

Bayesian estimation of correlation functions

Ángel Gutiérrez-Rubio,¹ Juan S. Rojas-Arias¹, Jun Yoneda², Seigo Tarucha^{1,3}, Daniel Loss,^{1,3,4} and Peter Stano^{3,5,*}

¹RIKEN Center for Quantum Computing, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

²Tokyo Tech Academy for Super Smart Society, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8552, Japan

³RIKEN Center for Emergent Matter Science, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

⁴Department of Physics, University of Basel, Klingelbergstrasse 82, CH-4056 Basel, Switzerland

⁵RCQI, Institute of Physics, Slovak Academy of Sciences, Dubravska cesta 9, 845 11 Bratislava, Slovakia



(Received 7 July 2022; accepted 1 November 2022; published 6 December 2022)

We apply Bayesian statistics to the estimation of correlation functions. We give the probability distributions of auto- and cross-correlations as functions of the data. Our procedure uses the measured data optimally and informs about the certainty level of the estimation. Our results apply to general stationary processes and their essence is a nonparametric estimation of spectra. It allows one to better understand the statistical noise fluctuations, assess the correlations between two variables, and postulate parametric models of spectra that can be further tested. We also propose a method to numerically generate correlated noise with a given spectrum.

DOI: [10.1103/PhysRevResearch.4.043166](https://doi.org/10.1103/PhysRevResearch.4.043166)

I. INTRODUCTION

In this article, we propose a rigorous method to estimate correlation functions. By “rigorous” we mean a nonparametric [1] estimation based on Bayesian statistics, being free of *ad hoc* procedures, *ansätze*, or assumptions going beyond the observed data. We use this notion to distinguish our approach from the orthodox methods, as we explain below. To motivate our work, we briefly introduce the concept of correlation and discuss why its estimation can benefit from a Bayesian approach.

Correlation is intimately related to causality and prediction. For this reason, in physics one often deals with detecting, quantifying, explaining, and testing correlations. A common way to analyze them is through correlation functions. Roughly, a correlation function $\langle A(t)B(t') \rangle$ is the expected value of the product of two variables (observables) A and B measured at times t and t' .

Correlation functions are ubiquitous in—and often the cornerstone of—different theories and formalisms. They play a major role in the study of noise, particularly regarding *spectra* [2–4]. Actually, a spectrum is the Fourier transform of a correlation function. Spectral analysis has been the main tool for the study of noise in very different contexts for more than a century [5–11]. In the past, the application of spectral analysis has ranged from the pioneering works in vacuum tubes [5] and electric circuits [6] to particle scattering [12], earthquakes and tides, music and geographic features [7,8], or astronomy [11,13,14], to name a few examples. Among the remarkable

discoveries made along that path, one can highlight the almost universal appearance of $1/f$ or *flicker* noise, whose origin is still under debate [8–11]. Today, noise is also in the spotlight in cutting-edge fields like quantum computing: It requires using error correction [15], it affects coherence times, and drives spatial correlations between qubits [16,17]. With emerging fields exploiting noisy intermediate-scale quantum (NISQ) devices [18,19], there is a need for a better understanding of the properties and origin of noise.

The *estimation* of correlation functions emerges then as a relevant task. However, the current standard techniques to estimate correlation functions can be subject to improvement. By “standard”, “traditional”, or “orthodox” we mean non-Bayesian. To cast more light on our motivations, we point out some issues in traditional parameter estimation. Later, we introduce Bayesian statistics as the way to sort them out and achieve a better (and actually optimal, as we argue below) estimation of correlation functions, which is the main goal of our paper.

Standard statistics estimates a variable by yielding a “best guess” of its value, called an *estimator*, and a *confidence interval* that reflects its uncertainty. Two familiar examples of estimators are the mean and the median. In the context of spectral analysis, we can mention the *periodogram* [1] (defined and discussed in the main text) as the most common estimator of the spectrum. A fundamental issue in standard estimation is the arbitrariness in the choice of an estimator. For example, there is no universally applicable ordering, preferring the mean over the median or vice versa. Several arguments can support either according to different criteria, but still, in the traditional approach, one estimator has to be chosen. Even the often adopted *unbiased* estimators [1] have drawbacks under certain circumstances [20]. Ultimately, there is no universal “best estimator”.

A similar arbitrariness also affects the choice of parameters to quantify errors, test hypotheses, and, most importantly for the subject of this article, assess correlations. For example,

*peter.stano@riken.jp

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

we will discuss in the main text the limitations of Pearson's r coefficient [21] to quantify the correlation between two variables. There is always the doubt that a different parameter could serve that purpose better. The search for "better" r coefficients leads to a series of more complicated definitions [1] like *nonparametric* or *rank correlation* [1], that end up raising the same kind of doubts: "It is precisely the uncertainty in interpreting the significance of the linear correlation coefficient r that leads us to the important concepts of *nonparametric* or *rank correlation*" [1]. Moreover, the statistical fluctuations of these coefficients—necessary to assess the confidence interval—is another matter that quickly gets similarly nontransparent: New (and potentially arbitrary) parameters are needed to analyze these fluctuations, making the problem scale in complexity. The whole process becomes less transparent and more dependent on assumptions not always met in practice, such as the asymptotic conditions required by the central limit theorem. For example, Ref. [22] studies the statistical properties of $1/f$ noise analyzing the "variance of the variance" of the periodogram, namely "the error of the error." Quantities like these are difficult to interpret and add elements of arbitrariness. The variance of the variance of the variance would be the next variable to look at in a presumably never ending progression [23].

This arbitrariness manifests a fundamental issue in traditional estimation: One cannot guarantee the optimal usage of the data, or optimal processing of information, when performing any inference. By inference, we mean, for instance, estimating a variable, assessing correlations, or testing a hypothesis. And with "optimal usage of the data," we mean that no relevant information contained in the data about the inference is lost in the inference process.

Bayesian probability theory [20,25], or Bayesian statistics, puts an end to these issues. This theory is solidly built as an extension of logic, departing from the minimal Cox's axioms—which include Bayes' rule, hence the name [20,25]. Throughout its construction, there is no arbitrary assumption or *ad hoc* choice of parameters, of estimators, or of statistical tests. Instead, Bayesian statistics focuses on the calculation of the *probability distributions* that encode all our *knowledge*—including measured data—, and nothing but our knowledge, about a variable [26]. In this way, Bayesian statistics guarantees the optimal usage of the data when making inferences. The interested reader is referred to Ref. [20] for a complete overview of fundamental historical problems and paradoxes of the orthodox theory that only a Bayesian approach can resolve [27].

Consequently, Bayesian statistics can enhance parameter estimation. Surprisingly, beyond a few exceptions in some specific problems [4,29,30], the Bayesian estimation of correlation functions has not been fully developed. In this article, we take up this task. Our paper can find the following applications. First and foremost, in the context of spectral analysis, our results allow one to estimate the spectrum *including the quantification of the corresponding statistical estimation errors* (encoded in probability distributions or in the error bars calculated from them). We emphasize that these statistical uncertainty measures (such as error bars) are produced also in the *nonparametric* estimation, which is the topic of this article. Although we do not pursue it here, one could in turn

use the results of a nonparametric estimation for a further parametric estimation, that is for estimation of parameters of a specific model of the spectrum [4]. Second, our results also allow one to discuss with generality several statistical properties of noise and the periodogram, like the signal-to-noise ratio (whose value we prove to be universal in the nonparametric estimation). Third, we develop a formalism to judge—with its corresponding uncertainty—the correlation between two generic variables, without involving arbitrary parameters. Finally, we provide a method to numerically generate correlated noise with an arbitrary spectrum.

The paper is organized as follows. In Sec. II we introduce the Bayesian formalism through a simple example and set the basis for the estimation of correlation functions. In Sec. III we study auto-correlations, namely the correlation functions of a time-dependent variable with itself. This includes a subsection discussing the statistical properties of the periodogram and the generation of uncorrelated noise. Section IV is parallel to Sec. III but focuses on the correlation between two variables, the cross-correlation. It also gives a method to generate correlated noise with an arbitrary spectrum. In Sec. V we illustrate our results numerically and in Sec. VI B we give some guidelines on estimating continuous spectra. Section VII contains the conclusions.

To improve the flow of the main text, topics that are technical or can be discussed separately are delegated to appendices: In Appendix A we prove that the multivariate Gaussian distribution maximizes entropy, in Appendix B we derive the prior distributions from invariance principles, in Appendix C we derive Eq. (24), Appendix D contains a useful identity for the zero-frequency case, and in Appendix E we give the estimating distribution for unnormalized correlation strength. The next three appendices are especially notable for practical usage of our formulas: in Appendix F we explain how to merge points in a spectral plot, superseding artificial "windowing" or batching the data, in Appendix G we explain how to rigorously correct the estimating distributions for errors on input, and in Appendix H we give the central-limit versions of the estimating distributions. Finally, in Appendix I we comment on the word "nonparametric".

II. VARIANCE IN ONE DIMENSION

In this section, we use a simple example to introduce Bayesian statistics and to set the basis of parameter estimation. However, the results presented here go beyond mere illustration purposes and directly apply to spectral analysis. The next sections will make the connection clear.

Before going into details, we recall a few definitions and conventions from probability theory and logic. Everything inside the probability symbol P is to be interpreted as a logical statement. $P(a)$ represents the probability that the statement a is true. $P(a \cup b)$ is the probability that either a or b (or both) are true. $P(a \cap b)$ is the probability that both a and b are true. However, usually the logical "and" operator is substituted by a comma or even omitted: $P(ab) = P(a, b) \equiv P(a \cap b)$. The *conditional probability* $P(a|b) \equiv P(ab)/P(b)$ denotes the probability that a is true provided that b is true. When considering the probability that a variable x has a value x' , one should write $P(x = x')$ (note that $x = x'$ is a logical statement), but

the notation is usually simplified by omitting the variable name: $P(x') \equiv P(x = x')$. At last, the mean of the expression $f(x)$ with the variable x having a probability distribution $P(x)$ is

$$\langle f(x) \rangle \equiv \int_{-\infty}^{\infty} dx P(x) f(x). \quad (1)$$

The mean and variance of x , the two central quantities in this article, are defined as

$$\nu \equiv \langle x \rangle, \quad (2a)$$

$$\Lambda \equiv \langle (x - \nu)^2 \rangle, \quad (2b)$$

respectively.

Having set the notation, consider a scalar variable x . Assume we want to estimate its mean ν and variance Λ given a data sample

$$\hat{x} = \{x^{(0)}, \dots, x^{(M-1)}\}$$

that contains M values $x^{(m)}$. One cannot directly apply the definitions $\nu = \langle x \rangle$ and $\Lambda = \langle (x - \nu)^2 \rangle$ because $P(x)$ in Eq. (1) is unknown. One usually resorts to standard formulas like

$$\bar{\nu} = \frac{1}{M} \sum_{m=0}^{M-1} x^{(m)}, \quad \bar{\Lambda} = \frac{1}{M} \sum_{m=0}^{M-1} (x^{(m)} - \bar{\nu})^2. \quad (3)$$

However, this estimation of ν and Λ is not very informative. To begin with, Eqs. (3) do not report on the confidence of the result and how this confidence depends on the size M of the sample. In other words, $\bar{\nu}$ and $\bar{\Lambda}$ are just specific estimator choices or “best guesses” of ν and Λ . We aim at more than that. Our goal is to get the probability distribution of ν and Λ given the observed data, namely, $P(\nu|\hat{x})$ and $P(\Lambda|\hat{x})$. By *estimating* the parameters ν and Λ , we mean computing these distributions. We will refer to them as the *estimating distributions* to make a clear distinction to the *sampling distributions* defined below. The estimating distributions encode all the information in the data, and nothing but the information in the data, that is relevant for the knowledge of ν and Λ . This includes, for example, the size of the data sample M .

Our goal is to calculate the distributions $P(\nu|\hat{x})$ and $P(\Lambda|\hat{x})$, but we start by looking at the *sampling distribution* $P(\hat{x}|\nu\Lambda)$. Its meaning is the following: If we knew *only* the mean ν and variance Λ of a variable x , what would be the probability to get the data sample \hat{x} ?

Shannon’s entropy theorem [28,31,32] is the key to construct sampling distributions. It is about the entropy S of a distribution $P(x)$, defined as

$$S = - \int_{-\infty}^{\infty} dx P(x) \log P(x).$$

The theorem proves that S measures the *information* about the variable x encoded in $P(x)$. Therefore, assuming nothing about x other than its mean ν and variance Λ , the distribution $P(x|\nu\Lambda)$ representing the knowledge about x must have maximal entropy allowed by the constraints

$$\nu = \int_{-\infty}^{\infty} dx P(x|\nu\Lambda) x, \quad \Lambda = \int_{-\infty}^{\infty} dx P(x|\nu\Lambda) (x - \nu)^2.$$

In Appendix A we prove that it is a Gaussian,

$$P(x|\nu\Lambda) = \frac{1}{\sqrt{2\pi\Lambda}} \exp\left(-\frac{(x - \nu)^2}{2\Lambda}\right). \quad (4)$$

It is important to emphasize the following. This result does not mean that the “true” distribution of x is Eq. (4). It means that our *knowledge* of x is described by Eq. (4). This interpretation is one of the core concepts of Bayesian statistics. In contrast, standard statistics would consider Eq. (4) an approximation or *ansatz*, even when it is not. The rest of the estimation process—and its validity for a generic variable x —relies on Eq. (4), and hence on this subtle point.

Equation (4) gives the probability to observe the single value x based on ν and Λ as the only knowledge about x . The probability to observe the whole data sample \hat{x} can be obtained from the definition of conditional probabilities, that is, $P(ab) = P(a)P(b|a)$. When b is logically independent on a , namely when $P(b|a) = P(b)$, we have $P(ab) = P(a)P(b)$. Therefore, if the values in the data sample \hat{x} are logically independent,

$$P(\hat{x}|\nu\Lambda) = \prod_{m=0}^{M-1} P(x^{(m)}|\nu\Lambda). \quad (5)$$

With this and Eq. (4), we have the sampling distribution $P(\hat{x}|\nu\Lambda)$. However, this is not what we want to calculate. For the estimation of ν and Λ , we rather need $P(\nu|\hat{x})$ and $P(\Lambda|\hat{x})$. It is possible to get these estimating distributions from the sampling distribution $P(\hat{x}|\nu\Lambda)$ by using elementary probability theory. We need to apply just two rules. The first is

$$\begin{aligned} \sum_i P(b_i) &= 1 \text{ and } P(b_i \cap b_j) = 0 \quad \forall i \neq j \\ \Rightarrow P(a) &= \sum_i P(a \cap b_i) \quad \forall a, \end{aligned} \quad (6)$$

which can be easily derived from (i) the axiom

$$P(a \cap b) = 0 \Rightarrow P(a \cup b) = P(a) + P(b);$$

(ii) the equivalence between the logical statement a and $a \cap (\cup_i b_i)$, because $\cup_i b_i$ is always true by virtue of $\sum_i P(b_i) = 1$ and $P(b_i \cap b_j) \propto \delta_{ij}$; and (iii) the identity $a \cap (\cup_i b_i) = \cup_i (a \cap b_i)$. The rule of Eq. (6) allows us to write

$$P(\nu|\hat{x}) = \int_0^\infty d\Lambda P(\nu\Lambda|\hat{x}),$$

$$P(\Lambda|\hat{x}) = \int_{-\infty}^\infty d\nu P(\nu\Lambda|\hat{x}).$$

The second rule we need is the Bayes’ rule,

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)}.$$

When applied to the integrands, one gets

$$P(\nu|\hat{x}) = \int_0^\infty d\Lambda \frac{P(\hat{x}|\nu\Lambda)P(\nu\Lambda)}{P(\hat{x})}, \quad (7a)$$

$$P(\Lambda|\hat{x}) = \int_{-\infty}^\infty d\nu \frac{P(\hat{x}|\nu\Lambda)P(\nu\Lambda)}{P(\hat{x})}. \quad (7b)$$

These equations express the estimating distributions as functions of the sampling distribution $P(\hat{x}|\nu\Lambda)$ and constitute the essence of parameter estimation from the perspective of Bayesian statistics together with the Shannon's entropy theorem.

The steps that led to Eq. (7) will be repeatedly used throughout the article. We can look at $P(\hat{x})$ in the denominators as a mere normalization constant. $P(\nu\Lambda)$ is the *prior* probability for the mean and variance, and it represents our knowledge about ν and Λ previous to the acquisition or analysis of any data. Assigning priors has been subject to extensive discussions, see Ref. [20]. We follow the approach advocated therein, and assign priors based on the requirements of symmetry (invariance under certain transformations). Accordingly, in Appendix B we prove that $P(\nu\Lambda) \propto 1/\Lambda$ keeps $P(\nu\Lambda)$ invariant under “frame” transformations of the variable x being the independence of $P(\nu\Lambda)$ on the units and the reference value chosen to measure x relative to [33].

Performing the integrations, we get Eq. (7) in a closed form

$$P(\nu|\hat{x}) = \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{M-1}{2})\sqrt{\pi\bar{\Lambda}}} \left(\frac{\bar{\Lambda}}{\bar{\Lambda} + (\nu - \bar{\nu})^2} \right)^{\frac{M}{2}}, \quad (8a)$$

$$P(\Lambda|\hat{x}) = \frac{1}{\Gamma(\frac{M-1}{2})} \left(\frac{M}{2} \frac{\bar{\Lambda}}{\Lambda} \right)^{\frac{M-1}{2}} \frac{1}{\Lambda} \exp\left(-\frac{M\bar{\Lambda}}{2\Lambda}\right). \quad (8b)$$

Remarkably, these distributions depend only on two variables that are functions of the data, or *sufficient statistics*, namely $\bar{\nu}$ and $\bar{\Lambda}$. Not surprisingly, they are given by the well-known formulas of Eqs. (3). However, the estimating distributions allow one to evaluate any desired estimator. For example, the *maximum-likelihood estimators* (mle) are the values of ν and Λ that maximize $P(\nu|\hat{x})$ and $P(\Lambda|\hat{x})$, respectively. Equation (8) gives

$$\text{mle}(\nu) = \bar{\nu}, \quad \text{mle}(\Lambda) = \frac{M}{M+1} \bar{\Lambda}.$$

Interestingly, $\text{mle}(\Lambda)$ significantly differs from $\bar{\Lambda}$ for low values of M . In a similar way, we could use the estimating distributions to calculate, for example, the confidence intervals [34].

Before illustrating the distributions, we examine Eq. (8) in certain special limits. The case $M = 1$ corresponds to measuring a single value $x^{(0)}$, giving $\bar{\nu} = x^{(0)}$ and $\bar{\Lambda} = 0$ by Eq. (3). Evaluating Eq. (8a) with $\bar{\Lambda} = 0$ and $M = 1$ is undefined, but let us take the limits $\bar{\Lambda} \rightarrow 0$ and then $M \rightarrow 1$:

$$P(\nu|x^{(0)}) = \lim_{M \rightarrow 1} \frac{1}{\Gamma(\frac{M-1}{2})} \frac{1}{|\nu - \bar{\nu}|^M}.$$

That is, the estimating distribution for the mean that we deduce from a single data point is centered at that point and falls off as $1/|x|$ symmetrically to both sides. Keeping M at a finite, albeit possibly small, distance above 1 is a way of regularization of this non-normalizable distribution [35]. The same happens with Eq. (8b): For $M = 1$, the formula should be interpreted as a limit towards a non-normalizable distribution $1/\Lambda$, with parameters $M \rightarrow 1$ and $\bar{\Lambda} \rightarrow 0$ possibly providing effective cut-offs for the divergencies at both $\Lambda \rightarrow 0$ and $\Lambda \rightarrow \infty$ tails of the distribution. Since a single point delivers

no information about the variance, our estimating distribution equals to the prior distribution.

Remarkably, one can meaningfully interpret Eq. (8) even for $M = 0$, even though the expressions in Eq. (3) are undefined. Taking them as regularization parameters instead, we get that for $M \rightarrow 0$ also Eq. (8a) reduces to the corresponding prior, $P(\nu|\text{no data}) = P(\nu)$. Recovering the prior distributions when the data deliver no information is an appealing property of Eq. (8).

As a final example, consider that two points were measured, $M = 2$, and they came out exactly the same, so that $\bar{\Lambda} = 0$. Equation (8a) now becomes a regularized delta function centered at $\bar{\nu}$ with the width $\bar{\Lambda} \rightarrow 0$. In reality, the values $x^{(m)}$ are measured with some finite precision. We therefore obtain another natural result that if $\bar{\Lambda} = 0$, one can be sure that $\nu = \bar{\nu}$, but only up to the precision with which $\bar{\Lambda} = 0$ is valid, which is never an infinite precision in reality.

As stressed in Ref. [20], this behavior is the hallmark of Bayesian statistics. The results such as Eq. (8), which are based on the Bayes rule and other rules backed by the probability-theory axioms used in Eq. (7), remain valid even under “extreme” cases. We will not repeatedly alert the reader on this fact for the formulas to come. Nevertheless, in Sec. VI B we discuss some paradoxes that occur for $M = 1$, which are of a different type.

To conclude the discussion of their general aspects, we note that having the estimating distributions $P(\nu|\hat{x})$ and $P(\Lambda|\hat{x})$, we know everything that can be known about ν and Λ from the observed data \hat{x} and—we emphasize again—nothing more than the data and the prior invariance principles: no model on x or ansatz on the “true” $P(x)$ has been invoked. These estimating distributions give more information than just the estimators $\bar{\nu}$ and $\bar{\Lambda}$: Spread or narrow, they express our confidence in the knowledge of ν and Λ . We will illustrate it with a figure shortly, but let us first consider a particular case relevant for spectral analysis.

Assume that the mean of x is known, $\nu = \nu_0$. Proceeding as before, we get the estimating distribution of the variance in the form

$$P(\Lambda|\nu = \nu_0, \hat{x}) = \frac{P(\hat{x}|\nu = \nu_0, \Lambda)P(\Lambda)}{P(\nu = \nu_0, \hat{x})} = \frac{1}{\Gamma(\frac{M}{2})} \left(\frac{M}{2} \frac{\bar{\Lambda}}{\Lambda} \right)^{\frac{M}{2}} \frac{1}{\Lambda} \exp\left(-\frac{M\bar{\Lambda}}{2\Lambda}\right). \quad (9)$$

Compared to Eq. (8b), there is a difference in the power exponent. Also, $\bar{\Lambda}$ should be here calculated according to the second expression in Eq. (3) replacing $\bar{\nu} \rightarrow \nu_0$. Although not needed here, for later convenience we note that this short calculation can be performed also in the following way,

$$P(\Lambda|\nu = \nu_0, \hat{x}) = \int d\nu \frac{P(\hat{x}|\nu\Lambda)P(\Lambda)}{P(\nu = \nu_0, \hat{x})} P(\nu), \quad (10)$$

and using the prior $P(\nu) = \delta(\nu - \nu_0)$.

