentist average, since it involves products of odd numbers of noises.

VIII. CONCLUSION

In this article I tried to provide a user’s manual for the Fisher information matrix. It seems clear that the Fisher-matrix formalism will continue to be featured prominently in research dealing with the parameter-estimation prospects of future GW observations, because of its compactness and accessibility, and because of the difficulty of computing higher-order corrections and of running full-blown simulations. Yet the three questions posed in the introduction loom over the credibility of Fisher-matrix results, which is all the more worrisome when these results are used to justify choices in science policy or experiment design.

I believe that the recipes provided in this paper to answer the initial questions can help assert (or falsify) the accuracy of the formalism for specific signal models. In particular:

1. As discussed in Sec. IV, ill-conditioned or singular Fisher matrices point to the need for increased numerical accuracy (and occasionally to a case for discarding a parameter or combination of parameters), but more often to suspicions about the appropriateness of the high-SN/LSA limit.

2. The necessity of including prior distributions for source parameters, perhaps as simple as uniform distributions over physically allowed ranges, can be assessed by verifying whether Fisher-matrix results change with the addition of simple Gaussian priors (as we show in Sec. V A), or even better by applying Monte Carlo techniques to the Gaussian integrals implicit in the formalism (as we do in Sec. V B).

3. The detected-signal strength (i.e., the SNR) necessary for Fisher-matrix results to be internally consistent can be evaluated with the likelihood-mismatch criterion of Sec. VI, or (at the price of some algebra) with the higher-order corrections of Sec. VII.

Acknowledgments

M.V.’s work was supported by the LISA Mission Science Office and by the Human Resources Development Fund at the Jet Propulsion Laboratory, California Institute of Technology, where it was performed under contract with the National Aeronautics and Space Administration.

APPENDIX A: LEMMA FOR THE CONDITIONAL AVERAGE OF JOINTLY NORMAL RANDOM VARIABLES

Assume the vector $x_j$ and the scalar $y$ are jointly normal random variables with mean zero and covariance matrix

$$C = \begin{pmatrix} F_{ij} & H_i \\ H_j & G \end{pmatrix}. \hspace{2cm} (A1)$$

From the standard Frobenius–Schur formula for the inverse of a block matrix [22],

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1} BS^{-1}CA^{-1} & -A^{-1}BS^{-1}A^{-1} \\ -S^{-1}CA^{-1} & S^{-1} \end{pmatrix} \hspace{2cm} (A2)$$
(with $S_A = D - CA^{-1}B$ the Schur complement of $A$), we find
\[
\mathcal{C}^{-1} = \begin{pmatrix}
F_{ij}^{-1} + S_A^{-1}(F_{ik}^{-1}H_k)(F_{jl}^{-1}H_l) & -S_A^{-1}(F_{ij}^{-1}H_k) \\
-S_A^{-1}(F_{ik}^{-1}H_k) & S_A^{-1}
\end{pmatrix},
\]  
(A3)
since in our case $F_{ij}^{-1}$ is symmetric and $S_A$ is the scalar $G - (F_{ij}^{-1}H_iH_j)$. Now, the joint distribution of $x_j$ and $y$ is given by
\[
p(x, y) \propto \exp - \left\{ (x, y) \cdot \mathcal{C}^{-1} \cdot \begin{pmatrix} x_i \\ y \end{pmatrix} \right\} / 2,
\]  
(A4)
while the conditional distribution of $y$ given $x_j$ is $p(y|x) = p(x, y)/p(x) = p(x, y)/[\int p(x, y) dy]$. Since however $p(x)$ can be a function only of $x$, by the properties of Gaussian integrals it must be that $p(x) \propto \exp (-\cdots)_{ij}x_ix_j$. It follows that $p(y|x)$ must be of the form
\[
p(y|x) \propto \exp - \left\{ S_A^{-1}y^2 - 2S_A^{-1}(x, H_j)y + \langle \cdots \rangle_{ij}x_ix_j \right\} / 2,
\]  
(A5)
from which, by inspection, we conclude that
\[
\langle y \rangle_{x_i} = \int y p(y|x) dy = x_iF^{-1}_{ij}H_j
\]  
(A6)
and that
\[
\text{var}_{x,y} = \int (y - \langle y \rangle_{x_i})^2 p(y|x) dx = S_A = G - F^{-1}_{ij}H_iH_j.
\]  
(A7)