As more important for spectral estimation, we plot Eq. (9) in Fig. 1 for different values of M and $\bar{\Lambda} = 1$. The information content of the full estimating distribution can be appreciated from the figure. Despite all the plotted distributions corresponding to the same variance estimator ($\bar{\Lambda} = 1$), they clearly convey a different knowledge on Λ . One way to illustrate it is

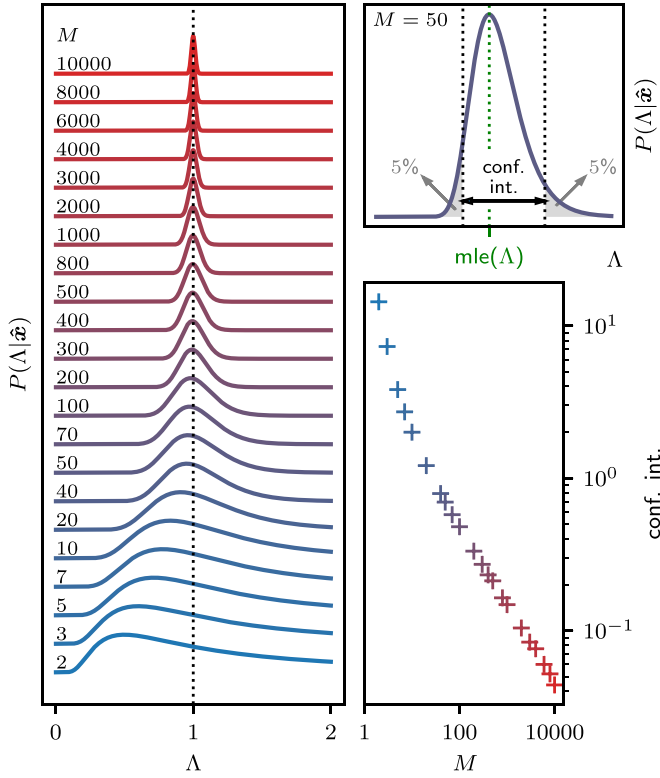


FIG. 1. Left: Estimating distributions—Eq. (9)—for $\bar{\Lambda} = 1$ and different sizes M of the data sample. The curves have been shifted and scaled in the vertical axis for the sake of visibility. Right: 90%-confidence intervals for the estimating distributions. The top figure shows how the maximum-likelihood estimator and confidence intervals are defined. The bottom figure plots the relation between confidence and sample size. The color code is consistent with the plot on the left.

to look at the dependence of the confidence interval size on the sample size M , plotted in the right panel of Fig. 1. The change of the information content with the size of the data sample M is obvious.

As a final remark, we emphasize that in this section, with the main result being Eqs. (8) and (9), we have indeed estimated a correlation function: The variance $\Lambda = \langle (x - \langle x \rangle)^2 \rangle$ is the correlation function of x with itself. What we do next is to extend this procedure to correlation functions of (possibly two different) variables that evolve in time. The essence of the parameter-estimation formalism remains identical: First, we find the sampling distribution that maximizes the entropy. Then, we apply Bayes' rule to calculate the estimating distribution. The main difference is the passage to Fourier space in order to simplify both distributions.

III. AUTO-CORRELATIONS

This section extends the estimation of Sec. II to a real variable A that evolves in time. This scenario includes the study of noise. The results that we present allow one to estimate the auto-correlation matrix of A (defined below) with Bayesian statistics. We establish connections with the common estimators used in the literature, such as the periodogram. Along the lines of Sec. II, our results contain more information.

Consider M independent batches of n measurements of A in time,

$$\hat{\mathbf{A}} \equiv (\mathbf{A}^{(0)}, \dots, \mathbf{A}^{(M-1)}), \quad \mathbf{A}^{(m)} \equiv (A_0^{(m)}, \dots, A_{n-1}^{(m)})^T.$$

The super- and subscripts are the batch and time index, respectively. Different batches correspond to uncorrelated copies or runs of the experiment. Let the time interval Δ between $A_j^{(m)}$ and $A_{j+1}^{(m)}$ be the same for all m and j [37]. Let

$$\mu = \langle A_j \rangle \quad (11a)$$

be the mean of the variable A (the same for all j) and $\hat{\Sigma}$ the correlation matrix, with components

$$\hat{\Sigma}_{jk} = \langle A_j A_k \rangle - \mu^2. \quad (11b)$$

Our goal now is to estimate these parameters.

We adopt two approximations. First, we impose periodic boundary conditions, $A_{j+n} = A_j$. These conditions are implicit in the discrete Fourier transforms used throughout this paper. Second, we consider *stationary noise* so that $\hat{\Sigma}_{jk}$ only depends on $j - k$. The scalar function Σ_j defined by $\hat{\Sigma}_{ij} = \Sigma_{i-j}$ is called the *auto-correlation function*, or simply correlation function. Stationary noise implies that $\Sigma_j = \Sigma_{-j}$. These two commonly adopted approximations are necessary for the correlation matrix to be diagonal in Fourier space.

In parallel with Sec. II, we start by giving the sampling distribution of the variable A . Next, we estimate the mean μ and the correlation matrix $\hat{\Sigma}$, and conclude by discussing some spectral properties of noise.

A. Sampling distribution

Since by definition the M data batches are independent, the sampling distribution given only μ and $\hat{\Sigma}$ reads

$$P(\hat{\mathbf{A}}|\mu, \hat{\Sigma}) = \prod_{m=0}^{M-1} P(\mathbf{A}^{(m)}|\mu, \hat{\Sigma}), \quad (12)$$

with, as we prove in Appendix A, the multivariate Gaussian

$$P(\mathbf{A}^{(m)}|\mu, \hat{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\hat{\Sigma}|}} \times \exp\left(-\frac{1}{2}(\mathbf{A}^{(m)} - \mu)^T \hat{\Sigma}^{-1}(\mathbf{A}^{(m)} - \mu)\right) \quad (13)$$

and $\mu = \mu(1, \dots, 1)^T$. These expressions are analogous to Eqs. (4) and (5). We insist on the fact that this is not an approximation of the distribution of A . Rather, it describes our knowledge based on the mean μ and correlation matrix $\hat{\Sigma}$.

It is convenient to perform a coordinate change to simplify the sampling distribution. The translational invariance encoded in $\hat{\Sigma}_{ij} = \Sigma_{j-i}$ hints on switching to Fourier space. The Fourier transform of each batch (m) reads

$$\alpha_j^{(m)} \equiv \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} A_k^{(m)} \exp\left(-i \frac{2\pi jk}{n}\right). \quad (14)$$

In agreement with our previous notation, we define

$$\hat{\alpha} \equiv (\alpha^{(0)}, \dots, \alpha^{(M-1)}), \quad \alpha^{(m)} \equiv (\alpha_0^{(m)}, \dots, \alpha_{n-1}^{(m)})^T.$$

Denoting by \hat{F} the Fourier-transform matrix, with components

$$\hat{F}_{jk} \equiv \frac{1}{\sqrt{n}} \exp\left(-i \frac{2\pi jk}{n}\right) \quad (15)$$

and thus unitary, we can write $\alpha^{(m)} = \hat{F}A^{(m)}$. Similarly, we define the vector $\mathbf{v} \equiv \hat{F}\mu$. All components of this vector except the first are zero, $v_k = v\delta_{k,0}$, where, for later convenience, we define $v \equiv \mu\sqrt{n}$. As mentioned, with periodic boundary conditions the correlation matrix is diagonal in Fourier space,

$$\hat{\Lambda} \equiv \hat{F} \hat{\Sigma} \hat{F}^\dagger, \quad \hat{\Lambda}_{jk} = \delta_{jk} \Lambda_k, \quad \Lambda_k \equiv \sum_{j=0}^{n-1} \Sigma_j e^{-i2\pi jk/n}. \quad (16)$$

Notice the difference in notation between the correlation matrix $\hat{\Lambda}$, with components $\hat{\Lambda}_{jk}$, and the scalar function (of k) Λ_k . Λ_k gets the same name as Σ_j , namely auto-correlation function or simply correlation function. Since it is the Fourier transform of Σ_j , it is also called spectrum [41].

Another notation caveat is pertinent. The bijections between $A^{(m)}$ and $\alpha^{(m)}$, and between $\hat{\Sigma}$ and $\hat{\Lambda}$, allow one to write the sampling distribution in the following equivalent forms:

$$P(\hat{A}|\mu\hat{\Sigma}) = P(\hat{A}|\mu\hat{\Lambda}) = P(\hat{\alpha}|\mu\hat{\Sigma}) = P(\hat{\alpha}|\mu\hat{\Lambda}).$$

Basically, data (in real or Fourier space) is on the left of the bar, whereas the mean and correlation matrices (in real or Fourier space) are on the right. These changes are trivial in the Bayesian formalism as introduced in Sec. II. We will express the sampling distribution in any of these forms according to our convenience. Also, distributions like $P(\alpha_k^{(m)}|\mu\Lambda_k)$ may be referred to as “sampling distributions” with no risk of confusion. Analogously, the estimating distributions will be written in different equivalent forms, easily identifiable because the mean and correlation matrix will lie on the left of the bar, and the data on the right. In a similar spirit, we can replace the mean μ by the vector \mathbf{v} or its only nonzero component v .

The sampling distribution of Eq. (12) becomes much simpler in the α coordinates,

$$P(\hat{\alpha}|\mu\hat{\Lambda}) = \prod_{k=0}^{\lfloor n/2 \rfloor} \prod_{m=0}^{M-1} P(\alpha_k^{(m)}|v_k\Lambda_k), \quad (17)$$

with $\lfloor x \rfloor$ the largest integer smaller or equal to x ,

$$P(\alpha_k^{(m)}|v_k\Lambda_k) = \left(\frac{d_k}{\pi\Lambda_k}\right)^{d_k} \exp\left(-d_k \frac{|\alpha_k^{(m)} - v_k|^2}{\Lambda_k}\right), \quad (18)$$

and [42]

$$d_k \equiv \begin{cases} 1/2 & \text{if } k = 0, \\ 1/2 & \text{if } n \text{ is even and } k = n/2, \\ 1 & \text{otherwise.} \end{cases}$$

The k th component of the vector \mathbf{v} is the mean of the variable α_k ; out of these, the only one with a nonzero mean is α_0 . This remark will be relevant in the next subsection for the choice of priors. It also allows us to write $P(\alpha_k^{(m)}|v_k\Lambda_k) = P(\alpha_k^{(m)}|\Lambda_k)$ for $k \neq 0$. A last detail worth discussing about Eq. (17) is that k ranges from 0 to $\lfloor n/2 \rfloor$ (and not n). This, as well as the

origin of the factor d_k , is due to (i) $\alpha_k^{(m)} = (\alpha_{-k}^{(m)})^*$, because the variable A is real and (ii) $\Lambda_{-k} = \Lambda_k$, with $\Lambda_k \in \mathbb{R}$, because $\Sigma_j \in \mathbb{R}$ is even in j . (In these relations, the subscript indexes are understood mod n .) In other words, the independent coordinates in $\alpha^{(m)}$ can be taken as $\alpha_k^{(m)}$ for $k = 0, \dots, \lfloor n/2 \rfloor$, and, similarly, $\hat{\Lambda}$ is fully parametrized by the independent values Λ_k with $k = 0, \dots, \lfloor n/2 \rfloor$. The sampling distribution already encodes this information.

B. Estimation of auto-correlations

The estimation of the correlation matrix $\hat{\Sigma}$ requires the calculation of $P(\hat{\Sigma}|\hat{A})$. Equivalently, we can estimate its transform $\hat{\Lambda}$ with $P(\hat{\Lambda}|\hat{\alpha})$. This is easier because $P(\hat{\alpha}|\mu\hat{\Lambda})$ manifestly factors in k , see Eq. (17). Assuming a factored prior (as argued in Appendix B)

$$P(\mu\hat{\Lambda}) = \prod_{k=0}^{\lfloor n/2 \rfloor} P(v_k\Lambda_k)$$

and using Bayes' rule, it is easy to see that $P(\hat{\Lambda}|\hat{\alpha})$ also factors,

$$\begin{aligned} P(\mu\hat{\Lambda}|\hat{\alpha}) &= \frac{P(\hat{\alpha}|\mu\hat{\Lambda})P(\mu\hat{\Lambda})}{P(\hat{\alpha})} \propto P(\hat{\alpha}|\mu\hat{\Lambda})P(\mu\hat{\Lambda}) \\ &= \prod_{k=0}^{\lfloor n/2 \rfloor} P(\alpha_k|v_k\Lambda_k)P(v_k\Lambda_k) \propto \prod_{k=0}^{\lfloor n/2 \rfloor} P(v_k\Lambda_k|\alpha_k), \end{aligned} \quad (19)$$

where we denoted $\alpha_k \equiv \{\alpha_k^{(0)}, \dots, \alpha_k^{(M-1)}\}$. Equation (19) greatly simplifies the estimation of $\hat{\Lambda}$: It decouples the estimation of the matrix $\hat{\Lambda}$, dependent on all the (Fourier transformed) data $\hat{\alpha}$, into the estimation of the diagonal elements Λ_k for $k = 0, \dots, \lfloor n/2 \rfloor$, each of which only depends on α_k . In this way, the problem reduces to the one solved in Sec. II. This connection is not unexpected: Since the change to Fourier space diagonalizes $\hat{\Sigma}$, the multivariate Gaussian of Eq. (13) is mapped to a product of one-variable Gaussians. Thus, spectral analysis reduces to the estimation of the variance in each of the $\lfloor n/2 \rfloor + 1$ independent one-dimensional problems labeled by $k = 0, \dots, \lfloor n/2 \rfloor$.

One could now in principle read off the result from Eqs. (8) and (9), noting that Eq. (4) gives Eq. (18) upon replacing $1/2 \rightarrow d_k$. However, it might be easier to repeat the calculation: Λ_k can be estimated with Bayes' rule,

$$\begin{aligned} P(\Lambda_k|\alpha_k) &= \int dv_k \frac{P(\alpha_k|v_k\Lambda_k)P(v_k\Lambda_k)}{P(\alpha_k)} \\ &\propto P(\Lambda_k) \prod_{m=0}^{M-1} \int dv_k P(\alpha_k^{(m)}|v_k\Lambda_k)P(v_k), \end{aligned} \quad (20)$$

the sampling distribution $P(\alpha_k^{(m)}|v_k\Lambda_k)$ of Eq. (18) and the prior (cf. Appendix B)

$$\begin{aligned} P(v_k\Lambda_k) &\propto \delta(v_k) \times 1/\Lambda_k \quad \text{if } k \neq 0, \\ P(v_k\Lambda_k) &\propto \text{const} \times 1/\Lambda_k \quad \text{if } k = 0, \end{aligned}$$

where in the first line δ is the Dirac delta function. The resulting estimating distribution of Λ_k reads

$$P(\Lambda_k|\alpha_k) = \frac{1}{\Gamma(Md_k - \frac{\delta_{k,0}}{2})} \left(Md_k \frac{\bar{\Lambda}_k}{\Lambda_k} \right)^{Md_k - \frac{\delta_{k,0}}{2}} \times \frac{1}{\Lambda_k} \exp \left(-Md_k \frac{\bar{\Lambda}_k}{\Lambda_k} \right), \quad (21)$$

with the sufficient statistics

$$\bar{\Lambda}_k \equiv \frac{1}{M} \sum_{m=0}^{M-1} |\alpha_k^{(m)} - \delta_{k,0} \bar{\alpha}_0|^2, \quad (22a)$$

$$\bar{\alpha}_0 \equiv \frac{1}{M} \sum_{m=0}^{M-1} \alpha_0^{(m)}. \quad (22b)$$

This is the main result of this section. In the literature, the quantities $\bar{\Lambda}_k$ (sometimes up to a factor) are called the *periodogram*. Whereas Λ_k is commonly estimated with $\bar{\Lambda}_k$, our distribution $P(\Lambda_k|\alpha_k)$ contains more information. Figure 1 and its discussion in Sec. II apply to Λ_k and make this evident. One important application of our results can take place in the design of experiments: As the right bottom plot of Fig. 1 reflects, one can figure out the amount of data needed in order to reach a certain confidence level in Λ_k (that is, in the spectrum). This precise assessment of confidence would be difficult without the estimating distributions. One would be limited, at most, to the asymptotic analysis of the periodogram fluctuations to give a rough idea of the error.

The maximum-likelihood estimator for Λ_k is

$$\text{mle}(\Lambda_k) = \frac{Md_k}{Md_k + 1 - \delta_{k,0}/2} \bar{\Lambda}_k,$$

and we repeat that it tends to the periodogram value $\bar{\Lambda}_k$ only for large M and that the estimating distribution $P(\Lambda_k|\alpha_k)$ allows the calculation of the corresponding confidence interval as explained in Fig. 1.

The estimation of the mean is completely identical to that of Sec. II with the result given in Eq. (8a). We copy it here converting to the notation used in this section,

$$P(v|\hat{\alpha}) = \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{M-1}{2})\sqrt{\pi}\bar{\Lambda}_0} \left(\frac{\bar{\Lambda}_0}{\bar{\Lambda}_0 + (v - \bar{\alpha}_0)^2} \right)^{\frac{M}{2}}. \quad (23)$$

For small values of M , this Lorentzian-like function significantly differs from the Gaussian that one might conjecture under (an abuse of) the central limit theorem. The maximum-likelihood estimator of μ is

$$\text{mle}(\mu) = \frac{\text{mle}(v)}{\sqrt{n}} = \frac{\bar{\alpha}_0}{\sqrt{n}} = \frac{1}{Mn} \sum_{m=0}^{M-1} \sum_{j=0}^{n-1} A_j^{(m)}.$$

As expected, $\text{mle}(\mu)$ is the mean of the variable A taken over all batches and all times.

C. Statistical properties of noise

Not only is the sampling distribution $P(\alpha_k^{(m)}|\mu\Lambda_k)$ necessary to estimate Λ_k as in Sec. III B, but it is a key to understand

statistical properties of noise. In this section, we discuss two applications. On the one hand, we show how to generate noise with a given mean μ and spectrum Λ_k . On the other, we find the exact expression for the signal-to-noise ratio (SNR) of the periodogram. Our derivation will clarify some aspects of the SNR that can be found in the literature [1,3,11,43] and make extensions. In particular, we elaborate on its value of order 1 irrespective of the data-sample size, arguing that it is universal for any kind of noise. We offer a simple general proof that avoids relying on asymptotic limits, since these are not necessarily met in practice.

We focus first on the generation of noise with a given mean μ and spectrum Λ_k . Since $\alpha_k^{(m)} \in \mathbb{C}$, it is convenient to get the sampling distributions of its real and imaginary parts, $\Re\alpha_k^{(m)}$ and $\Im\alpha_k^{(m)}$, separately. From Eq. (18),

$$P(\Re\alpha_k^{(m)}|v_k\Lambda_k) = \sqrt{\frac{d_k}{\pi\Lambda_k}} \exp \left(-d_k \frac{(\Re\alpha_k^{(m)} - v_k)^2}{\Lambda_k} \right),$$

$$P(\Im\alpha_k^{(m)}|v_k\Lambda_k) = \sqrt{\frac{d_k}{\pi\Lambda_k}} \exp \left(-d_k \frac{(\Im\alpha_k^{(m)})^2}{\Lambda_k} \right).$$

The distribution for $\Im\alpha_k^{(m)}$ only applies to $k \neq 0$, $[n/2]; \alpha_k^{(m)} \in \mathbb{R}$ otherwise. These expressions are all we need: To generate the batch (m), generate the normally distributed and independent variables $\Re\alpha_k^{(m)}$ and $\Im\alpha_k^{(m)}$ for $k = 0, \dots, [n/2]$. Then, Fourier transform $\alpha^{(m)}$ to get $A^{(m)}$. Repeat the process to generate M batches.

This is in agreement with Ref. [11], which restricts itself to $1/f^\beta$ noise and departs from the statistical properties of white noise ($\Lambda_k \propto 1$) [44]. Our approach does not impose any condition on noise other than being stationary. In fact, the statistical properties of noise, given by the sampling distribution $P(\hat{\alpha}|\mu\hat{\Lambda})$ and common to every function Λ_k , should not be confused with the particular shape of the function Λ_k , which describes a specific kind of noise. With our sampling distribution, one can generate noise with arbitrary spectra. This discussion sets the basis for Sec. IV C, which will extend to the generation of correlated noise.

We proceed to discuss some statistical properties of the periodogram. The literature reports that the periodogram $\bar{\Lambda}_k$ for one batch, given by Eq. (22a) with $M = 1$, has a SNR ~ 1 irrespective of the size n of the data sample [11]. In other words, its value fluctuates a lot for different measurements of the same phenomenon (i.e., when repeating an experiment or measuring a new batch), displaying a relative error around 1. We offer a simple and general proof that this holds for any kind of noise. Once again, the key point is to deal with distributions instead of estimators. According to Appendix C, the distribution for the periodogram value $\bar{\Lambda}_k$ depends only on Λ_k and reads

$$P(\bar{\Lambda}_k|\Lambda_k) = \frac{1}{\Gamma(Md_k)} \left(Md_k \frac{\bar{\Lambda}_k}{\Lambda_k} \right)^{Md_k} \frac{1}{\bar{\Lambda}_k} \exp \left(-Md_k \frac{\bar{\Lambda}_k}{\Lambda_k} \right). \quad (24)$$

This distribution [46] has the moments [48]

$$\langle \bar{\Lambda}_k^n \rangle = \left(\frac{\Lambda_k}{Md_k} \right)^n \frac{\Gamma(n + Md_k)}{\Gamma(Md_k)},$$

the integral converging for $n + Md_k > 0$. We observe that while the maximum of the distribution is not at $\bar{\Lambda}_k = \Lambda_k$, the mean of the observed values of $\bar{\Lambda}_k$ is equal to Λ_k . For the signal to noise ratio, we get

$$\text{SNR} = \frac{\Lambda_k}{[\text{var}(\bar{\Lambda}_k)]^{1/2}} = \sqrt{Md_k}. \quad (25)$$

For $k \neq 0$ and taking $M = 1$, we finally recover the result $\text{SNR} = 1$. Our derivation confirms that the SNR does not depend on the batch size n , but only on the number of data batches M . The paradox that the SNR does not increase with n is discussed in Sec. VIC in the context of continuous spectra. We also point out that Eq. (25) agrees with the dependence \sqrt{M} when $M \rightarrow \infty$, expected from the law of large numbers [3].

IV. CROSS CORRELATIONS

This section extends the estimation of Sec. III to two real variables A and B that evolve in time. Such estimation includes the study of their noise, which might be correlated. The results that we present allow the estimation of the correlation strength and phase (defined below) with Bayesian statistics. We discuss the connection with common estimators used in the literature, like Pearson's r coefficient. Similarly to the previous cases, our results contain more than just suggesting a specific estimator.

The notation of the variable A is identical to that of Sec. III and trivially extends to the variable B . Let us use $X = (A, B)$ as a composite variable. The batch (m) contains the data

$$X^{(m)} \equiv (A_0^{(m)}, \dots, A_{n-1}^{(m)}, B_0^{(m)}, \dots, B_{n-1}^{(m)})^T,$$

and the whole data set \hat{X} contains M batches,

$$\hat{X} = (X^{(0)}, \dots, X^{(M-1)}).$$

Let the means of A and B be μ_A and μ_B , respectively, and write the correlation matrix as

$$\hat{\Sigma} \equiv \begin{bmatrix} \hat{\Sigma}^A & \hat{\Sigma}^{AB} \\ \hat{\Sigma}^{BA} & \hat{\Sigma}^B \end{bmatrix}.$$

Here, $\hat{\Sigma}_{jk}^A \equiv \langle A_j A_k \rangle - \mu_A^2$ and $\hat{\Sigma}_{jk}^{AB} \equiv \langle A_j B_k \rangle - \mu_A \mu_B$, with analogous expressions for $\hat{\Sigma}_{jk}^B$ and $\hat{\Sigma}_{jk}^{BA}$. As in Sec. III, our only approximations are (i) periodic boundary conditions: $A_{j+n} = A_j$ and the same for B ; and (ii) stationary noise: $\hat{\Sigma}_{jk}^C = \hat{\Sigma}_{j-k}^C$, for $C = A, B, AB, BA$. As before, it follows that the scalar functions Σ_j^A and Σ_j^B are even in j . Here, Σ_j^A , Σ_j^B are auto-correlation functions and Σ_j^{AB} , Σ_j^{BA} the so-called *cross-correlation functions*. Often, all of them are called correlation functions without the risk of confusion. We note that from now on, the matrix $\hat{\Sigma}$ refers to the variable X , and not to A or B separately.

As in Sec. III, we first give the sampling distribution of the variable X , then estimate the correlation matrix $\hat{\Sigma}$ and the means μ_A , μ_B , and finish with the procedure to generate correlated noise.

A. Sampling distribution

Since the M data batches are independent, the sampling probability given only by μ_A , μ_B , and $\hat{\Sigma}$ factors in the batch index,

$$P(\hat{X}|\mu\hat{\Sigma}) = \prod_{m=0}^{M-1} P(X^{(m)}|\mu\hat{\Sigma}).$$

As we prove in Appendix A, $P(X^{(m)}|\mu\hat{\Sigma})$ is a multivariate Gaussian,

$$P(X^{(m)}|\mu\hat{\Sigma}) = \frac{1}{\sqrt{(2\pi)^{2n}|\hat{\Sigma}|}} \times \exp\left(-\frac{1}{2}(X^{(m)} - \mu)^T \hat{\Sigma}^{-1}(X^{(m)} - \mu)\right). \quad (26)$$

The vector μ has the components

$$\mu_j = \begin{cases} \mu_A & \text{for } 0 \leq j < n, \\ \mu_B & \text{for } n \leq j < 2n. \end{cases}$$

As in Sec. III A, it is convenient to Fourier transform the variables for each batch (m). We denote by Z the conjugate variable to X ,

$$Z^{(m)} = \tilde{F} X^{(m)}, \quad \tilde{F} = \begin{bmatrix} \hat{F} & 0 \\ 0 & \hat{F} \end{bmatrix}.$$

\hat{F} is the Fourier-transform matrix given by Eq. (15), so \tilde{F} is unitary. For each batch, the coordinates $Z^{(m)}$ gather the Fourier transform of $A^{(m)}$ and $B^{(m)}$,

$$Z^{(m)} = (\alpha_0^{(m)}, \dots, \alpha_{n-1}^{(m)}, \beta_0^{(m)}, \dots, \beta_{n-1}^{(m)})^T.$$

The notation for $\alpha_k^{(m)}$ is also shared with Sec. III and trivially extends to $\beta_k^{(m)}$ for B .

In the transformed coordinates Z , the correlation matrix takes the form

$$\hat{\Lambda} \equiv \tilde{F} \hat{\Sigma} \tilde{F}^\dagger = \begin{bmatrix} \hat{\Lambda}^A & \hat{\Lambda}^{AB} \\ \hat{\Lambda}^{BA} & \hat{\Lambda}^B \end{bmatrix},$$

with

$$\begin{aligned} \hat{\Lambda}_{jk}^A &= \delta_{jk} \Lambda_k^A, & \Lambda_k^A &\equiv \sum_j \Sigma_j^A e^{-i2\pi jk/n}, \\ \hat{\Lambda}_{jk}^{AB} &= \delta_{jk} \Lambda_k^{AB}, & \Lambda_k^{AB} &\equiv \sum_j \Sigma_j^{AB} e^{-i2\pi jk/n} \end{aligned}$$

and analogous expressions for B and BA . Therefore, $\hat{\Lambda}$ is block diagonal in k , with blocks

$$\hat{\Lambda}_k = \begin{bmatrix} \Lambda_k^A & \Lambda_k^{AB} \\ \Lambda_k^{BA} & \Lambda_k^B \end{bmatrix}. \quad (27)$$

To prevent confusion, the reader should take note of the difference in notation for the $2n \times 2n$ matrix $\hat{\Lambda}$, the 2×2 matrix $\hat{\Lambda}_k$ and the scalar functions (of k) Λ_k^C for $C = A, B, AB, BA$. The functions Λ_k^C are correlation functions in Fourier space and are thus termed spectra. Λ_k^{AB} and Λ_k^{BA} are called *cross spectra* or simply spectra.

Below, we need also the Fourier transform of the vector μ , which is a vector with $2n$ components, only two of which

are nonzero. In line with the notation for Λ_k , we collect the two nonzero components into $\mathbf{v}_0 = \sqrt{n}(\mu_A, \mu_B)^T$ and put $\mathbf{v}_k = (0, 0)^T$ for $k \neq 0$.

For a convenient parametrization of $\hat{\Lambda}_k$, it is necessary to consider the following constraints on its matrix elements: (i) $\Lambda_k^C = (\Lambda_{-k}^C)^*$ for $C = A, B, AB, BA$ since the variables A and B are real; (ii) $\Lambda_k^A, \Lambda_k^B \in \mathbb{R}$ because, moreover, Σ_j^A and Σ_j^B are even in j due to the assumption of stationary auto-correlations; and (iii) $\Lambda_k^{AB} = (\Lambda_k^{BA})^*$, which follows from $\hat{\Sigma}_{jk}^{AB} = \hat{\Sigma}_{kj}^{BA}$ together with the assumption of stationary cross-correlations, namely from $\Sigma_j^{AB} = \Sigma_{-j}^{BA}$. The relations (i)–(iii) imply that the only independent variables in $\hat{\Lambda}$ are the values $\{\Lambda_k^A, \Lambda_k^B, \Lambda_k^{AB}\}$ for $k = 0, \dots, \lfloor n/2 \rfloor$. It also follows that the elements of $\hat{\Lambda}_k$ are complex in general, but $\Lambda_k^A, \Lambda_k^B, \Lambda_0^{AB}$, and $\Lambda_{n/2}^{AB}$ for even n are real.

We can now parametrize $\hat{\Lambda}_k$. Writing $\Lambda_k^{AB} = |\Lambda_k^{AB}| \exp(i\phi_k)$ giving $\Lambda_k^{BA} = |\Lambda_k^{AB}| \exp(-i\phi_k)$, each block $\hat{\Lambda}_k$ of the correlation matrix is given by the real parameters $\{\Lambda_k^A, \Lambda_k^B, |\Lambda_k^{AB}|, \phi_k\}$: the auto-correlation functions or spectra Λ_k^A, Λ_k^B , the cross-correlation modulus $|\Lambda_k^{AB}|$, and the cross-correlation phase ϕ_k of the (cross) spectrum. Alternatively, we can consider the real parameters $\{\Lambda_k^A, \Lambda_k^B, s_k, \phi_k\}$, with

$$s_k \equiv \frac{|\Lambda_k^{AB}|}{\sqrt{\Lambda_k^A \Lambda_k^B}}. \quad (28)$$

The dimensionless parameter s_k quantifies the degree of correlation between A and B , but normalized to the auto-correlations of A and B . We will refer to s_k as the *correlation strength*. It hints at a connection with *Pearson's r* coefficient or *linear correlation coefficient*, which will be defined in Eq. (37). Far from being an arbitrary choice to quantify cross-correlations, we stress that the definition of s_k has emerged in a natural parametrization of the matrix $\hat{\Lambda}_k$.

The caveat regarding notations made in Sec. III also applies here: The sampling distribution will be denoted by the equivalent forms

$$P(\hat{X}|\mu\hat{\Sigma}) = P(\hat{X}|\mathbf{v}\hat{\Lambda}) = P(\hat{Z}|\mu\hat{\Sigma}) = P(\hat{Z}|\mathbf{v}\hat{\Lambda}),$$

according to convenience. In these forms, the data lies on the left of the bar, and the means and correlation function on the right. In this sense, it should not be confusing to refer to distributions such as $P(Z_k^{(m)}|\mathbf{v}_k\hat{\Lambda}_k)$ as sampling distributions. Analogous considerations, with sides swapped, apply to the estimating distribution.

After the passage to the Fourier space, the sampling probability of Eq. (26) takes the form

$$P(\hat{Z}|\mathbf{v}\hat{\Lambda}) = \prod_{k=0}^{\lfloor n/2 \rfloor} \prod_{m=0}^{M-1} P(Z_k^{(m)}|\mathbf{v}_k\hat{\Lambda}_k). \quad (29a)$$

Here,

$$P(Z_k^{(m)}|\mathbf{v}_k\hat{\Lambda}_k) = \frac{1}{[(\pi/d_k)^2 \Lambda_k^A \Lambda_k^B (1-s_k^2)]^{d_k}} \times \exp[-d_k(\mathbf{Z}_k^{(m)} - \mathbf{v}_k)^\dagger \hat{\Lambda}_k^{-1}(\mathbf{Z}_k^{(m)} - \mathbf{v}_k)], \quad (29b)$$

with the obvious notation $\mathbf{Z}_k^{(m)} \equiv (\alpha_k^{(m)}, \beta_k^{(m)})^T$. The inverse of the block $\hat{\Lambda}_k$ reads

$$\hat{\Lambda}_k^{-1} = \frac{1}{1-s_k^2} \begin{bmatrix} \frac{1}{\Lambda_k^A} & -\frac{s_k \exp(i\phi_k)}{\sqrt{\Lambda_k^A \Lambda_k^B}} \\ -\frac{s_k \exp(-i\phi_k)}{\sqrt{\Lambda_k^A \Lambda_k^B}} & \frac{1}{\Lambda_k^B} \end{bmatrix}. \quad (29c)$$

Equation (29) is the main result of this section. As in Eq. (18), note that in Eq. (29b) the means μ_A, μ_B (inside \mathbf{v}) only appear for $k = 0$. This fact is important for choosing the priors, see Appendix B. Finally, we note that k ranges from 0 to $\lfloor n/2 \rfloor$, and not n , in Eq. (29a). The reason is that $\gamma_k^{(m)} = (\gamma_{-k}^{(m)})^*$ for $\gamma = \alpha, \beta$. That is, the only independent coordinates in $\mathbf{Z}^{(m)}$ are $\mathbf{Z}_k^{(m)}$ for $k = 0, \dots, \lfloor n/2 \rfloor$. We also recall that in the parametrization of $\hat{\Lambda}$, only the values of $k = 0, \dots, \lfloor n/2 \rfloor$ are needed.

B. Estimation of cross-correlations

In this section, we estimate the correlation matrix $\hat{\Sigma}$, or equivalently its transform $\hat{\Lambda}$, which in turn contains the spectra and cross spectra. As previously [see Eq. (19)], the factoring in k of the sampling distribution—Eq. (29a)—implies the factoring of the estimating distribution,

$$P(\mathbf{v}\hat{\Lambda}|\hat{Z}) = \prod_{k=0}^{\lfloor n/2 \rfloor} P(\mathbf{v}_k\hat{\Lambda}_k|\mathbf{Z}_k). \quad (30)$$

This means that the estimation of the correlation matrix $\hat{\Lambda}$ decouples into $\lfloor n/2 \rfloor + 1$ independent estimations, one for each $k = 0, \dots, \lfloor n/2 \rfloor$, of the set of four parameters $\{\Lambda_k^A, \Lambda_k^B, s_k, \phi_k\}$.

We start by estimating the spectrum Λ_k^A . The estimation of Λ_k^B is identical upon the exchange of A and B . Our goal is to calculate the estimating distribution $P(\Lambda_k^A|\hat{X})$ —equivalent to $P(\Lambda_k^A|\mathbf{Z}_k)$ by virtue of the k factoring, see Eq. (30). We do not expect the result to be exactly the same as Eq. (21), because now, due to possible correlations, the data about B might give extra information about A and vice versa. The estimating distribution is

$$P(\Lambda_k^A|\mathbf{Z}_k) = \int_{-\infty}^{\infty} d\mu_A \int_{-\infty}^{\infty} d\mu_B \int_0^{\infty} d\Lambda_k^B \int_0^1 ds_k \times \int_0^{2\pi} d\phi_k P(\Lambda_k^A \Lambda_k^B s_k \phi_k | \mathbf{Z}_k) \propto \int_0^{\infty} d\Lambda_k^B \int_0^1 ds_k \int_0^{2\pi} d\phi_k \overline{P(\mathbf{Z}_k|\mathbf{v}_k\hat{\Lambda}_k)} P(\hat{\Lambda}_k). \quad (31)$$

In these two steps, we applied Eq. (6) and the Bayes' rule. Note also the obvious replacement $\Lambda_k^A \Lambda_k^B s_k \phi_k \rightarrow \hat{\Lambda}_k$ inside the probability sign P , since both contain the same information. We have also introduced the notation

$$\overline{f(\mathbf{v})} = \int_{-\infty}^{\infty} d\mu_A \int_{-\infty}^{\infty} d\mu_B f(\mathbf{v}) P(\mu_A) P(\mu_B), \quad (32)$$

which makes it easier to provide general formulas covering both $k = 0$ and $k \neq 0$ cases. The sampling distribution $P(\mathbf{Z}_k|\mathbf{v}_k\hat{\Lambda}_k)$ entering Eq. (31) is given by Eq. (29). As for the

priors, according to Appendix B,

$$P(\mu_A, \mu_B, \hat{\Lambda}) = \prod_{k=0}^{\lfloor n/2 \rfloor} P(\mathbf{v}_k, \hat{\Lambda}_k), \quad (33)$$

with

$$P(\mathbf{v}_k, \hat{\Lambda}_k) = P(v_k^A \Lambda_k^A) P(v_k^B \Lambda_k^B) P(s_k) P(\phi_k) \quad (34)$$

for all k and with the notation $\mathbf{v}_k \equiv (v_k^A, v_k^B)$. In turn,

$$P(v_k^A \Lambda_k^A) = P(v_k^A) P(\Lambda_k^A) = \begin{cases} \text{const} \times 1/\Lambda_k^A & \text{for } k=0, \\ \delta(v_k^A) \times 1/\Lambda_k^A & \text{otherwise,} \end{cases} \quad (35)$$

with analogous expressions for the variable B , and

$$P(s_k) = 1 \quad \forall k, \\ P(\phi_k) = \begin{cases} [\delta(\phi_0) + \delta(\phi_0 - \pi)]/2 & \text{for } k=0, n/2, \\ 1/(2\pi) & \text{otherwise.} \end{cases}$$

Inserting all this into Eq. (31) and performing the integrations in Λ_k^B and ϕ_k , we get

$$P(\Lambda_k^A | \mathbf{Z}_k) \propto \frac{1}{(\Lambda_k^A)^{(M-\delta_{k,0})d_k+1}} \times \int_0^1 ds_k \exp\left(\frac{-Md_k}{1-s_k^2} \frac{\bar{\Lambda}_k^A}{\Lambda_k^A}\right) \times F\left(\begin{matrix} (M-\delta_{k,0})d_k \\ d_k \end{matrix} \middle| Md_k \frac{\bar{s}_k^2 s_k^2}{1-s_k^2} \frac{\bar{\Lambda}_k^A}{\Lambda_k^A}\right). \quad (36)$$

Here, F is the hypergeometric function [49]. We could not find a primitive for the integral in s_k and leave it for numerical evaluation. The sufficient statistics in Eq. (36) are the periodograms $\bar{\Lambda}_k^A$ and $\bar{\Lambda}_k^B$, given by applying Eq. (22a) to A and B , respectively. Further,

$$\bar{s}_k = \frac{|\bar{\Lambda}_k^{AB}|}{\sqrt{\bar{\Lambda}_k^A \bar{\Lambda}_k^B}}, \quad (37)$$

where

$$\bar{\Lambda}_k^{AB} \equiv \frac{1}{M} \sum_{m=0}^{M-1} (\alpha_k^{(m)} - \delta_{k,0} \bar{\alpha}_0) (\beta_k^{(m)} - \delta_{k,0} \bar{\beta}_0)^*. \quad (38)$$

$\bar{\alpha}_0$ is defined in Eq. (22b) and $\bar{\beta}_0$ analogously. The sufficient statistic \bar{s}_k is the Pearson's r coefficient, a parameter commonly used in the literature to analyze cross-correlations. We have arrived at it as the natural parameter of the estimating distribution.

As follows from Eq. (36), the data about B enter the estimation of the auto-correlations Λ_k^A , through \bar{s}_k and the integration in s_k . On the other hand, the noise of B is integrated out in the estimation, see Eq. (31). In contrast, applying Eq. (21) would implicitly assign all the noise observed in the data on A directly to Λ_k^A . It is reassuring to note that Eq. (36) reduces to Eq. (21) in the absence of correlations, that is in the limit of $\bar{s}_k \rightarrow 0$ and $M \rightarrow \infty$. Figure 2 visualizes the difference between the two formulas. One can see that Eq. (36) corrects Eq. (21) appreciably when M is small, while the difference

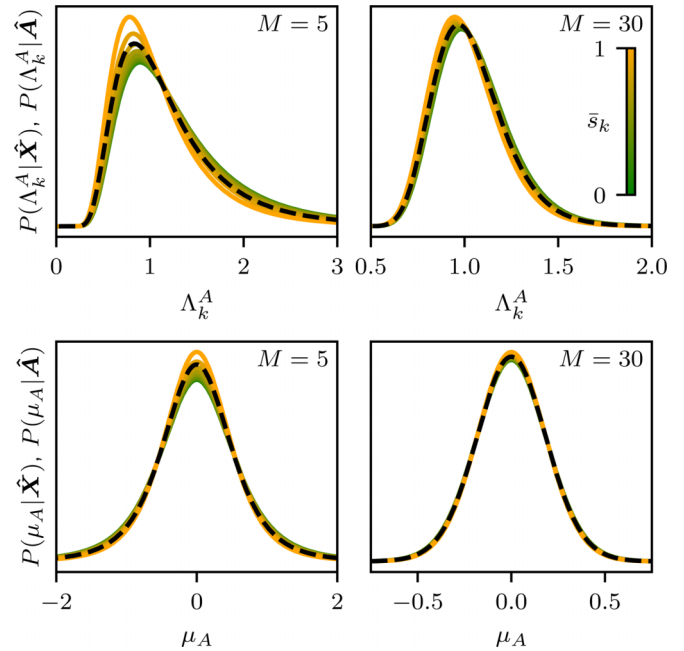


FIG. 2. Comparison of estimating distributions of the variable A when considering the partial data \hat{A} (dashed lines) vs the whole data \hat{X} (colored-solid lines). The different colors indicate the values of \bar{s}_k , ranging from 0 to 0.9 in steps of 0.1. The batch size M appears on the top right corners, $\bar{\Lambda}_k^A = 1$ and $\bar{\alpha}_0 = 0$. Top: Estimating distributions of Λ_k^A , namely Eq. (21) versus Eq. (36). Bottom: Estimating distributions of μ_A , namely Eq. (23) vs Eq. (46). The deviations from the dashed lines are significant for low values of M and \bar{s}_k close to 0 or 1.

diminishes for larger M . Nevertheless, the difference proves the need to consider *all* the available data (\hat{X}) to perform an estimation, even when it may seem that only a part of it (\hat{A}) is relevant. After these remarks, our study of auto-correlations is complete.

So far, we have estimated the diagonal elements of $\hat{\Lambda}_k$, namely the auto-correlations of the variables A and B . Now, we focus on estimating the remaining parameters, s_k and ϕ_k , which account for cross-correlations. To do so, we calculate the estimating distribution $P(s_k \phi_k | \hat{X})$ —equivalently $P(s_k \phi_k | \mathbf{Z}_k)$ —once again by virtue of the k factoring in Eq. (30). Performing the same manipulations used to get Eq. (31), we can write

$$P(s_k \phi_k | \mathbf{Z}_k) = \int_0^\infty d\Lambda_k^A \int_0^\infty d\Lambda_k^B P(\Lambda_k^A \Lambda_k^B s_k \phi_k | \mathbf{Z}_k) \propto \int_0^\infty d\Lambda_k^A \int_0^\infty d\Lambda_k^B P(\mathbf{Z}_k | \mathbf{v}_k \hat{\Lambda}_k) P(\hat{\Lambda}_k). \quad (39)$$

The sampling distribution is given by $P(\mathbf{Z}_k | \mathbf{v}_k \hat{\Lambda}_k) = \prod_m P(\mathbf{Z}_k^{(m)} | \mathbf{v}_k \hat{\Lambda}_k)$ and Eq. (29b) and the prior in Eq. (35). Performing the integrations one gets

$$P(s_k \phi_k | \mathbf{Z}_k) \propto P(\phi_k) (1 - s_k^2)^m (1 - q_{s_k \phi_k})^{\frac{1}{2}-2m} \times F\left(\frac{\frac{1}{2}, \frac{1}{2}}{2m+\frac{1}{2}} \middle| \frac{1+q_{s_k \phi_k}}{2}\right) \quad (40)$$

Here,

$$q_{s_k \phi_k} \equiv s_k \bar{s}_k \cos(\phi_k - \bar{\phi}_k), \quad (41)$$

where the angle $\bar{\phi}_k$ is the argument of the complex number $\bar{\Lambda}_k^{AB}$. We have also introduced

$$m = (M - \delta_{k,0})d_k, \quad (42)$$

to ease the formulas in this section. In sum, the only sufficient statistics relevant to inferences about cross-correlations are \bar{s}_k and $\bar{\phi}_k$ [50].

The estimating distribution for the correlation strength, $P(s_k | \mathbf{Z}_k)$, can be obtained by integrating out ϕ_k in Eq. (40), what yields

$$P(s_k | \mathbf{Z}_k) \propto (1 - s_k^2)^m \times F\left(\begin{matrix} m, m \\ d_k \end{matrix} \middle| s_k^2 \bar{s}_k^2\right). \quad (43)$$

We note in passing that for $d_k = 1$, so that m is a positive integer, this hypergeometric function can be rewritten in terms of the Legendre polynomial L_{m-1} of order $m - 1$,

$$P(s_k | \mathbf{Z}_k) \propto \left(\frac{1 - s_k^2}{1 - \bar{s}_k^2 s_k^2}\right)^m L_{m-1}\left(\frac{1 + \bar{s}_k^2 s_k^2}{1 - \bar{s}_k^2 s_k^2}\right).$$

The advantage of this form is that for small m one gets the result in the form of a simple polynomial in the two rational expressions appearing in the previous equation.

Integrating out s_k instead gives the estimating distribution for the phase,

$$P(\phi_k | \mathbf{Z}_k) \propto P(\phi_k) \left[F\left(\begin{matrix} m, m \\ m + \frac{3}{2} \end{matrix} \middle| q_{\phi_k}^2\right) + \frac{2q_{\phi_k}}{\sqrt{\pi}} \frac{\Gamma(m + \frac{3}{2})}{\Gamma(m + 2)} \right. \\ \left. \times \left(\frac{\Gamma(m + \frac{1}{2})}{\Gamma(m)} \right)^2 F\left(\begin{matrix} m + \frac{1}{2}, m + \frac{1}{2}, 1 \\ m + 2, \frac{3}{2} \end{matrix} \middle| q_{\phi_k}^2\right) \right], \quad (44)$$

with

$$q_{\phi_k} \equiv \bar{s}_k \cos(\phi_k - \bar{\phi}_k).$$

With Eqs. (36), (40), (43), and (44), the estimation of the correlation matrix $\hat{\Lambda}$ is complete [51]. Since its diagonal elements (auto-correlations) were already discussed, now we focus on the off-diagonal ones (cross-correlations). We proceed to describe the estimating distributions of s_k and ϕ_k in terms of M and of the sufficient statistics \bar{s}_k , $\bar{\phi}_k$. As expected from the dependence on $q_{s_k \phi_k}$ —defined in Eq. (41)—, $P(s_k \phi_k | \mathbf{Z}_k)$ is radially symmetric: Its shape does not depend on $\bar{\phi}_k$ but only on \bar{s}_k and M . The angle $\bar{\phi}_k$ is just a polar shift, see Fig. 3(a). The dependence on \bar{s}_k is analyzed with $P(s_k | \hat{\mathbf{X}})$ in Fig. 3(b) and with the contour lines of $P(s_k \phi_k | \mathbf{Z}_k)$ in Fig. 3(c). We point out that the estimating distribution is “conservative” in the assessment of the correlation strength. We mean that, (i) as Fig. 3(b) shows, for low or moderate values of \bar{s}_k —say, in the range $0 < \bar{s}_k \lesssim 0.5$ — $P(s_k | \mathbf{Z}_k)$ is quite spread along the s_k axis (the radial axis); and (ii) $P(s_k \phi_k | \mathbf{Z}_k)$ is skewed towards low values of s_k , see Fig. 3(c). As for the dependence on M for a fixed \bar{s}_k , Fig. 3(c) shows that greater values reduce the spread of $P(s_k \phi_k | \mathbf{Z}_k)$. Figure 4 displays the transition upon changing M for the estimating distributions $P(s_k | \mathbf{Z})$ and $P(\phi_k | \mathbf{Z})$.

The case $M = 1$ does not appear in Fig. 4 but it is worth of discussion. For any k , Eq. (43) reduces to a

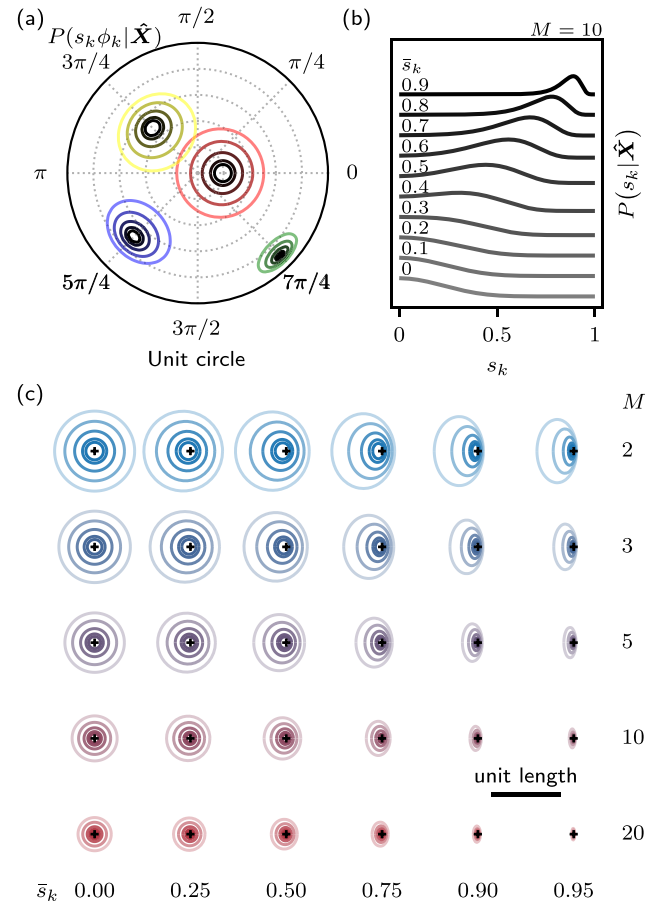


FIG. 3. Distributions of cross-correlations as a function of \bar{s}_k , $\bar{\phi}_k$, and M . (a) Contours of $P(s_k \phi_k | \hat{\mathbf{X}})$ —Eq. (40)—enclosing the 75, 50, 25, 10, and 5%-probability regions. The plot is in the complex plane, with s_k as the radial coordinate and ϕ_k as the polar angle. The unit length is given by the radius of the circle. For red, yellow, blue, and green, $\bar{s}_k = \{0.2, 0.5, 0.7, 0.9\}$ and $\bar{\phi}_k = \{0, 3\pi/4, 5\pi/4, 7\pi/4\}$, respectively. In all cases, $M = 10$. (b) Dependence of $P(s_k | \hat{\mathbf{X}})$ —Eq. (43)—on \bar{s}_k for $M = 10$. The curves have been shifted and scaled in the vertical axis. (c) Contours like in (a) as a function of \bar{s}_k and M . The complex-plane axes have been omitted (but note the given length scale) and the plots arranged in a matrix for the sake of visibility and ease of comparison. We take $\bar{\phi}_k = 0$, but recall that the value of $\bar{\phi}_k$ is irrelevant for the shape of $P(s_k \phi_k | \hat{\mathbf{X}})$ up to polar rotations. The crosses mark the maximum-likelihood estimators.

uniform distribution, $P(s_k | \mathbf{Z}_k) \propto 1$. It follows from identities [52] ${}_2F_1(1, 1, 1, x) \propto (1 - x)^{-1}$, ${}_2F_1(1/2, 1/2, 1/2, x) \propto (1 - x)^{-1/2}$, and ${}_2F_1(0, 0, 1/2, x) = 1$, for $m = 1$, $m = 1/2$, and $m = 0$, respectively. Consequently, a single batch of data does not allow one to infer anything about the degree of correlation between the variables A and B . This can be explained as follows: The variables α and β (the transforms of A and B) are decoupled in k , and for each k , we only have the values $\{\alpha_k^{(0)}, \beta_k^{(0)}\}$. With only two values, it is impossible to tell whether α and β are correlated, anticorrelated or not correlated at all. Hence the posterior for s_k is equal to its prior. Section VID contains more on this point.

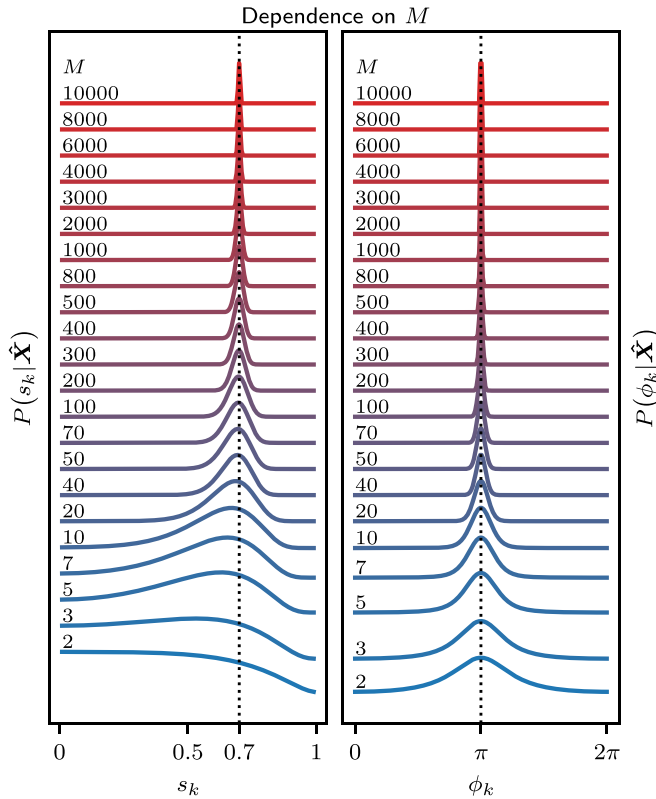


FIG. 4. Distributions of cross-correlations as a function of M for $\bar{s}_k = 0.7$ and $\bar{\phi}_k = \pi$. The curves are shifted and scaled in the vertical axis for visibility.

In parallel with the previous sections, we discuss next the importance and possible applications of our results. To begin with, one could study cross-correlations with coefficients like Pearson's r , which is equal to \bar{s}_k —cf. Eq. (37)—, and with the phase $\bar{\phi}_k$ of $\bar{\Lambda}_k^{AB}$. But we stress once again that it is more informative to make Bayesian inferences, that is, describing s_k and ϕ_k with $P(s_k \phi_k | \hat{\mathbf{Z}})$, or their marginals $P(s_k | \hat{\mathbf{Z}})$ and $P(\phi_k | \hat{\mathbf{Z}})$. Figures 3(c) and 4 illustrate this point: Even for the same values of the sufficient statistics \bar{s}_k and $\bar{\phi}_k$, the estimating distributions convey a different degree of confidence depending on the amount of available data. Not only the confidence level, but also other features of the estimating distribution (like the skewness that we have pointed out) are possibly relevant pieces of information that our estimation retains. As we discussed with Fig. 1, the explicit dependence of the estimating distributions on the number of batches M can help to design experiments aiming at a certain confidence level. In turn, this enables to detect low correlations or correlations contaminated by strong noise: When a variable (like s_k) is close to 0, one can only resolve it from its fluctuations by knowing how its error scales with the amount of data.

Moreover, there is an important conceptual difference between choosing a coefficient like Pearson's r (or, for example, Kendall's τ) [1] and our approach. Sometimes, these coefficients are not defined from the parametrization of the correlation matrix $\hat{\Lambda}_k$, but instead from other arguments. For example, Pearson's r is introduced in the context of linear regressions, hence the name “linear correlation coefficient” [1].

Consequently, one might wonder whether using other parameters would be better to analyze cross-correlations. However, this concern does not apply to our estimation: The sufficient statistic \bar{s}_k appeared as a native parametrization of the correlation matrix $\hat{\Lambda}_k$ in Sec. IV A. Actually, with our set of parameters $\{\Lambda_k^A, \Lambda_k^B, s_k, \phi_k\}$ for $k = 0, \dots, \lfloor n/2 \rfloor$, we are sure not to lose any information about the correlation matrix $\hat{\Lambda}$.

The estimation of cross-correlations is now complete. We conclude this section by discussing the estimation of μ_A . It holds identically for μ_B . Similarly to what we pointed out about Λ_k^A , the result for the estimating distribution $P(\mu_A | \hat{\mathbf{X}})$ is not expected to be exactly the same as $P(\mu_A | \hat{\mathbf{A}})$ —Eq. (23). The estimating distribution $P(\mu_A | \hat{\mathbf{X}})$ —equivalent to $P(\mu_A | \mathbf{Z}_0)$ due to the k factoring in Eq. (30)—reads

$$P(\mu_A | \mathbf{Z}_0) = \int_0^\infty d\Lambda_0^A \int_0^\infty d\Lambda_0^B \sum_{\phi_0 \in \{0, \pi\}} \int_0^1 ds_0 \times \int_{-\infty}^\infty d\mu_B P(\mu_A \mu_B \Lambda_0^A \Lambda_0^B s_0 \phi_0 | \mathbf{Z}_0) \propto \int_0^\infty d\Lambda_0^A \int_0^\infty d\Lambda_0^B \sum_{\phi_0 \in \{0, \pi\}} \times \int_0^1 ds_0 \int_{-\infty}^\infty d\mu_B P(\mathbf{Z}_0 | \mathbf{v} \hat{\Lambda}_0) P(\mathbf{v} \hat{\Lambda}_0). \quad (45)$$

Once more, we have applied Eq. (6) and Bayes' rule to get this expression. As for the presence of the means \mathbf{v} in this formula, recall our previous remark that \mathbf{v} only appears in the sampling distribution for $k = 0$. The sampling distribution $P(\mathbf{Z}_0 | \mathbf{v} \hat{\Lambda}_0)$ is given by Eq. (29a), and the prior $P(\mathbf{v} \hat{\Lambda}_0)$ by Eq. (34). After these substitutions, we can integrate analytically in all variables but s_0 , which remains for numerical evaluation,

$$P(\mu_A | \mathbf{Z}_0) \propto \int_0^1 ds_0 \left(\sqrt{\frac{1}{1-s_0^2} + \frac{(\bar{\alpha}_0 - v_A)^2}{\bar{\Lambda}_0^A}} \right)^{-1} \times \sum_{\kappa=\pm 1} \left(\frac{\kappa s_0 \bar{s}_0}{\sqrt{1-s_0^2}} + \sqrt{\frac{1}{1-s_0^2} + \frac{(\bar{\alpha}_0 - v_A)^2}{\bar{\Lambda}_0^A}} \right)^{1-M} \quad (46)$$

There is a clear resemblance to the estimating distribution of Sec. III: As a check, Eq. (46) reduces to Eq. (23) in the absence of cross-correlations, namely using the prior $P(s_0) = \delta(s_0 - 0)$. The comparison between both results appears in Fig. 2. An identical discussion to that comparing Eqs. (21) and (36) applies also here.

C. Generation of noise

While the method to generate noise given in Sec. III C is known [11], we are not aware of comparable proposals for *correlated noise*. In this section, we describe how to generate noise given means μ_A, μ_B and spectra Λ_k^A, Λ_k^B , and Λ_k^{AB} (or equivalently, $\Lambda_k^A, \Lambda_k^B, s_k$, and ϕ_k). In analogy to Sec. III C, in our method the two noisy signals are generated solely from this information and do not rely on further choices, such as on admixing or filtering other signals, distinguishing it from existing methods to generate correlated noise.

We proceed analogously to Sec. III C, using now the sampling distribution given by Eq. (29b). However, it is necessary to transform the variables $\alpha_k^{(m)}$ and $\beta_k^{(m)}$ —inside $\mathbf{Z}_k^{(m)}$ —into two independently distributed scalars. To do so, we perform the transformation $\mathbf{W}_k^{(m)} = \hat{T}_k(\mathbf{Z}_k^{(m)} - \mathbf{v}_k)$ that diagonalizes the matrix inside the sampling distribution,

$$\hat{T}_k \hat{\Lambda}_k^{-1} \hat{T}_k^{-1} = \begin{bmatrix} H_k & 0 \\ 0 & J_k \end{bmatrix}. \quad (47)$$

\hat{T}_k exists and $H_k, J_k \in \mathbb{R}$ because $\hat{\Lambda}_k^{-1}$ is Hermitian, see Eq. (29c). In the coordinates $\mathbf{W}_k^{(m)} = (h_k^{(m)}, j_k^{(m)})$, the sampling distribution reads

$$P(h_k^{(m)} | \mathbf{v}_k \hat{\Lambda}_k) = \left(\frac{d_k}{\pi H_k} \right)^{d_k} \exp \left(-d_k \frac{|h_k^{(m)}|^2}{H_k} \right).$$

The formula applies for $j_k^{(m)}$ upon the substitution $H_k \rightarrow J_k$. Splitting to the real and imaginary parts,

$$P(\Re h_k^{(m)} | \mathbf{v}_k H_k) = \sqrt{\frac{d_k}{\pi H_k}} \exp \left(-d_k \frac{(\Re h_k^{(m)})^2}{H_k} \right), \quad (48)$$

the same for $\Im h_k^{(m)}$, and also for $\Re j_k^{(m)}$ and $\Im j_k^{(m)}$ after the change $H_k \rightarrow J_k$. The distribution for the imaginary parts only applies to $k \neq 0, n/2$, because $h_k^{(m)}$ and $j_k^{(m)}$ are real otherwise.

These distributions are all we need to generate two variables A and B with given means μ_A, μ_B and spectra parametrized by $\Lambda_k^A, \Lambda_k^B, s_k$, and ϕ_k . The procedure is as follows. To generate a batch (m), (i) for each k , diagonalize the matrix $\hat{\Lambda}_k$, finding the coordinate-change matrix \hat{T}_k and the eigenvalues H_k and J_k ; (ii) generate the independent and normally distributed variables $\Re h_k^{(m)}$ and $\Im h_k^{(m)}$ with their sampling distributions—Eq. (48) and similar—, and do the same for $j_k^{(m)}$; and finally, (iii) perform the transformation $\mathbf{W}_k^{(m)} \rightarrow \mathbf{Z}_k^{(m)} - \mathbf{v}_k$ with \hat{T}_k^{-1} , and $\mathbf{Z}_k^{(m)} \rightarrow \hat{\mathbf{A}}^{(m)}, \hat{\mathbf{B}}^{(m)}$ with \tilde{F}^{-1} (the inverse Fourier transform from the frequency space back to the time space).

V. NUMERICAL EXAMPLE

In this section, we test and illustrate our analytical Bayesian estimation formulas applying them on numerically generated uncorrelated and correlated noise. Specifically, we generate $M = 10$ batches of noisy signals for three variables A, B , and C . Each batch contains $n = 1000$ data points. We work in arbitrary units for all magnitudes, including time. We choose spectra,

$$\Lambda_k^A = \text{const.}, \quad \Lambda_k^B \propto \frac{1}{\lambda^2 + k^2}, \quad \Lambda_k^C \propto \frac{1}{k}, \quad (49)$$

corresponding to white noise; to noise with an exponentially decaying correlation function, $\langle x(s)x(s+t) \rangle \propto e^{-t/\lambda}$, taking $\lambda = 10$; and to $1/f$ noise, respectively. For $1/f$ noise, we take a frequency cutoff smaller than the lowest nonzero k frequency. A is generated independently from B and C with the method presented in Sec. III C. B and C are anticorrelated with a constant strength $s_k = 0.7$ and phase $\phi_k = \pi$. They are generated according to Sec. IV C. Finally, A and B have the same total spectral power (that is, the same area under the spectral-power curve), while C is hundred times weaker.

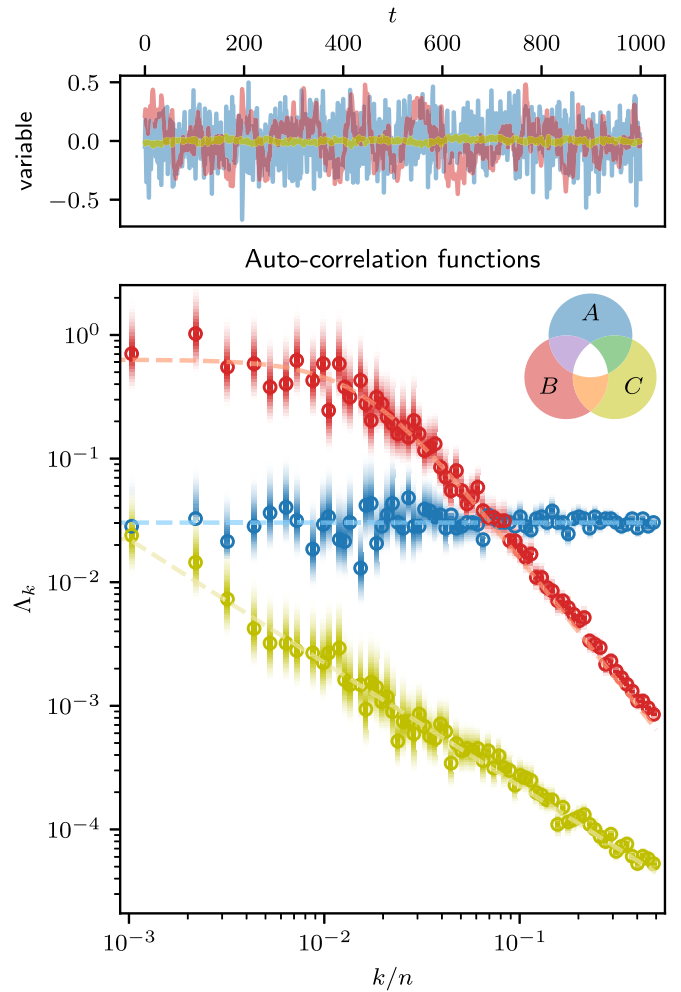


FIG. 5. Top: Time evolution of the generated variables A, B , and C for batch $m = 1$. Bottom: Estimation of the spectra Λ_k^A with Eq. (21); and of Λ_k^B, Λ_k^C with Eq. (36). The maximum-likelihood estimators appear circled, and the estimating distributions for Λ_k^A, Λ_k^B , and Λ_k^C are overlaid as a linear color gradient. The dashed lines plot the spectral functions used to generate the data, see Eqs. (49). Note that adopting the log scale for the horizontal axis makes the distributions for high k overlap. To declutter the plot and keep its information content, we locally average the points using the procedure described in Appendix F.

We chose to simulate this data according to the following purposes. First, we aim at analyzing auto-correlations for some typical noise types. Second, we intend to check that the cross-correlation between the variables B and C is correctly quantified in strength and phase, and also assess the lack of correlation between A and B , and A and C . In other words, we set a test for the positive and negative assessment of correlation. At last, we want to check that our estimation of discrete spectra is sensible for data generated from continuous dispersions, at least qualitatively. The next section discusses this point further.

Figure 5 plots the time evolution of the first batch (top panel) and the resulting estimated auto-correlations (bottom panel). The estimation of Λ_k uses Eq. (36) of Sec. IV B for B and C , and Eq. (21) of Sec. III B for A . These

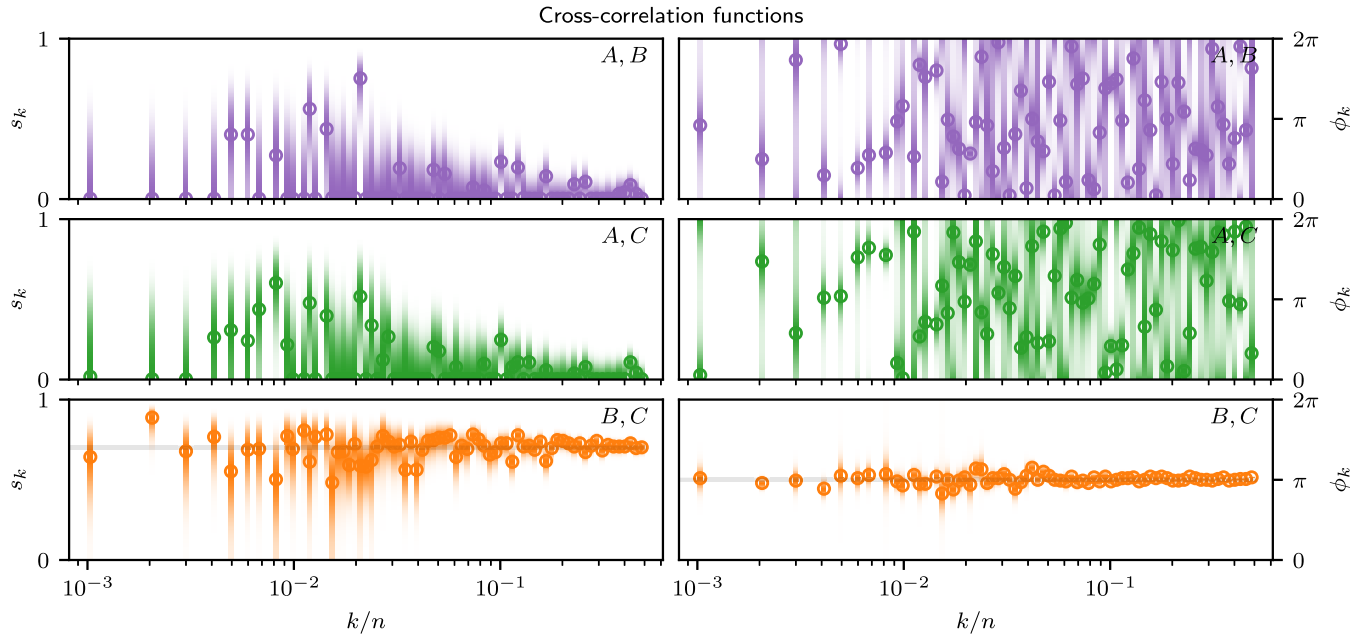


FIG. 6. Left: Estimation of the correlation strength for all possible pairs in the set $\{A, B, C\}$, with $P(s_k|\hat{X})$ as a linear color gradient. The legend appears in the upper right corner and the color code is consistent with Fig. 5. The maximum-likelihood estimators appear circled. The results reflect that only the variables B and C are correlated. The value $s_k = 0.7$, used to generate the data, is marked with a faint gray line. Right: The same for the correlation phase with the estimating distribution $P(\phi_k|\hat{X})$. The gray line marks $\phi_k = \pi$. The points were averaged according to the procedure described in Appendix F for better visibility.

estimating distributions are plotted in a color scale. The results fit well the continuous curves used to generate the data.

Figure 6 analyzes cross-correlations. The estimating distributions for s_k and ϕ_k appear plotted in a color scale for every pair in the set $\{A, B, C\}$. The degree of correlation is correctly estimated in all cases: While $P(s_k|\hat{X})$ accumulate around 0.7 for $\{B, C\}$, they peak at 0 for the uncorrelated pairs $\{A, B\}$ and $\{A, C\}$ at majority of the frequencies [53]. The results for $P(\phi_k|\hat{X})$ are consistent with those for the correlation strength: For the uncorrelated cases of $\{A, B\}$ and $\{A, C\}$, the estimating distributions peak at random values and mostly are spread over $[0, 2\pi)$. For the correlated case $\{B, C\}$, the estimating distribution is narrow and peaks at π .

Our results succeed in estimating the parameters that were used to generate the data. Some comments regarding the convergence from discrete to continuous spectra, oriented towards the application to actual experimental data, appear in the next section.

VI. DISCUSSION

We now discuss several aspects of our results concerning their application. In addition, a reader with experience in spectral estimation might find some of our results counter-intuitive. We clarify those seeming contradictions, assigning them mostly to the difference between parametric and non-parametric estimation. The technical results of our article have been already given in previous sections and they can be used without following the discussion here.

A. Nonparametric versus parametric estimation

Let us return to a single variable of time $A(t)$ observed at discretized times $t_i = i\Delta$, resulting in a discrete set $\{A_i \equiv A(t_i)\}_{i=1}^n$. This discrete set allows one to infer properties of the auto-correlation $\Sigma(\tau)$ evaluated at discrete values of the time delay $\tau_d = d\Delta$: $\{\Sigma_d \equiv \Sigma(\tau_d)\}_{d=-[n/2]+1}^{[n/2]}$. Due to the discrete sampling of A one has no access to $\Sigma(\tau)$ for other values of τ . Since the Fourier transform is a bijection, the discrete set of the Fourier coefficients $\{\Lambda_k\}$ is equivalent to $\{\Sigma_d\}$: It is a minimal-size set of parameters required to describe what can be known about $\Sigma(t)$ from the discretized values of A . However, one can introduce a different set of parameters, as a *model* to explain or predict $\{\Sigma_d\}$. Any such model can be then tested with respect to the data, in the sense that the parameters of the model can be estimated.

If the function $\Sigma(\tau)$ is defined for any real τ , what requires $A(t)$ to be defined for any real t , it is natural to consider $\Lambda(\kappa)$, the continuum of the Fourier components of $\Sigma(\tau)$, as such an extended model. Introducing a continuum of redundant variables, there would be no hope of estimating them from $\{\Sigma_d\}$. The redundancy is compensated by considering simple functions for $\Lambda(\kappa)$, describable by only a few parameters. A typical example would be $\Lambda(\kappa) = c/\kappa^\alpha$ for $\kappa \in [\kappa_{\text{low}}, \kappa_{\text{high}}]$ and zero otherwise. Here, the overall strength c , the power α and the cutoffs κ_{low} and κ_{high} would be the four (instead of a continuum of) parameters to be estimated. In general, one considers a family of continuous curves $C_{\vec{p}}(\kappa)$ defined by N parameters gathered in vector \vec{p} . Since the goal is to estimate the parameters \vec{p} , one calls the procedure *parametric* estimation. We do not introduce such functional models. We estimate the natural parametrization of $\{\Sigma_d\}$, namely the discrete set

$\{\Lambda_k\}$. To distinguish it from the other approach, we call it *nonparametric* estimation.

Now we come to the central point of this section. The motivation to introduce the continuum version of Λ_k is not so much the existence of $\Sigma(\tau)$ as a “more-true” entity than its discretized version $\{\Sigma_d\}$ [54]. Rather, the motivation is to compress the information contained in the results of nonparametric estimation [55]. The latter typically results in hundreds to millions of data points in a plot such as Fig. 5. To interpret such a plot requires that one or a few of the typical shapes are recognized in it [57]. Whether knowingly or not, anyone looking at a figure such as Fig. 5 performs intuitively at least a rough parametric estimation and draws conclusion from there. That is, the global properties of the spectral data are judged, meaning relations between different frequencies. *Our formulas do not do this; they describe each frequency k in isolation.* Once this point is appreciated, the “paradoxes” that we discuss below quickly resolve.

B. Continuous spectra

Let us now shortly comment on the parametric estimation, which is fitting the parameters of a continuous curve $\Lambda(\kappa) \equiv \mathcal{C}_{\bar{p}}(\kappa)$. Unfortunately, the issue with it is that, unlike the relation of $\{A_i\}$ to $A(t)$ and $\{\Sigma_d\}$ to $\Sigma(\tau)$, which are trivial, the relation of $\{\Lambda_k\}$ to $\Lambda(\kappa)$ is not: Λ_k is not equal to $\Lambda(\kappa_k)$ evaluated at the corresponding real frequency $\kappa_k = k/n\Delta$. Namely, the function Λ has to be additionally transformed,

$$\Lambda(\kappa) \rightarrow \mathcal{F}[\Lambda](\kappa) \rightarrow \{\mathcal{F}[\Lambda](\kappa_k)\}_{k=1}^n \equiv \{\Lambda_k\}_{k=1}^n, \quad (50)$$

The transformation \mathcal{F} reflects two well-known complications: the continuous frequencies, which are beyond the Nyquist frequency $|\kappa| > 1/2\Delta$ are aliased (folded over) into this range and the values of $\Lambda(\kappa')$ for which $n\Delta\kappa'$ is not integer are smeared throughout the discrete set $\{\Lambda_k\}$ with a weight falling off only as $1/(\kappa - \kappa')^2$. We do not give formulas for \mathcal{F} , they can be found in Ref. [1]. Equation (50) then states that $\Lambda(\kappa)$ has to be first aliased and smeared (the first arrow) and then sampled (the second arrow), to arrive at a discrete set $\{\Lambda_k\}$, the equivalent discrete spectrum.

Nevertheless, once the candidate model function $\mathcal{C}_{\bar{p}}(\kappa)$ has been decided for and the above transformation is performed, one can estimate the parameters \bar{p} from the likelihood

$$L \equiv P(\mathcal{F}[\mathcal{C}_{\bar{p}}]|\hat{\mathbf{X}}) = \prod_{k=0}^{\lfloor n/2 \rfloor} P(\Lambda_k = \mathcal{F}[\mathcal{C}_{\bar{p}}](\kappa_k)|\hat{\mathbf{X}}). \quad (51)$$

Here, $P(\Lambda_k = \mathcal{F}[\mathcal{C}_{\bar{p}}](\kappa_k)|\hat{\mathbf{X}})$ is given by Eq. (21) or Eq. (36) evaluated at $\Lambda_k = \mathcal{F}[\mathcal{C}_{\bar{p}}](\kappa_k)$. In practice, the parametric fitting is often done either simply ignoring the \mathcal{F} in Eq. (51), or moving it on the data, by changing the equation $\Lambda = \mathcal{F}[\mathcal{C}]$ into $\mathcal{F}_{\text{inv}}[\Lambda] = \mathcal{C}$, using an abstract notation without indexes. In the first approach the quality of the approximation $\mathcal{F} \approx 1$ relies on the properties of the function $\mathcal{C}_{\bar{p}}(\kappa_k)$ being estimated, and is thus unknown. In the second approach, the effective inverse \mathcal{F}_{inv} is implemented by windowing [61], filtering, and procedures of similar kind [62]. Again, since \mathcal{F} certainly does not have an exact inverse (being a map from a real axis to a finite set), the quality of the effective inverse is difficult to

predict and will depend on the function \mathcal{C} itself. Therefore, we suggest using Eq. (51) instead: there, the data are, in principle, processed independently from the parametric model, by a nonparametric estimation. The likelihood is given by evaluating the resulting estimating distributions at points given by the model function transformed into an effective discrete spectrum [63]. The advantage is that this likelihood is exact, and thus allows one to assign rigorous error bars to the estimated parameters \bar{p} .

We note that one can do parametric estimation directly, skipping the nonparametric one. Reference [4] is one example, where the model likelihood is calculated in the time domain (that is, directly from the data A_i). Nevertheless, the key to be able to do so is to have the candidate model, equivalent to $\mathcal{C}_{\bar{p}}$ in the notation here [64]. Without that, to propose a reasonable candidate $\mathcal{C}_{\bar{p}}$ one has to start with a plot such as Fig. 5 anyway. Our approach can, therefore, be viewed as splitting the parametric estimation to two steps, what allows one to delay conjecturing about $\mathcal{C}_{\bar{p}}$ till we can take a more informed decision.

Finally, analogous considerations hold for the cross-correlation, where formulas such as Eq. (51) will apply for Λ_k^A , Λ_k^B , and Λ_k^{AB} together. We point out here that the estimating distribution in Eq. (30) is actually nonseparable in the four variables. There are thus dependencies in the plausible values of the parameters. Going to marginals, which we have done in calculating, for example, the estimating distributions $P(s_k)$ and $P(\phi_k)$, means that those dependencies were lost. One could consider fitting parametric models to the more-variable estimating functions such as $P(s_k\phi_k)$ and similar, instead of the marginals. Nevertheless, we leave investigations of parametric estimation as outlined in this subsection for future works.

C. Paradox 1: Precision does not increase upon acquiring more data?

Let us again focus on a single variable with discrete spectrum Λ_k . Consider that we increase the batch size n keeping the number of batches M and the time interval Δ fixed. While the number of k points, being $\lfloor n/2 \rfloor + 1$, increases, the number of data points $\{\alpha_k^{(0)}, \dots, \alpha_k^{(M-1)}\}$ for each k does not, remaining at M . Given that our estimating distributions (and the error that they convey) depend only on M —see e.g. $P(\Lambda_k^A|\alpha_k)$ in Eq. (21)—, it looks paradoxical that the acquisition of more data by increasing n does not have any effect on the precision of the estimation (given by M).

This paradox is resolved by considering our previous discussion of Eq. (51): While it is true that after increasing n , the precision of Λ_k^A does not increase *for given k* , the fact that there are more values of k results in a more precise parametric estimation. For example, in Fig. 5, doubling n but keeping M fixed would correspond to roughly doubling the number of red distributions—and circles—, but keeping their shape—and thus the length of the corresponding error bar—fixed. We could fit a continuous curve better than with a lower n . The same point has arisen in Sec. IIIC: It might have looked paradoxical that the signal to noise ratio given in Eq. (25) does not increase with the number of measured data, proportional

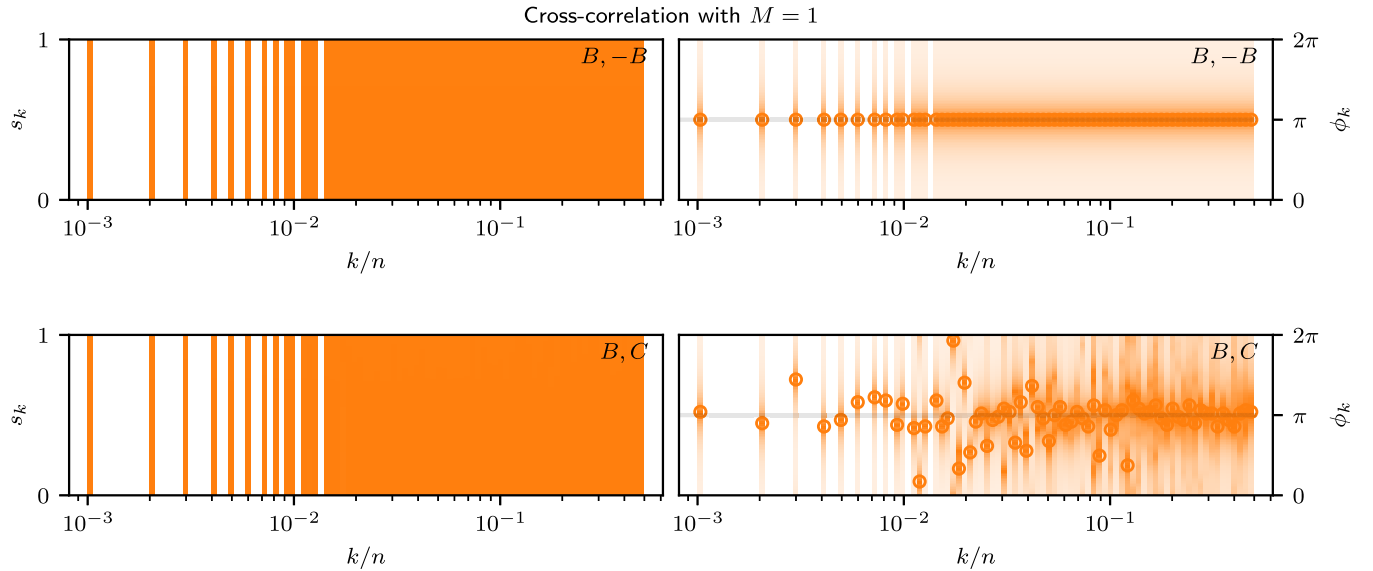


FIG. 7. The analog to Fig. 6 for fully or partially anticorrelated signals for a single batch, $M = 1$. The left panels show the estimating distributions for the correlation strength s_k , the right ones for the correlation phase ϕ_k . We generate B and C with spectra as shown in Fig. 5, with $s_k = 0.7$ and $\phi_k = \pi$ for all k . In the first row we calculate correlation of B with its inverse $-B$, while in the second row the correlation of B and C . To declutter the plot the horizontal axis was divided into 100 bins equally spaced in the logarithmic scale and all distributions falling within the same bin were arithmetically averaged. To deliver the point of the figure, we intentionally do not use here a more efficient averaging described in Appendix F.

to n . Again, the increase of the number of k points delivers a higher precision of a parametric fit.

D. Paradox 2: A single trace can not reveal anything about cross-correlations?

The paradox from the previous subsection has a variant appearing for cross-correlations. Namely, we found that the estimating distribution of the correlation strength s_k , Eq. (43), is constant, $P(s_k|\mathbf{Z}_k) = 1$ irrespective of the data, if $M = 1$. This seems to imply that irrespective of what was measured, from a single batch one can not draw any conclusions about a possible correlation of two variables A and B . One can easily find an example where the intuition points against such a statement: For example, imagine that a million of pairs B_i, C_i were measured in a single batch, and while wildly fluctuating individually, they happen to be opposite in all instances $B_i = -C_i$. One can obviously judge the perfect anticorrelation of B and C from such data.

The explanation is, again, in the global properties of the estimating distributions, that is, their relations for different frequencies k . While it is true that looking at a single k , one can conclude nothing about s_k^{BC} , the fact that all the estimated phases ϕ_k are the same in the given example makes it obvious that B and C are correlated. The parametric estimation would confirm it, for example examining a model such as $B(t) = aA(t) + cC(t)$, finding $c \approx -1, a \approx 0$. To illustrate, Fig. 7 shows the resulting estimating distributions in the first row. A similar situation arises if the correlation is imperfect, see the second row of the figure. Again, while nothing can be concluded from the estimating distributions for s_k as they are all constant, the clustering of phases ϕ_k shows that they

are not randomly distributed and thus the two signals are not uncorrelated.

E. Paradox 3: Are formulas for $M > 1$ ever relevant?

In practice one has seldom (if ever) access to statistically independent batches. In majority of cases, the whole set of $n \times M$ data points is measured on a single system, after which the data are separated into M batches. There might or might not be time delays between measurements taking data assigned to different batches m . Nevertheless, data measured in this way correspond to a single batch, possibly with some portions of the data omitted. As omitting valid data never improves any estimation, the best estimation one can hope for in this scenario is to retain all the data without omissions and use it as a single batch, not splitting it into artificial batches. This is the best (most informative) procedure for a subsequent parametric estimation: Despite our formulas taking trivial forms (for $M = 1$) at each k , for example $s_k = 1$ irrespective of the data, there is no issue in using the $M = 1$ nonparametric estimation results for the parametric estimation; we have explained it in the previous two subsections.

Nevertheless, calculating the likelihood in the parametric estimation, Eq. (51), is not the only use of the nonparametric estimation. As we pointed out in Sec. VIB, the latter is required to assist in formulating the parametric model in the first place. This task might be difficult based on seeing only the $M = 1$ plots. The artificial splitting of a single measurement into M batches can then be seen as a visualization tool [65]. Once the parametric model is formulated, one should use the true number of batches, without splitting them artificially to smaller ones, even for $M = 1$.

A natural question is the error stemming from splitting a single batch into several smaller ones, which are then treated as statistically independent (even though they are not). While we have investigated this question, we conclude that quantifying the error arising in this way in the nonparametric estimation requires additional assumptions on the correlation function. We have not been able to find assumption(s) that would be general and lead to useful formulas. We leave this question open. In any case, we expect that the “averaging” described in Appendix F makes this issue irrelevant, since it replaces the artificial splitting of batches by a more efficient procedure.

VII. CONCLUSIONS

In this article, we have applied Bayesian statistics to estimate correlation functions. We have given the estimating probability distributions to calculate (i) the variance for scalar variables; (ii) the auto-correlation functions for time-dependent variables, which includes the study of noise spectra; and (iii) the cross-correlations for possibly coupled pairs of variables, which covers the study of correlated noise. Our approach guarantees an optimal usage of the information contained in the data, makes no assumptions on the distribution of the variables, reduces to a minimum the assumptions on the priors, and avoids the arbitrary choice of estimators. Every result is expressed in terms of a distribution, from which confidence intervals can be calculated afterwards.

The relevance of our results is manifest when applied to spectral analysis. They allow one to estimate the spectrum in terms of the estimating distributions. This inference is more informative than using standard estimators such as the periodogram. It allows one to assess the certainty of the spectrum estimation (for example, by error bars) and fit continuous spectra in terms of maximum likelihood. Our analysis also covers the study of cross-correlations from a Bayesian perspective: We offer a systematic way to assess the correlation between two variables without choosing arbitrary parameters to quantify it and with distributions that assess the certainty level. We presented some numerical tests that support our calculations.

Beyond these fundamental results, our paper discusses other aspects of noise from a Bayesian perspective. As an example concerning the statistical properties of the periodogram, we derived its exact signal-to-noise ratio and proved that it is universal. We also proposed a method to numerically simulate correlated stationary noise with an arbitrary spectrum.

ACKNOWLEDGMENTS

We acknowledge the support from CREST JST (Grants No. JPMJCR15N2 and No. JPMJCR1675), Swiss National Science Foundation (SNSF), NCCR SPIN, JST PRESTO Grant No. JPMJPR21BA, MEXT Quantum Leap Flagship Program (MEXT Q-LEAP) Grant No. JPMXS0118069228, JST Moonshot R&D Grant No. JPMJMS2065, JSPS KAKENHI Grants No. 16H02204, No. 17K14078, No. 18H01819, No. 19K14640, and No. 20H00237, The Precise Measurement

Technology Promotion Foundation, Suematsu Fund, and Advanced Technology Institute Research Grants.

APPENDIX A: MULTIVARIATE GAUSSIANS

In this Appendix, we prove that for the set of variables $\vec{x} = \{x_0, \dots, x_{n-1}\}$ and given the expected values $\mu_i = \langle x_i \rangle$ and $\hat{\Omega}_{ij} = \langle x_i x_j \rangle$, the multivariate Gaussian distribution maximizes the entropy [66]. Note that giving $\hat{\Omega}_{ij}$ is equivalent to giving $\hat{\Sigma}_{ij}$ defined as $\hat{\Sigma}_{ij} = \hat{\Omega}_{ij} - \mu_i \mu_j$.

We start by the simple case $n = 1$, for which \vec{x} , $\vec{\mu}$, $\hat{\Sigma}$ and $\hat{\Omega}$ are the scalars x , μ , Σ and Ω . For notational simplicity, only in this section we denote $P(x|\mu\Sigma)$ by $P(x)$. The goal is to find the distribution $P(x)$ that maximizes the entropy

$$S = \int_{-\infty}^{\infty} dx P(x) \log P(x), \quad (\text{A1})$$

subject to the conditions

$$\begin{aligned} \int_{-\infty}^{\infty} dx P(x) &= 1, \quad \mu = \int_{-\infty}^{\infty} dx P(x)x, \\ \Omega &= \int_{-\infty}^{\infty} dx P(x)x^2. \end{aligned} \quad (\text{A2})$$

We apply the method of Lagrange to $P(x)$ with multipliers $\lambda - 1$, λ_1 and λ_{11} ,

$$\int_{-\infty}^{\infty} \delta P(x) (\log P(x) + \lambda + \lambda_1 x + \lambda_{11} x^2) = 0,$$

and thus

$$P(x) = \frac{1}{e^\lambda} \exp(-\lambda_1 x - \lambda_{11} x^2). \quad (\text{A3})$$

To determine the multipliers λ , λ_1 , and λ_{11} , we insert Eq. (A3) in Eqs. (A2). With this, we get

$$P(x) = \frac{1}{\sqrt{2\pi\Sigma}} \exp\left(-\frac{(x-\mu)^2}{2\Sigma}\right), \quad (\text{A4})$$

which ends the proof for $n = 1$.

Now, consider $n > 1$. Once again, only in this section for notational simplicity, we denote $P(\vec{x}|\vec{\mu}\hat{\Sigma})$ by $P(\vec{x})$. The goal is to find the distribution $P(\vec{x})$ that maximizes the entropy

$$S = \int_{\mathbb{R}^n} d^n x P(\vec{x}) \log P(\vec{x}), \quad (\text{A5})$$

subject to the conditions

$$\begin{aligned} \int_{\mathbb{R}^n} d^n x P(\vec{x}) &= 1, \quad \mu_i = \int_{\mathbb{R}^n} d^n x P(\vec{x}) x_i, \\ \hat{\Omega}_{ij} &= \int_{\mathbb{R}^n} d^n x P(\vec{x}) x_i x_j \quad \text{for } i, j = 0, \dots, n-1. \end{aligned} \quad (\text{A6})$$

It is convenient to perform a coordinate change $\vec{y} = U\vec{x}$, with U unitary such that $\hat{\Xi} = U\hat{\Omega}U^\dagger$ is diagonal. Because $\hat{\Omega}$ is symmetric and real, U exists and can be chosen real. Define $\vec{v} = U\vec{\mu}$ and note that $P(\vec{x}) = P(\vec{y})$, since the Jacobian of the change $\vec{x} \rightarrow \vec{y}$ is $|U| = 1$. Rewriting Eqs. (A5) and (A6) in the \vec{y} coordinates is then trivial: It only requires the changes $\mu_i \rightarrow v_i$ and $\hat{\Omega}_{ij} \rightarrow \hat{\Xi}_{ij}$.

We apply the method of Lagrange to $P(\vec{y})$ with multipliers $\lambda - 1$, λ_i , and λ_{ij} ,

$$\int_{\mathbb{R}^n} d^n y \delta P(\vec{y}) \left(\log P(\vec{y}) + \lambda + \sum_i \lambda_i y_i + \sum_{i,j} \lambda_{ij} y_i y_j \right) = 0,$$

and thus

$$P(\vec{y}) = \frac{1}{e^\lambda} \exp \left(- \sum_i \lambda_i y_i - \sum_{i,j} \lambda_{ij} y_i y_j \right).$$

To determine the multipliers, we insert this expression into Eqs. (A6) (rewritten for the y variable), $\int d^n y P(\vec{y}) = 1$ allows us to write

$$P(\vec{y}) = \frac{\exp(-\sum_i \lambda_i y_i - \sum_{i,j} \lambda_{ij} y_i y_j)}{\int d^n y' \exp(-\sum_i \lambda_i y'_i - \sum_{i,j} \lambda_{ij} y'_i y'_j)}. \quad (\text{A7})$$

$\hat{\Xi}_{ij} = \int d^n y P(\vec{y}) y_i y_j$ for $i \neq j$ implies

$$\frac{\partial}{\partial \lambda_{ij}} \int d^n y \exp \left(- \sum_i \lambda_i y_i - \sum_{i,j} \lambda_{ij} y_i y_j \right) = 0,$$

so the exponentials in Eq. (A7) are independent on λ_{ij} for $i \neq j$. Therefore,

$$P(\vec{y}) = \frac{\exp[-\sum_i (\lambda_i y_i + \lambda_{ii} y_i^2)]}{\int d^n y' \exp[-\sum_i (\lambda_i y'_i + \lambda_{ii} y'^2_i)]}. \quad (\text{A8})$$

From this equation

$$P(\vec{y}) = \prod_i P(y_i),$$

with

$$P(y_i) = \frac{\exp(-\lambda_i y_i - \lambda_{ii} y_i^2)}{\int d^n y' \exp(-\lambda_i y'_i - \lambda_{ii} y'^2_i)},$$

which in turn yields

$$S = \sum_i S_i, \quad S_i \equiv \int_{-\infty}^{\infty} dy_i P(y_i) \log P(y_i).$$

In other words, the maximization of S can be decoupled in n independent maximizations of S_i , each of them respective to $P(y_i)$ and subject to the conditions $\langle y_i \rangle = v_i$ and $\langle y_i^2 \rangle = \hat{\Xi}_{ii}$. The solution for each i is given by Eq. (A4),

$$P(y_i) = \frac{1}{\sqrt{2\pi} \Lambda_{y_i}} \exp \left(- \frac{(y_i - v_i)^2}{2\Lambda_{y_i}} \right),$$

with $\Lambda_{y_i} = \langle y_i^2 \rangle - \langle y_i \rangle^2$. Transforming back to the original \vec{x} coordinates, we get

$$P(\vec{x} | \mu \hat{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\hat{\Sigma}|}} \exp \left(- \frac{1}{2} (\vec{x} - \vec{\mu}) \hat{\Sigma}^{-1} (\vec{x} - \vec{\mu}) \right),$$

with $\hat{\Sigma}_{ij} = \langle x_i x_j \rangle - \mu_i \mu_j$. This completes the proof.

APPENDIX B: PRIORS

In this Appendix, we discuss how to calculate noninformative priors. By a noninformative prior, we mean a prior distribution that contains no information other than certain symmetry arguments. Rather than being assumptions, these

symmetry arguments, encoded in transformation groups, are necessary for a consistent description of the problem. This point will become clear throughout the calculation. We start by computing the priors of the mean and the variance of a single variable, needed for Sec. II. After that, we give the priors of the mean and the correlation matrix that appear in Secs. III and IV.

1. Single variable

Let us address the case of a single scalar variable x . We aim at estimating the prior probability distribution $P(v, \Lambda)$ of its mean v and variance Λ . The only condition or “symmetry” that we impose on $P(v, \Lambda)$ is that it should not depend on the arbitrary choice of the reference frame and units to measure x . In other words, $P(v, \Lambda)$ must be invariant under the “frame” transformations $x \rightarrow x'$ with

$$x' = a(x - \langle x \rangle) + v + b. \quad (\text{B1})$$

Here, b shifts the origin of the x reference frame, whereas the scaling factor a changes the units in which we measure the displacements of x from its mean. Under this transformation, the mean and variance change to

$$v' = v + b, \quad \Lambda' = a^2 \Lambda.$$

The invariance of the prior is expressed as

$$P(v, \Lambda) dv d\Lambda = P(v', \Lambda') dv' d\Lambda',$$

which implies

$$P(v, \Lambda) = a^2 P(v + b, a^2 \Lambda) \quad (\text{B2})$$

for all a, b . Assuming that $P(v, \Lambda)$ is differentiable in the interior of its domain ($v \in \mathbb{R}$, $\Lambda \geq 0$), Eq. (B2) is equivalent to the system of differential equations

$$\begin{aligned} \partial_v P(v, \Lambda) &= 0, \\ 2\Lambda \partial_\Lambda P(v, \Lambda) + 2P(v, \Lambda) &= 0. \end{aligned} \quad (\text{B3})$$

This equivalence can be proved by differentiating Eq. (B2) with respect to a and b and then setting the values $a = 1$, $b = 0$. The only solution reads

$$P(v, \Lambda) = \frac{\text{const}}{\Lambda}. \quad (\text{B4})$$

The result can be alternatively written as

$$P(v, \Lambda) \propto P(v)P(\Lambda), \quad P(v) \propto 1, \quad P(\Lambda) \propto \frac{1}{\Lambda}. \quad (\text{B5})$$

We note that the scale-invariant prior that we have obtained for the variance is known as the *Jeffreys prior* [20]. We still need to repeat the calculation for a case relevant for spectral analysis, namely when it is known that the mean v of x is 0. We thus consider the transformation involving only scaling, $x' = ax$. Proceeding identically to the previous case, we again get Eq. (B4). The derivation of the priors for the one-dimensional case is complete [69].

2. Many variables, factorizable priors

Let us consider the priors for the multivariate case, that is priors of μ and $\hat{\Lambda}$ —defined by Eq. (16)—used in Sec. III. We

proceed to argue that

$$P(\mu, \hat{\Lambda}) = P(\mu, \Lambda_0) \prod_{k=1}^{\lfloor n/2 \rfloor} P(\Lambda_k), \quad (\text{B6})$$

with $P(\Lambda_k)$ given by Eq. (B5). Equation (B6) seems reasonable due to the decoupling of the variables A_j into the $\lfloor n/2 \rfloor + 1$ independent variables α_k , with $k = 0, \dots, \lfloor n/2 \rfloor$ —see the discussion after Eq. (18). This decoupling is manifest in the sampling distribution $P(\hat{\alpha}|\mu, \hat{\Lambda})$ —Eq. (17)—and suggests to consider priors that also factor in k . However, let us take up a more systematic approach to get Eq. (B6). The invariance of the prior under the frame transformation Eq. (B1) of A reads

$$P(\mu', \hat{\Lambda}') d\mu' d\Lambda'_0 \dots d\Lambda'_{\lfloor n/2 \rfloor} = P(\mu, \hat{\Lambda}) d\mu d\Lambda_0 \dots d\Lambda_{\lfloor n/2 \rfloor},$$

which implies

$$P(\mu, \hat{\Lambda}) = a^{2\lfloor n/2 \rfloor + 2} P(\mu + b, a^2 \hat{\Lambda}). \quad (\text{B7})$$

It is immediate to see that Eq. (B6) together with Eqs. (B4) fulfills this condition. However, we now demonstrate that it is not the only solution. We proceed as before in this section: We rewrite Eq. (B7) as a system of differential equations and solve it. The differential equations read

$$\begin{aligned} \partial_\mu P(\mu, \hat{\Lambda}) &= 0, \\ 2\hat{\Lambda} \cdot \partial_{\hat{\Lambda}} P(\mu, \hat{\Lambda}) + (2\lfloor n/2 \rfloor + 2)P(\mu, \hat{\Lambda}) &= 0. \end{aligned}$$

Here, we used the obvious notation $\hat{\Lambda} \cdot \partial_{\hat{\Lambda}} \equiv \sum_k \Lambda_k \partial_{\Lambda_k}$. The first equation is trivial. It implies the independence of $P(\mu, \hat{\Lambda})$ on μ . We can thus focus on the second,

$$2\hat{\Lambda} \cdot \partial_{\hat{\Lambda}} P(\hat{\Lambda}) + (2\lfloor n/2 \rfloor + 2)P(\hat{\Lambda}) = 0. \quad (\text{B8})$$

A canonical way to solve linear differential equations in manifolds is the method of separation of variables [70]. It starts by a factored ansatz like Eq. (B6), leaving for the end the proof that it is the unique solution. Here, the ansatz leads to the system of uncoupled equations

$$\Lambda_k \partial_{\Lambda_k} P(\Lambda_k) + \lambda_k P(\Lambda_k) = 0, \quad k = 0, \dots, \lfloor n/2 \rfloor, \quad (\text{B9})$$

with the constants λ_k fulfilling

$$\sum_{k=0}^{\lfloor n/2 \rfloor} \lambda_k = \lfloor n/2 \rfloor + 1. \quad (\text{B10})$$

We can solve these equations to conclude that any function of the form

$$P(\mu, \hat{\Lambda}) \propto \prod_{k=0}^{\lfloor n/2 \rfloor} \frac{1}{\Lambda_k^{\lambda_k}} \quad (\text{B11})$$

is a solution of Eq. (B7). That is, such a function is invariant under the frame transformations of A . We also impose $\lambda_k > 0$ to have $\lim_{\Lambda_k \rightarrow \infty} P(\mu, \hat{\Lambda}) = 0$.

In many classical problems that require solving a differential equation like the one in Eq. (B3), one has a set of boundary conditions (along a so-called characteristic curve) [70] that (i) narrow down the solution to unique values of λ_k and (ii) allow one to prove that the factored ansatz is actually the only solution. However, we do not have such boundary conditions for the prior $P(\mu, \hat{\Lambda})$ and thus we find

that the frame invariance does not lead to a unique factorizable prior.

We now invoke an additional invariance argument: the prior for the noise at a fixed *physical* frequency should not depend on how many data points the batch contains. The Fourier index k corresponds to the frequency $k/n\Delta$, where Δ is the time step with which the data are measured. Crucially, changing n changes the assignments of the indexes k to the physical frequencies. If the reassignment is to leave the priors the same, it must be that the constants λ_k for $k \neq 0$ actually do not depend on k . Requiring in addition that the prior for $\Lambda_{k=0}$, is described by Eq. (B4), we arrive at $\lambda_k = 1$ for all k . With this additional invariance requirement, we conclude that there is a unique factorizable prior,

$$P(\mu, \hat{\Lambda}) \propto \frac{1}{\Lambda_0} \prod_{k=1}^{\lfloor n/2 \rfloor} \frac{1}{\Lambda_k}, \quad (\text{B12})$$

where we separated $\Lambda_{k=0}$, which is a somewhat different variable than all other $\Lambda_{k \neq 0}$.

3. Many variables, nonfactorizable priors

It is not too difficult to find a nonfactorizable prior fulfilling Eq. (B8). We first note the following function family:

$$P(\hat{\Lambda}) = ((\Lambda_0^p + \Lambda_1^p + \dots + \Lambda_{\lfloor n/2 \rfloor}^p)^{1/p})^{-\lfloor n/2 \rfloor - 1}. \quad (\text{B13})$$

Here, p can be any real number except zero, though probably only $p \geq 1$, or at least $p > 0$, can ever be relevant for scenarios realistic in physics. As the expression in the outer bracket is the p -norm of the vector $\hat{\Lambda}$, this family already includes appealing choices, such as an expression that depends only on the sum of Λ_k (for $p = 1$), the Euclidean distance in the space of vectors $\hat{\Lambda}$ (for $p = 2$), or the maximum $\max_k \Lambda_k$ (for $p \rightarrow \infty$). Yet another possibility worth considering is

$$P(\hat{\Lambda}) = \frac{1}{\Lambda_0} ((\Lambda_1^p + \dots + \Lambda_{\lfloor n/2 \rfloor}^p)^{1/p})^{-\lfloor n/2 \rfloor}, \quad (\text{B14})$$

a prior factorizable in Λ_0 and the rest.

A second look on Eq. (B8) reveals that it specifies the prior $P(\hat{\Lambda})$ as a homogeneous function of variables $\hat{\Lambda}$ of degree $-\lfloor n/2 \rfloor - 1$, by the Euler theorem on homogeneous functions [71]. Equations (B11), (B12), (B13), and (B14) are instances of such homogeneous functions, and there are many more. Still, we speculate that the requirement of the symmetry of the prior, either factorizable or not, with respect to permutation of the indexes k within the set $k \neq 0$, is a relevant invariance requirement. It leads to a unique factorizable prior and reduces the number of possibilities to consider for a nonfactorizable one, suggesting Eqs. (B13) or (B14) as (perhaps the only) viable candidates.

To conclude, without imposing extra conditions beyond the frame invariance, the prior is not unique. We adopt Eq. (B12) and its analogs as the prior for the main text, following the common practice of the separation of variables technique. We leave the investigations of the consequences of nonfactorizable priors as a future research avenue.

4. Priors for cross-correlations

At last, we make some remarks about the priors needed for Sec. IV. Following the notation of the main text, beware that now $\hat{\Lambda}$ refers to Eq. (27) and not Eq. (16). An analogous discussion to that of Eq. (B6) leads to

$$P(\mu_A, \mu_B, \hat{\Lambda}) = \prod_{k=0}^{[n/2]} P(\mu_A, \mu_B, \hat{\Lambda}_k), \quad (\text{B15a})$$

with

$$P(\mu_A, \mu_B, \hat{\Lambda}_k) = P(\Lambda_k^A) P(\Lambda_k^B) P(s_k) P(\phi_k), \quad (\text{B15b})$$

for all k . $P(\Lambda_k^A)$ and $P(\Lambda_k^B)$ are given by Eq. (B5). Note that we have assumed the factoring of the prior in k and also in the independent quantities $\{\Lambda_k^A, \Lambda_k^B, s_k, \phi_k\}$ that parametrize $\hat{\Lambda}_k$ for each k . Analogously to the previous case, the factorization does not follow from the invariance requirements that we impose.

In Eq. (B15), it remains to determine the priors $P(s_k)$ and $P(\phi_k)$. The correlation strength s_k and the correlation phase ϕ_k are the modulus and the argument of $\Lambda_k^{AB}/(\Lambda_k^A \Lambda_k^B)^{1/2}$, respectively. Similarly to the quantity Λ of the one-dimensional case, discussed in the beginning of this Appendix, Λ_k^A , Λ_k^B , and Λ_k^{AB} transform by a multiplicative factor under the frame transformation. Therefore, s_k and ϕ_k remain invariant and there are no nontrivial conditions such as Eqs. (B2) that apply to $P(s_k)$ and $P(\phi_k)$. We can impose only entropy maximization on these priors. The result, of a trivial proof, is a uniform distribution,

$$P(s_k) = 1 \quad \forall k,$$

$$P(\phi_k) = \begin{cases} [\delta(\phi_0) + \delta(\phi_0 - \pi)]/2 & \text{for } k = 0, n/2, \\ 1/(2\pi) & \text{otherwise.} \end{cases}$$

The phase prior for $k = 0, n/2$ follows from Λ_0^{AB} and $\Lambda_{n/2}^{AB}$ (for even n) being real.

APPENDIX C: DERIVATION OF THE PERIDOGRAM DISTRIBUTION

In this Appendix, we derive the distribution $P(\bar{\Lambda}_k|\Lambda_k)$, necessary to understand the statistical properties of the periodogram $\bar{\Lambda}_k$. The distribution can be calculated as follows. We start by applying some probability-theory rules on $P(\bar{\Lambda}_k|\Lambda_k)$,

$$\begin{aligned} P(\bar{\Lambda}_k|\Lambda_k) &= \int_{-\infty}^{\infty} d\mu \prod_{m=0}^{M-1} d\alpha_k^{(m)} P(\bar{\Lambda}_k \alpha_k \mu | \Lambda_k) \\ &= \int_{-\infty}^{\infty} d\mu \prod_{m=0}^{M-1} d\alpha_k^{(m)} P(\mu | \Lambda_k) \\ &\quad \times P(\alpha_k | \mu \Lambda_k) P(\bar{\Lambda}_k | \alpha_k \mu \Lambda_k). \end{aligned}$$

The first equality uses Eq. (6). In the second, we apply the conditional-probability definition $P(ab) = P(a)P(b|a)$ twice. The differentials in the variables α read

$$d\alpha_k^{(m)} = \begin{cases} d\Re \alpha_k^{(m)} & \text{for } k = 0, n/2, \\ d\Re \alpha_k^{(m)} d\Im \alpha_k^{(m)} & \text{otherwise,} \end{cases}$$

because $\alpha_k^{(m)}$ is real for $k = 0, n/2$ and complex (in general) otherwise. Consequently, the number of one-dimensional integrals over α is $2Md_k$. Next, we use that $P(\mu|\Lambda_k) = P(\mu)$ and define the average

$$\overline{P(\alpha_k|\mu\Lambda_k)} = \int d\mu P(\mu) P(\alpha_k|\mu\Lambda_k). \quad (\text{C1})$$

Also, using the definition of Eq. (22a), we write $P(\bar{\Lambda}_k|\alpha_k\mu\Lambda_k)$ as a delta function. These two replacements give

$$\begin{aligned} P(\bar{\Lambda}_k|\Lambda_k) &= \int \prod_{m=0}^{M-1} d\alpha_k^{(m)} \overline{P(\alpha_k|\mu\Lambda_k)} \\ &\quad \times \delta\left(\bar{\Lambda}_k - \frac{1}{M} \sum_{m=0}^{M-1} |\alpha_k^{(m)} - \delta_{k,0}\bar{\alpha}_0|^2\right). \end{aligned}$$

We now consider $k \neq 0$ and $k = 0$ separately. For $k \neq 0$ the sampling distribution $P(\alpha_k|\mu\Lambda_k)$ is independent on μ . The integral in μ is then trivial, giving $\int d\mu P(\mu) = 1$ by the normalization of the probability distribution $P(\mu)$ and the overline can be simply omitted. We can now substitute the condition imposed by the delta function in $P(\alpha_k|\mu\Lambda_k)$, in turn given by $\prod_m P(\alpha_k^{(m)}|\mu\Lambda_k)$ and Eq. (18). This yields

$$\begin{aligned} P(\bar{\Lambda}_k|\Lambda_k) &= \exp\left(-Md_k \frac{\bar{\Lambda}_k}{\Lambda_k}\right) \left(\frac{d_k}{\pi \Lambda_k}\right)^{Md_k} \int \prod_{m=0}^{M-1} d\alpha_k^{(m)} \\ &\quad \times \delta\left(\bar{\Lambda}_k - \frac{1}{M} \sum_{m=0}^{M-1} |\alpha_k^{(m)} - \delta_{k,0}\bar{\alpha}_0|^2\right). \end{aligned}$$

For $k = 0$, a simple integration in μ is needed to evaluate Eq. (C1) and confirm that the previous equation holds also for $k = 0$, and thus for all k . Using $\delta(x) = d\Theta(x)/dx$ and taking the derivative out of the integrals, we recognize the formula for the volume of a $2Md_k$ -dimensional sphere and write

$$P(\bar{\Lambda}_k|\Lambda_k) \propto \exp\left(-Md_k \frac{\bar{\Lambda}_k}{\Lambda_k}\right) \frac{\partial}{\partial \bar{\Lambda}_k} V_{2Md_k}(\sqrt{M\bar{\Lambda}_k}),$$

where the N -dimensional volume $V_N(r)$ reads

$$V_N(r) = \frac{\pi^{\frac{N}{2}}}{\Gamma(\frac{N}{2} + 1)} r^N.$$

Substituting this expression, taking the derivative with respect to $\bar{\Lambda}_k$ and normalizing the result, we get Eq. (24).

APPENDIX D: A USEFUL IDENTITY TO INTEGRATE OUT THE MEAN

Concerning zero-frequency spectral components that have finite mean, the following formula is useful. The $k = 0$ sampling probability from Eq. (29b) is

$$\begin{aligned} P(\mathbf{Z}_0|\mathbf{v}_0\hat{\Lambda}_0) &= \frac{1}{[(2\pi)^2 \Lambda_0^A \Lambda_0^B (1 - s_0^2)]^{M/2}} \\ &\quad \times \exp\left[-\frac{1}{2} \sum_m \left(\frac{\alpha_0^{(m)} - v_A}{\beta_0^{(m)} - v_B}\right)^\dagger \right. \\ &\quad \left. \cdot \hat{\Lambda}_0^{-1} \cdot \left(\frac{\alpha_0^{(m)} - v_A}{\beta_0^{(m)} - v_B}\right)\right], \quad (\text{D1a}) \end{aligned}$$

where we have used $d_k = 1/2$, and, with $s_\phi = \exp(i\phi)$,

$$\hat{\Lambda}_0^{-1} = \begin{pmatrix} \frac{1}{\Lambda_0^A} & -\frac{s_0 s_\phi}{\sqrt{\Lambda_0^A \Lambda_0^B}} \\ -\frac{s_0 s_\phi^*}{\sqrt{\Lambda_0^A \Lambda_0^B}} & \frac{1}{\Lambda_0^B} \end{pmatrix}, \quad (\text{D1b})$$

We note that all quantities in the above equations are real. It means that s_ϕ is either $+1$ or -1 . The variable \mathbf{Z}_0 denotes together the two sets $\{\alpha_0^{(m)}\}$ and $\{\beta_0^{(m)}\}$ that contain M real numbers each. It is now an algebraic exercise to cast the expression into the following form:

$$P(\mathbf{Z}_0 | \mathbf{v}_0 \hat{\Lambda}_0) = \frac{\exp\left[-\frac{M}{2(1-s_0^2)} X\right]}{[(2\pi)^2 \Lambda_0^A \Lambda_0^B (1-s_0^2)]^{M/2}}, \quad (\text{D2a})$$

where

$$X = \frac{\bar{\Lambda}_0^A}{\Lambda_0^A} + \frac{\bar{\Lambda}_0^B}{\Lambda_0^B} - \frac{2s_0 s_\phi \bar{\Lambda}_0^{AB}}{\sqrt{\Lambda_0^A \Lambda_0^B}} + \frac{1-s_0^2}{\Lambda_0^A} (v_A - \bar{\alpha}_0)^2 + \frac{1}{\Lambda_0^B} \left(v_B - \bar{\beta}_0 - s_0 s_\phi \sqrt{\frac{\Lambda_0^B}{\Lambda_0^A}} (v_A - \bar{\alpha}_0) \right)^2. \quad (\text{D2b})$$

In the derivation, it is helpful to notice that $\bar{\Lambda}_0^A$ and $\bar{\Lambda}_0^B$ are the variances of the sets $\{\alpha_0^{(m)}\}$ and $\{\beta_0^{(m)}\}$, and $\bar{\Lambda}_0^{AB}$ is the covariance of the two sets. Particularly, the (co)variance is not changed upon a uniform shift of all input numbers. One can use Eq. (D2) to calculate the sampling distributions averaged over the mean, $P(\cdots | \mathbf{v}_0 \hat{\Lambda}_0)$. Indeed, setting $s = 0$ and ignoring the B -related variables, one can easily integrate out the mean μ_A in Eq. (D2), deriving Eq. (21) from Eq. (20) for $k = 0$. Similarly, integrating out both v_A and v_B in Eq. (D2) gives the $k = 0$ sampling distribution entering Eq. (31).

APPENDIX E: POSTERIOR FOR UNNORMALIZED CROSS-CORRELATION STRENGTH

Here we consider an alternative parametrization of the cross-correlation, using $\{\Lambda_k^A, \Lambda_k^B, |\Lambda_k^{AB}|, \phi_k\}$ instead of the set $\{\Lambda_k^A, \Lambda_k^B, s_k, \phi_k\}$ considered in the main text. These sets differ only in one variable, being $|\Lambda_k^{AB}|$ instead of s_k . Therefore, the only additional marginal posterior to calculate is $p(|\Lambda_k^{AB}|)$, the remaining three posteriors remain as given in the main text. Indeed, we have

$$p(\Lambda_k^A, \Lambda_k^B, s_k, \phi) d\Lambda_k^A d\Lambda_k^B ds_k d\phi = p(\Lambda_k^A, \Lambda_k^B, |\Lambda_k^{AB}|, \phi) \frac{ds_k}{d|\Lambda_k^{AB}|} d\Lambda_k^A d\Lambda_k^B d|\Lambda_k^{AB}| d\phi, \quad (\text{E1})$$

and one can see that the marginals of Λ_A , Λ_B , and ϕ are the same using any of the coordinates sets, since integrating out s_k is the same as integrating out $|\Lambda_k^{AB}|$. However, the marginals of s_k versus $|\Lambda_k^{AB}|$ are different and can not be obtained from each other, as these two variables define different hyperplanes in the four-dimensional space.

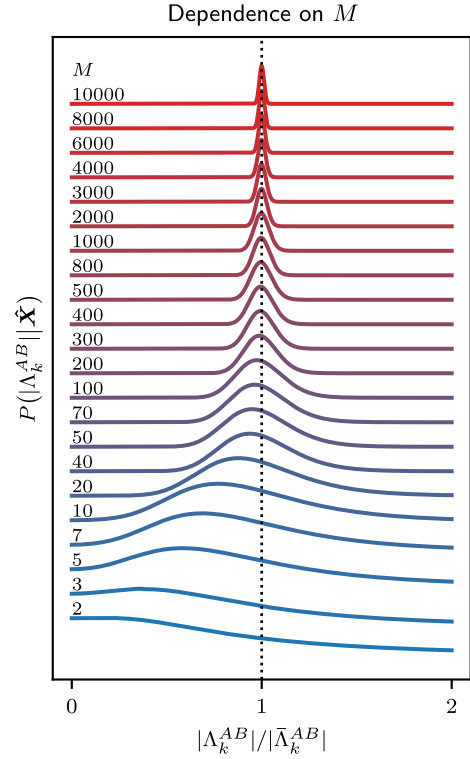


FIG. 8. Distribution of the magnitude of unnormalized cross-correlations as a function of M for $\bar{s}_k = 0.7$. The curves are shifted and scaled in the vertical axis for visibility.

For completeness, we repeat the definition of the marginal,

$$P(|\Lambda_k^{AB}| | \mathbf{Z}_k) = \int_0^\infty d\Lambda_k^A \int_0^\infty d\Lambda_k^B \int_0^{2\pi} d\phi_k \overline{P(\Lambda_k^A \Lambda_k^B s_k \phi_k | \mathbf{Z}_k)} = \int_0^\infty d\Lambda_k^A \int_0^\infty d\Lambda_k^B \int_0^{2\pi} d\phi_k \overline{P(\mathbf{Z}_k | \mathbf{v}_k \hat{\Lambda}_k) P(\hat{\Lambda}_k)}. \quad (\text{E2})$$

We could cast it as the following one-dimensional integral

$$P(|\Lambda_k^{AB}| | \mathbf{Z}_k) \propto \frac{1}{|\Lambda_k^{AB}|^{2d_k(M-\delta_{k,0})+1}} \int_0^\infty dx \frac{x^{d_k(M-\delta_{k,0})}}{x^{\frac{1}{2}}(1+x)^{\frac{3}{2}}} \times I_0\left(\frac{2d_k M x}{|\Lambda_k^{AB}|/|\bar{\Lambda}_k^{AB}|}\right) K_0\left(\frac{2d_k M \sqrt{x(x+1)}}{\bar{s}_k |\Lambda_k^{AB}|/|\bar{\Lambda}_k^{AB}|}\right), \quad (\text{E3})$$

where \bar{s}_k and $|\bar{\Lambda}_k^{AB}|$ are as defined in the main text, K_n is the modified Bessel function of the second kind, and I_n is the modified Bessel function of the first kind. Note how $|\bar{\Lambda}_k^{AB}|$ sets the scale for $|\Lambda_k^{AB}|$. In Fig. 8 we illustrate this distribution for different values of M .

APPENDIX F: AVERAGING OVERLAPPING POINTS IN LOG-LOG PLOTS

The results of the estimation, the spectral plots, are often presented using the log-log scale, such as in Fig. 5. It is

useful in revealing a possible power-law frequency dependence of the spectral densities, which is often the case. On the other hand, it comes with a disadvantage of overcrowding the figure at higher frequencies: since the density of the points on the horizontal axis is constant on the linear scale, it grows exponentially on the log scale. Due to a finite resolution of the printer, computer screen, and the human eye, one can often see just an impenetrable mess in such plots above a certain frequency. Here we propose a solution of this problem.

The idea is to replace several neighboring points on the plot by a single representative point. While this idea seems too obvious, there are different ways how one can merge several points into one. We suggest the following way. We aim to represent a (dimensionfull) frequency interval $\mathcal{F} = \langle \kappa_{\text{low}}, \kappa_{\text{high}} \rangle$, where for simplicity $\kappa_{\text{low}} > 0$. It contains a set \mathcal{K} of integer Fourier indexes k , that is, $k \in \mathcal{K} \Leftrightarrow \kappa_k \equiv k/n\Delta \in \mathcal{F}$. Let the number of integers in \mathcal{K} be K . For each of these integers, one has the associated set $\{\alpha_k^{(m)}\}$. To get the single representative, one should merge these sets into one set containing KM entries, $\{\alpha_{\text{eff}}^{(m)}\} = \bigcup_{k \in \mathcal{K}} \{\alpha_k^{(m)}\}$. This merged set is interpreted as an input to Eq. (22) corresponding to an effective batch size KM . Evaluating Eq. (22a) gives $\bar{\Lambda}_{\text{eff}} = (1/K) \sum_{k \in \mathcal{K}} \bar{\Lambda}_k$, a simple average of the sufficient statistics. One then evaluates Eq. (21) with effective parameters $\bar{\Lambda}_k \rightarrow \bar{\Lambda}_{\text{eff}}$ and $M \rightarrow KM$, to obtain the representative estimating distribution.

This procedure has the following advantages: (1) One can merge a variable number of Fourier components throughout the plot, for example, plotting nonmerged values ($K = 1$) at low frequencies, and increasing K towards high frequencies. Therefore, one does not have to sacrifice the lowest frequencies, usually of primary interest, a deficiency that results from artificially splitting the data to shorter batches [72]. (2) In the limit of a narrow frequency window (more specifically, in the limit where the spectral density is constant within the covered frequency interval), the resulting estimating distribution is *exact*. (3) Since one gets an estimating distribution for the representative point, one still has the freedom of presenting it in various ways, for example, plotting its maximum likelihood estimator or average, in a way which is consistent with the nonmerged points on the plot. (4) Most crucially, due to (2), a rigorous error bar can be assigned to each representative point.

Let us point out another paradox related to this approach. The number of points assigned to a single “representative” point grows for larger frequency. As one expects, and as we prove in Appendix H on the central limits, the estimating distributions shrink, in proportion to $1/\sqrt{K}$. It would seem that when such points are used in some subsequent estimation (say, parametric estimation of the continuous spectrum), these points with minuscule error bars will dominate the fit, making the low-frequency points irrelevant. The resolution of the paradox is in realizing that an “effective point” *represents an interval* rather than a precise frequency, irrespective of where exactly it is placed. The interval width is proportional to K . Therefore, to use an “effective point” correctly, one needs to use error bars for both its coordinates, the vertical as well as the horizontal one.

We have used this averaging in plotting Fig. 5. We have split the horizontal axis to intervals of constant length on the log-scale and merged the frequencies within each such

interval. This means the K starts at 1 on that plot, becomes larger than 1 at $k = 47$ and grows towards high frequencies up to 31 at the highest frequency.

APPENDIX G: HOW TO TREAT ERRORS (FINITE PRECISION) ON INPUT

In this article, we have considered stochastic variables, denoted as A , B , or C . Their fluctuations (or noise) is what we aim at extracting. However, apart from their “inherent” fluctuations, the values that we use as input in the above formulas are often subject to additional errors. Namely, there is no realistic case where a continuous quantity, such as A_j , would be known precisely. In line with the spirit of this article, the values A_j will be given as probability distributions rather than as exact numbers. These probability distributions are results of some measurement protocol, which is nothing else than another case of estimation. This estimation comes with its own precision, and thus nonzero width of those probability distributions on the input. We introduce the following notation:

$$A'_i = A_i + \delta A_i, \quad (\text{G1})$$

with A' the available signal, A the quantity of interest, and δA the unwanted error.

To illustrate, consider that A is the energy of a qubit. As there is no way to measure it directly, one needs to extract it from observing, for example, qubit precession, fitting it to damped harmonic oscillations. While there is no issue with the formulas in the main text, since they work whatever is the nature, character, or origin of the fluctuations in the signal, in this case one is interested in the inherent noise of the qubit energy, without the contribution from the imperfect energy-estimation. We now give a simple recipe how to subtract the latter contribution. The subtraction is possible if we can *separately estimate the spectrum of the additional error δA and this error is uncorrelated with the remaining part of the signal A* . In the following we give the necessary formulas for correcting the auto- and cross-spectra.

1. Auto-correlation case

In this scenario, one has individual estimates of the auto-correlation for the available input A' and for the error itself δA . The estimating distribution of the auto-correlation for the signal with the error removed is

$$\begin{aligned} \tilde{P}(\Lambda_k) &\propto \int_0^\infty d\Lambda'_k P'(\Lambda'_k) \int_0^\infty d\lambda_k p(\lambda_k) \delta(\Lambda_k + \lambda_k - \Lambda'_k) \\ &= \int_0^\infty d\lambda_k P'(\lambda_k + \Lambda_k) p(\lambda_k). \end{aligned} \quad (\text{G2})$$

Here, the probability distributions P' and p are the estimating distributions given in Eq. (21) or Eq. (36) for the error-polluted signal A' and the error δA , respectively. In the simple case of the error being white noise with variance σ^2 , one can use $p(\lambda_k) = \delta(\lambda_k - \sigma^2)$ in the above, and the only calculation required is to get the overall normalization factor as a one-dimensional integral.

In practice, one can not sample the error together with the signal (since that would mean one could simply subtract

the error and use precise inputs). Therefore, the assumption on their zero correlation can not be checked, and will hardly be met in practice. For this reason, we do not delve on this topic more [73] and stop at Eq. (G2) as a rough approximation for the signal stripped of the additional error. We put tilde on the resulting estimating distribution to emphasize that the result of Eq. (G2) relies on an explicit assumption (of zero correlation of the signal of interest and the error) that would not be met in practice, unlike all other estimating distributions derived in this article, which are, in a well defined sense, exact. Nevertheless, even if only approximate, we stress that Eq. (G2) produces a valid estimating distribution (a well-defined probability distribution), whatever is the relation of the signal and error: The error might be comparable or even much bigger than the typical value of signal auto-correlation as estimated from the available data.

2. Cross-correlation case

We now extend the input-error-removal procedure to cross-correlations. Equation (G1) then applies to each of two stochastic variables A and B . The cross-correlator we can access from the error-polluted input data is

$$\langle A'_i B'_j \rangle = \langle A_i B_j \rangle + \langle \delta A_i \delta B_j \rangle. \quad (\text{G3})$$

As before, this result assumes that errors are uncorrelated with signals. We now discuss two typical scenarios.

First, consider that A and B are estimated independently, either using separate experimental equipment or unrelated measurement and estimation procedures. In this case, one expects mutually independent errors $\langle \delta A_i \delta B_j \rangle = 0$, and there is nothing to correct for [74].

Second, consider that A and B are composite quantities: Not directly accessible, they are derived from other measured or estimated variables. In this situation the errors will often be correlated by common elements in that derivation. Assuming that one can somehow estimate the cross-spectrum $\langle \delta A_i \delta B_j \rangle$, it can be removed analogously as before for auto-correlations. The main difference is that cross-correlations are complex quantities, having both phase and magnitude. The subtraction needs to be done for estimating distributions describing both, that is, before any these two variables (phase or magnitude) is marginalized out. Let us consider that the removal is done for the estimating distribution expressed using the unnormalized cross-correlation strength $|\Lambda_k|$ and the phase ϕ_k . It corresponds to Eq. (E3) before integrating out the phase and reads

$$\begin{aligned} P(\Lambda_k) &\propto \frac{1}{|\Lambda_k|^{2d_k(M-\delta_{k,0})+1}} \int_0^\infty dx \frac{x^{d_k(M-\delta_{k,0})}}{x^{\frac{1}{2}}(1+x)^{\frac{3}{2}}} \\ &\times \exp\left(\frac{2d_k M \cos(\phi_k - \bar{\phi}_k)}{|\Lambda_k|/|\bar{\Lambda}_k|} x\right) \\ &\times K_0\left(\frac{2d_k M \sqrt{x(x+1)}}{\bar{s}_k |\Lambda_k|/|\bar{\Lambda}_k|}\right), \end{aligned} \quad (\text{G4})$$

where the complex number $\Lambda_k = |\Lambda_k| \exp(i\phi_k)$. We now use this distribution in the following variants: $P'(\Lambda')$ for the cross-correlation evaluated with the available—error-polluted—data, $p(\lambda)$ for the cross-correlation of the errors,

and $P(\Lambda)$ for the cross-correlation of interest. Using tilde with the same meaning as explained below Eq. (G2), the desired estimating distribution is

$$\tilde{P}(\Lambda_k) \propto \int d\lambda_k P'(\lambda_k + \Lambda_k) p(\lambda_k), \quad (\text{G5})$$

with the integration being over the complex plane. Otherwise, the equation is completely analogous to Eq. (G2). The marginal distribution for the magnitude is obtained by integrating out the phase in $\tilde{P}(\Lambda_k)$, and vice versa.

Let us finish with comments on using Eq. (G5) in practice. The formula contains two explicit one-dimensional integrals, while the distributions P' and p require another integral each (they are given by two instances of Eq. (G4), which do not have to be normalized individually as only the final distribution needs to be normalized). Counting also the integration required for the marginalized distributions, we get a procedure requiring five nested integrals, making it computationally demanding. Still, this is the most favorable scenario: The removal can be applied using alternative parametrizations, for example, the set $\{\Lambda_k^A, \Lambda_k^B, \Re \Lambda_k^{AB}, \Im \Lambda_k^{AB}\}$. However, using normalized correlation strength s_k among the parameters, the presence of auto-correlations in the denominator [see Eq. (28)] makes the procedure too complex to be practical in our experience. Therefore, we do not give formulas for that variant: the errors on input can be removed only for the cross-correlation parametrized with the unnormalized correlation strength described in Appendix E.

APPENDIX H: CENTRAL-LIMIT RESULTS

Here we consider how estimating distributions scale in the limit $M \rightarrow \infty$. Compared to exact expressions, the limiting approximations have two benefits. First, this limit corresponds to a large number of observations for which the formulas should display the “central-limit” behavior: each estimating distribution should become a Gaussian with the variance falling off as $1/M$. Deriving such variance stands for a consistency check and gives insight on how does the distribution width (the uncertainty) depend on sufficient statistics. Second benefit is practical: the exact expressions are complicated functions, often given only as integrals, which become exceedingly unwieldy in this limit. Standard numerical libraries might fail to reliably evaluate hypergeometric functions with parameters of order thousand or more. While one rarely has so many data batches available, very high values of M easily result when plotting representative values according to Appendix F. Here, the limit formula might become numerically more precise than the exact one (unless the latter is evaluated in a special way). Perhaps even more importantly, the Gaussian-like posteriors drastically simplify the evaluation of the figure of merit in a parametric fit of the spectrum: Instead of possibly involving complicated integrals, the logarithm of Eq. (51) becomes a simple sum of weighted squared differences [75]. Therefore, the central-limit expressions can be used to quickly locate the approximate position of the minimum for a parametric fit of spectra, followed by a refinement using exact formulas.

In this part, we prefer simplicity to rigor: We are interested only in the leading order results. Among other, we neglect

factors of order one compared to M , so that we drop any small integer, $\delta_{k,0}$ among them, accompanying M in sum. We skip the derivations, give results only for the most important estimating distributions, and do not specify the error arising from the approximations. Concerning the last point, it is important to understand the following: the results given below are valid in the limit $M \rightarrow \infty$ unless the sufficient statistics are singular, for example $\bar{s} = 0$ or $\bar{s} = 1$. The singular cases might not display the central-limit behavior. This fact implies that when M is fixed and the estimating distributions are considered as functions of their sufficient statistics, the limit expressions will fail once the sufficient statistics become close enough to the singular value, for example once \bar{s} becomes small enough. In other words, at what value the parameter M can be considered “large” so that the formulas below apply, depends not only on M itself but also on the values of sufficient statistics.

1. The auto-correlation

The estimating distribution in Eq. (21) is the Poisson distribution, simple enough to work with so that no approximation is needed. Nevertheless, to be consistent with the remainder of this section, we approximate it by the following form valid in the limit $M \rightarrow \infty$ [76],

$$P(\Lambda_k | \alpha_k) d\Lambda_k \equiv p(\lambda; m) d\lambda, \quad (\text{H1})$$

where we introduced the following variables:

$$p(\lambda; m) = \frac{(m\lambda)^m}{m!} \exp(-m\lambda), \quad (\text{H2a})$$

$$\lambda = \frac{\bar{\Lambda}_k}{\Lambda_k}, \quad (\text{H2b})$$

$$m = Md_k. \quad (\text{H2c})$$

The distribution $p(\lambda; m)$ in Eq. (H2a) is the probability to observe m events over time m for a Poisson process with rate λ . It peaks at rate $\lambda = 1$, corresponding to $\Lambda_k = \bar{\Lambda}_k$. At large m , the distribution is well approximated by

$$P(\Lambda_k | \alpha_k) \simeq \frac{1}{\sqrt{2\pi}\sigma_{\Lambda_k}} \exp\left(-\frac{(\Lambda_k - \bar{\Lambda}_k)^2}{2\sigma_{\Lambda_k}^2}\right), \quad \sigma_{\Lambda_k}^2 = \frac{\bar{\Lambda}_k^2}{Md_k}. \quad (\text{H3})$$

In the first panel of Fig. 9, we compare Eq. (21) to its central-limit approximation in Eq. (H3).

While it was not exact for Eq. (21), the appeal of Eq. (H2) is that it covers also the case of correlation measurement considered in Sec. IV B. Namely, Eq. (36) reduces to it in the limit $M \rightarrow \infty$,

$$P(\Lambda_k^A | \mathbf{Z}_k) d\Lambda_k^A \equiv p(\lambda; m) d\lambda; \quad (\text{H4})$$

the only change needed is to put the upper subscript A on the two quantities on the right-hand side of Eq. (H2b).

2. The normalized cross-correlation

For large M , the hypergeometric function in Eq. (40) can be set to 1, which greatly simplifies the expression. Integrating

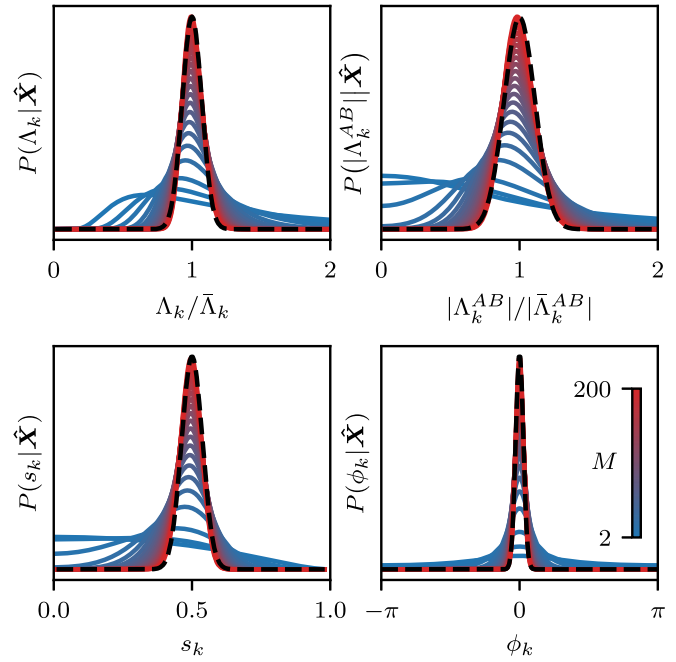


FIG. 9. Central-limit (large- M) approximations to estimating distributions. In each panel, exact estimating distributions for the variable given on the x axis are plotted for several values of m as given by the color bar. The black-dashed curves give the approximate (central-limit) estimating distributions for $m = 200$.

out the variable s_k , we got

$$P(\phi_k | \mathbf{Z}_k) \propto \exp\left(-Md_k \frac{\bar{s}_k^2}{1 - \bar{s}_k^2} \sin^2(\phi_k - \bar{\phi}_k)\right), \quad (\text{H5})$$

a form suitable when the explicit periodicity in the phase variable is beneficial. Alternatively, the expression explicitly Gaussian in the phase variable is

$$P(\phi_k | \mathbf{Z}_k) \simeq \frac{1}{\sqrt{2\pi}\sigma_{\phi_k}} \exp\left(-\frac{(\phi_k^A - \bar{\phi}_k)^2}{2\sigma_{\phi_k}^2}\right), \quad \sigma_{\phi_k}^2 = \frac{\bar{s}_k^{-2} - 1}{2Md_k} = \frac{1}{2Md_k} \frac{\bar{\Lambda}_k^A \bar{\Lambda}_k^B - |\bar{\Lambda}_k^{AB}|^2}{|\bar{\Lambda}_k^{AB}|^2}. \quad (\text{H6})$$

Similarly, integrating out the phase gives

$$P(s_k | \mathbf{Z}_k) \simeq \frac{1}{\sqrt{2\pi}\sigma_{s_k}} \exp\left(-\frac{(s_k - \bar{s}_k)^2}{2\sigma_{s_k}^2}\right), \quad \sigma_{s_k}^2 = \frac{(1 - \bar{s}_k^2)^2}{2Md_k} = \frac{1}{2Md_k} \left(\frac{\bar{\Lambda}_k^A \bar{\Lambda}_k^B - |\bar{\Lambda}_k^{AB}|^2}{\bar{\Lambda}_k^A \bar{\Lambda}_k^B}\right)^2. \quad (\text{H7})$$

The two approximations are compared to the exact distributions in the lower two panels of Fig. 9.

3. The unnormalized cross-correlation

We could cast Eq. (E3) into the following form:

$$P(|\Lambda_k^{AB}| | Z_k) \simeq \frac{1}{\sqrt{2\pi} \sigma_{\Lambda_k^{AB}}} \exp \left(-\frac{(|\Lambda_k^{AB}| - |\bar{\Lambda}_k^{AB}|)^2}{2\sigma_{\Lambda_k^{AB}}^2} \right). \quad (\text{H8a})$$

We were not able to obtain an analytic estimate for the variance $\sigma_{\Lambda_k^{AB}}$ directly from Eq. (E3). However, our numerical investigations suggested

$$\sigma_{\Lambda_k^{AB}}^2 = \frac{\bar{\Lambda}_k^A \bar{\Lambda}_k^B}{2M d_k} (1 + \bar{s}_k^2). \quad (\text{H8b})$$

As we could derive the same formula from Eq. (E3) in the limit $\bar{s}_k \rightarrow 0$, we believe that it is valid generally. The correspondence between the exact distribution and its approximate form is illustrated in the upper-right panel of Fig. 9. Among other, Eq. (H8) implies that one needs $M \sim \Lambda_k^A \Lambda_k^B / (\Lambda_k^{AB})^2$ batches to discriminate a finite value of Λ_k^{AB} from zero, that is, to detect a cross-correlation.

APPENDIX I: THE MEANING OF “NONPARAMETRIC”

We would like to comment on the meaning of “nonparametric” since it varies in the literature. With “nonparametric statistics” a subfield of current statistical sciences [77], it contains works denoted as “Bayesian”, including those aiming at spectral estimation; for example, see Ref. [78] and references therein. There is some tension in using both qualifiers together: A recent review [79] starts with (in this Appendix we

use the *Italic* font for a text cited literally) *Nonparametric Bayesian inference is an oxymoron and a misnomer.*, and later adds ... *a commonly used technical definition of nonparametric Bayesian models is probability models with infinitely many parameters (Bernardo and Smith, 1994). Equivalently, nonparametric Bayesian models are probability models on function spaces.* Another review [80] puts it differently: *Nonparametric models constitute an approach... where the sizes of models are allowed to grow with data size.... In the context of Bayesian nonparametric models, “infinite-dimensional” can therefore be interpreted as “of finite but unbounded dimension”.*

We chose the name “nonparametric” with a motivation expressed by the opening sentence of Ref. [81]: *The basic idea of nonparameteric inference is to use data to infer an unknown quantity while making as few assumptions as possible.* However, the next sentences from that reference: *Usually, this means using statistical models that are infinite-dimensional. Indeed, a better name for nonparameteric inference might be infinite-dimensional inference.* do not apply to our model.

To conclude on the naming, we stress that we are neither affected by nor interested in the issues inherent in infinite-dimensional models. Our model is finite dimensional. For example, when we derive the behavior for large M (the number of batches) in Appendix H, we perform a limit $M \rightarrow \infty$ from results where M is large but finite at each step of the limiting process. One could similarly investigate the limit of large N (the length of one data batch), though we do not do it here.

-
- [1] See Appendix I for a comment on the meaning of “nonparametric”.
 - [2] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. (Cambridge University Press, Cambridge, 2007).
 - [3] M. B. Priestley, *Spectral Analysis and Time Series: Probability and Mathematical Statistics* (Academic Press, New York, 1981).
 - [4] G. L. Bretthorst, *Bayesian Spectrum Analysis and Parameter Estimation*, Lecture notes in Statistics No. 48 (Springer-Verlag, New York, 1988).
 - [5] W. Schottky, Über spontane Stromschwankungen in verschiedenen Elektrizitätsleitern, *Ann. Phys.* **362**, 541 (1918).
 - [6] J. B. Johnson, The Schottky effect in low frequency circuits, *Phys. Rev.* **26**, 71 (1925).
 - [7] M. Gardner, Mathematical games, *Sci. Am.* **233**, 174 (1975).
 - [8] E. Milotti, 1/f noise: A pedagogical review, [arXiv:physics/0204033](https://arxiv.org/abs/physics/0204033).
 - [9] D. Halford, A general mechanical model for $|f|^\alpha$ spectral density random noise with special reference to flicker noise $1/|f|$, *Proc. IEEE* **56**, 251 (1968).
 - [10] R. F. Voss, Linearity of 1/f Noise Mechanisms, *Phys. Rev. Lett.* **40**, 913 (1978).
 - [11] J. Timmer and M. König, On generating power law noise, *Astron. Astrophys.* **300**, 707 (1995).
 - [12] P. H. Handel, 1/f Noise—An “Infrared” Phenomenon, *Phys. Rev. Lett.* **34**, 1492 (1975).
 - [13] S. Gull and G. Daniell, Image reconstruction from incomplete and noisy data, *Nature (London)* **272**, 686 (1978).
 - [14] W. H. Press, Flicker noises in astronomy and elsewhere, *Comments Phys., C - Comments Astrophys.* **7**, 103 (1978).
 - [15] J. M. Epstein, A. W. Cross, E. Magesan, and J. M. Gambetta, Investigating the limits of randomized benchmarking protocols, *Phys. Rev. A* **89**, 062321 (2014).
 - [16] T. E. O’Brien, B. Tarasinski, and B. M. Terhal, Quantum phase estimation of multiple eigenvalues for small-scale (noisy) experiments, *New J. Phys.* **21**, 023022 (2019).
 - [17] J. M. Boter, X. Xue, T. Krähenmann, T. F. Watson, V. N. Premakumar, D. R. Ward, D. E. Savage, M. G. Lagally, M. Friesen, S. N. Coppersmith *et al.*, Spatial noise correlations in a Si/SiGe two-qubit device from Bell state coherences, *Phys. Rev. B* **101**, 235133 (2020).
 - [18] J. Preskill, Quantum computing in the NISQ era and beyond, *Quantum* **2**, 79 (2018).
 - [19] B. Bauer, S. Bravyi, M. Motta, and G. K.-L. Chan, Quantum algorithms for quantum chemistry and quantum materials science, *Chem. Rev.* **120**, 12685 (2020).
 - [20] E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, 2003).

- [21] J. Lee Rodgers and W. A. Nicewander, Thirteen ways to look at the correlation coefficient, *Am. Stat.* **42**, 59 (1988).
- [22] W. E. Purcell, Variance noise spectra of $1/f$ noise, *J. Appl. Phys.* **43**, 2890 (1972).
- [23] On the other hand, if it is not a (two-point) correlation function but noise as a phenomenon that is in the focus, one should note that the former does not fully describe the latter. Various procedures have been put forward to test if the noise—or the noisy system—is equilibrium, Markovian, Gaussian, linear, time-reversal symmetric, or stationary [8,10,24].
- [24] M. Nelkin and A.-M. S. Tremblay, Deviation of $1/f$ voltage fluctuations from scale-similar Gaussian behavior, *J. Stat. Phys.* **25**, 253 (1981).
- [25] R. T. Cox, Probability, frequency and reasonable expectation, *Am. J. Phys.* **14**, 1 (1946).
- [26] Remarkably, Bayesian statistics rejects the idea of observation as a sampling of the “true” probability distribution of a variable. Actually, the “true” probability distribution is an ill-defined concept, in contrast to a probability distribution that just reflects our knowledge. These distinctions might seem irrelevant at first sight, but they become decisive on numerous occasions. The method presented in this paper is one of them, as we will explicitly point out.
- [27] Another outstanding achievement of Bayesian statistics is the formulation of the whole statistical mechanics from information theory, requiring a minimal set of axioms [28].
- [28] E. T. Jaynes, Information theory and statistical mechanics, *Phys. Rev.* **106**, 620 (1957).
- [29] E. T. Jaynes, Bayesian spectrum and chirp analysis, in *Maximum-Entropy and Bayesian Spectral Analysis and Estimation Problems*, edited by C. R. Smith and G. J. Erickson (Springer Netherlands, Dordrecht, 1987), pp. 75–84.
- [30] C. Granade, J. Combes, and D. G. Cory, Practical Bayesian tomography, *New J. Phys.* **18**, 033024 (2016).
- [31] C. E. Shannon, A mathematical theory of communication, *ACM SIGMOBILE Mob. Comput. Commun. Rev.* **5**, 3 (2001).
- [32] E. Jaynes, On the rationale of maximum-entropy methods, *Proc. IEEE* **70**, 939 (1982).
- [33] To illustrate, assume that the variable x is a voltage. The prior invariance that we impose means the requirement that the statistical inference on x will not be changed upon assigning its mean to be 0 volts instead of, say, 10 volts, or measuring the fluctuations of x away from its mean in millivolts instead of volts.
- [34] As yet another illustration, keeping the lowest two terms in the Taylor expansion of $\ln P(\nu|\hat{x})$ around its maximum, one gets the central limit expression $P(\nu|\hat{x}) \propto \exp[-M(\nu - \bar{\nu})^2/2\bar{\Lambda}]$, showing that $\bar{\Lambda}/M$ is the variance of the posterior in the large- M limit.
- [35] Sometimes one can use directly the non-normalizable distribution, here $1/|\nu - \bar{\nu}|$, for example in algorithms using sampling where only likelihood ratios enter [36].
- [36] A. Doucet, N. Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice* (Springer, New York, 2001).
- [37] Thus, the results below do not apply for unevenly sampled data, a scenario typical in, for example, astronomy [38–40].
- [38] N. R. Lomb, Least-squares frequency analysis of unequally spaced data, *Astrophys. Space Sci.* **39**, 447 (1976).
- [39] J. D. Scargle, Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data, *Astrophys. J.* **263**, 835 (1982).
- [40] J. T. VanderPlas, Understanding the Lomb–Scargle Periodogram, *Astrophys. J. Suppl. Ser.* **236**, 16 (2018).
- [41] Another usual name for Λ_k is the power spectrum. If the frequency index is continuous rather than discrete (see Sec. VI B for more on models with continuous k), the usual name is power spectral density (PSD). In this case, one might add a qualifier, calling Λ_k the auto-PSD, to distinguish it from the cross-PSD discussed in the next section.
- [42] As a mnemonic, d_k equals one half of the number of “(real) degrees of freedom” of the (in general complex) variable α_k .
- [43] M. A. Caloyannides, Microcycle spectral estimates of $1/f$ noise in semiconductors, *J. Appl. Phys.* **45**, 307 (1974).
- [44] Another method to generate $1/f$ noise from white noise is discussed in Ref. [14]. Reference [45] gives a different method and includes a list of references to further ones in its introduction.
- [45] E. Milotti, Exact numerical simulation of power-law noises, *Phys. Rev. E* **72**, 056701 (2005).
- [46] For example, Ref. [47] investigated this distribution experimentally and compared the findings with a model where the distribution is Gaussian (which it is not).
- [47] U. J. Strasilla and M. J. O. Strutt, Narrow-band variance noise, *J. Appl. Phys.* **45**, 1423 (1974).
- [48] As a side remark, we note that from Eqs. (24) and (21) we can get the prior distribution for the periodogram value $P(\bar{\Lambda}_k) \propto P(\bar{\Lambda}_k|\Lambda_k)/P(\Lambda_k|\bar{\Lambda}_k) \propto 1/\bar{\Lambda}_k$ for $k \neq 0$ and $\propto 1/\bar{\Lambda}_k^{1/2}$ for $k = 0$. Thus, the prior for a finite frequency element of the periodogram is $1/\bar{\Lambda}_k$, as expected for a scale variable. However, the symmetry between the priors for Λ_k and $\bar{\Lambda}_k$ is not valid for $k = 0$.
- [49] Here and in further we use the hypergeometric function notation putting the upper parameters up, the lower parameters down, and the function argument behind a vertical bar. In case there are multiple upper or lower parameters, they are separated by commas. We do not use the p and q indexes such as ${}_pF_q$ as the number of the upper and lower parameters can be easily counted.
- [50] That the scale variables $\bar{\Lambda}_k^A$ and $\bar{\Lambda}_k^B$ are not needed is another advantage of the parametrization through the dimensionless variable s_k . See Appendix E for a different choice.
- [51] One might be interested in different parametrizations, most often using the unnormalized correlation strength Λ_k^{AB} instead of s_k . The single additional estimating distribution needed for this parametrization is given in Appendix E.
- [52] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*, 9th ed., Dover books on Mathematics (Dover Publishing, New York, 2013).
- [53] But not at every k : as a result of statistical fluctuations, the noise of uncorrelated variables shows false correlations (a peak away from $s_k = 0$) at some frequencies. This behavior is to be expected. At these values, the estimating distribution of the phase is localized (not spread to whole $[0, 2\pi)$ interval), which is consistent with a finite value of $\text{mle}(s_k)$. One could come up with a statistical distribution for their number analogous to Eq. (24). However, one can judge whether the given two variables are or are not correlated even without such tests, see Sec. VI D.

- [54] Indeed, a continuum model can be adopted even if A_i are undefined for intermediate “times” (or if i does not represent time, but an index that is inherently integer, such as the rank). There is no difference from the point of view of the estimation. The difference would be in the interpretation: it might be unclear what a continuum function $\Lambda(\kappa)$, even if fitting a simple shape well, represents. In this case, one might be better off postulating a candidate shape directly in the discrete space, with the benefit that no \mathcal{F} discussed in the next subsection arises.
- [55] While Ref. [56] discusses a slightly different “frequency distribution”, the following citation from page 60 there is exactly to the point here: “But it is often possible to find a simple form that fits the frequency distribution approximately. If this can be done it has the advantage of describing the results of the statistics briefly. In some cases it is suggestive of the causes that lie behind the results. But the main reason, in general, for looking for a simple mathematical “law” of this type, is that if it is found it is believed to have predictive value. That is to say the simple law, if it is a very good approximation to the distribution function F of the original sample, is likely to describe the distribution function of another sample (or of the whole population) *even better than F would*. This is partly because it is likely that there are a few predominating causes lying behind the statistics, even though these cases are unknown”.
- [56] I. J. Good, *Probability and the Weighing of Evidence*, 1st ed. (Charles Griffin, London, 1950).
- [57] The typical shapes are: a constant representing a white noise, that is a Markovian environment; a Lorentzian for an environment equivalent to a two-level system with fixed transition rates [58], a $1/f$ for a collection of two-level systems with a specific distribution of transition rates (Ref. [8] attributes it to Johnson 1925 and Schottky 1926); $1/f^2$ for a random walk [59]. Further, a substantial volume of results exists on ARMA models [60], with spectrum a general rational function of k .
- [58] S. Machlup, Noise in semiconductors: Spectrum of a two-parameter random signal, *J. Appl. Phys.* **25**, 341 (1954).
- [59] M. C. Wang and G. E. Uhlenbeck, On the Theory of the Brownian Motion II, *Rev. Mod. Phys.* **17**, 323 (1945).
- [60] S. Kay and S. Marple, Spectrum analysis—A modern perspective, *Proc. IEEE* **69**, 1380 (1981).
- [61] F. Harris, On the use of windows for harmonic analysis with the discrete Fourier transform, *Proc. IEEE* **66**, 51 (1978).
- [62] As an illustration, Ref. [43], which measured $1/f$ noise down to one of the lowest frequencies ever (for a semiconductor), found the need for “pre-whitening”, “windowing”, “de-aliasing”, and “post-greening” of the measured data.
- [63] The evaluation of $\mathcal{F}[C](\kappa_k)$ does not need high precision: a precision comparable to the error bar of Λ_k suffices.
- [64] In the nomenclature of Ref. [4], it requires “prior knowledge... to supplement the data”.
- [65] The “windowing”, et cetera, can also be viewed as useful visualization tools.
- [66] While the entropy-maximization property of Gaussian distributions is known (see Refs. [32,67], Theorem 9.6.5 on page 234 in Ref. [68] or Sec. 7 and Sec. 11 in Ref. [20]), we include the proof for the article completeness.
- [67] B. S. Choi and T. M. Cover, A Proof of Burg’s Theorem, in *Maximum-Entropy and Bayesian Spectral Analysis and Estimation Problems*, edited by C. R. Smith and G. J. Erickson (Springer Netherlands, Dordrecht, 1987), pp. 75–84.
- [68] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications (Wiley, New York, 1991).
- [69] We note in passing that an affine transformation, $x \rightarrow x' = ax + b$, would give a different prior for the case where the mean is nonzero, namely $P(v, \Lambda) \propto 1/\Lambda^{3/2}$.
- [70] R. Haberman, *Elementary Applied Partial Differential Equations* (Prentice Hall Englewood Cliffs, NJ, 1983), Vol. 987.
- [71] Wikipedia contributors, Homogeneous function (2021), page version ID: 1036390550.
- [72] P. Welch, The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms, *IEEE Trans. Audio Electroacoust.* **15**, 70 (1967).
- [73] First, one could consider more general relations, for example $\Lambda' = \Lambda + \lambda + s\sqrt{\Lambda\lambda}$ from where the correlation strength s would be integrated out according to some prior or estimated distribution for s . But more importantly, Eq. (G2) is only an *ad hoc* approximation; the rigorous removal requires a more complicated procedure, the explanation of which we leave for elsewhere.
- [74] First, this result still relies on the assumption of the errors being uncorrelated with signals. Otherwise the errors do become correlated through the correlation between the signals; such errors we simply can not correct for. Second, the result is counter intuitive, since acquiring additional relevant information must change the Bayes posterior. The lack of such correction uncovers a deficiency of the approximate procedure expressed by Eq. (G2) and its analogues. See Footnote 27.
- [75] For example, in the case of correlation spectrum Λ_k , Eq. (51) becomes $\log L \propto \sum_k \sigma_{\Lambda_k}^{-2} [\bar{\Lambda}_k - \mathcal{F}[C_p](\kappa_k)]^2$, see Eq. (H3), and analogously for other variables.
- [76] The exact expression for the density $p(\lambda; m)$ contains an additional factor $1/\lambda^{\delta_{k_0}/2+1}$.
- [77] A. Richardson, A comparative review of nonparametric statistics textbooks, *Am. Stat.* **73**, 360 (2019).
- [78] N. Choudhuri, S. Ghosal, and A. Roy, Bayesian estimation of the spectral density of a time series, *J. Am. Stat. Assoc.* **99**, 1050 (2004).
- [79] P. Müller and F. A. Quintana, Nonparametric Bayesian data analysis, *Stat. Sci.* **19**, 95 (2004).
- [80] P. Orbanz and Y. W. Teh, Bayesian nonparametric models, *Encyclopedia of machine learning* **1**, 1 (2010).
- [81] L. Wasserman, *All of Nonparametric Statistics*, Springer Texts in Statistics (Springer, New York, 2006).