Model-independent time-delay interferometry based on principal component analysis

Quentin Baghi, John Baker, Jacob Slutsky, and James Ira Thorpe

Model-independent time-delay interferometry based on principal component analysis

Quentin Baghi[*]

IRFU, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France

John Baker, Jacob Slutsky, and James Ira Thorpe

Goddard Space Flight Center
Mail Code 663
8800 Greenbelt Rd, Greenbelt, Maryland 20771, USA

(Dated: October 26, 2021)

With a laser interferometric gravitational-wave detector in separate free flying spacecraft, the only way to achieve detection is to mitigate the dominant noise arising from the frequency fluctuations of the lasers via postprocessing. The noise can be effectively filtered out on the ground through a specific technique called time-delay interferometry (TDI), which relies on the measurements of time-delays between spacecraft and careful modeling of how laser noise enters the interferometric data. Recently, this technique has been recast into a matrix-based formalism by several authors, offering a different perspective on TDI, particularly by relating it to principal component analysis (PCA). In this work, we demonstrate that we can cancel laser frequency noise by directly applying PCA to a set of shifted data samples, without any prior knowledge of the relationship between single-link measurements and noise, nor time-delays. We show that this fully data-driven algorithm achieves a gravitational-wave sensitivity similar to classic TDI.

I. INTRODUCTION

Gravitational waves can be detected through a constellation of satellites forming a network of laser interferometers that probe phase differences in beams exchanged between spacecraft. This concept is the core of future space-based detectors like the Laser Interferometer Space Antenna (LISA) [1] and TianQuin [2]. It was soon realized [3] that the noise caused by the stochastic fluctuations of laser frequencies cannot be canceled at the photometric detectors onboard the satellites. The reason lies in the difference between the delays they experience when being transmitted along two interferometric arms. Rather, interferometry must be performed once the telemetry is retrieved on the ground. By taking advantage of the fact that several delayed versions of the same noise component appear in different measurements, it is possible to cancel laser noise by combining and interpolating the received data streams at the right timestamps. We call this post-processing technique time-delay interferometry (TDI) [4]. TDI has been undergoing a constant development to account for increasing degrees of complexity, including optical bench motion noise [5], varying armlengths [6, 7], noise correlations [8], on-board filtering [9], clock jitters [10] and the preferred representative units of the processed data (phase or fractional-frequency) [11]. Other efforts focused on finding new noise-canceling combinations [12].

Romano and Woan [13] showed that we could think of TDI combinations as low-variance principal components of the laser noise covariance matrix. This idea relied on a matrix representation of the interferometric measurements and was also explored in Ref. [14]. Computing TDI variables is equivalent to projecting the data onto the null space of the covariance. Recently, this concept regained interest in the community. Vallisneri *et al.* [15] presented a similar algebraic computation (called TDI-∞) where they perform the singular value decomposition (SVD) of the design matrix $\boldsymbol{M}$, i.e., the matrix translating individual laser noise sources $\boldsymbol{c}$ into measured noises $\boldsymbol{y}$ so that $\boldsymbol{y} = \boldsymbol{M}\boldsymbol{c}$. They approximate the model likelihood by restricting it to the dominating terms, using the singular vectors that correspond to the null space of $\boldsymbol{M}$. Concomitantly, we derived a frequency-domain version of the covariance eigenvector decomposition [16] called PCI (for principal component interferometry), where we focused on the orthogonalization of the data with respect to secondary (i.e., non-laser) noises. We showed that this method generalizes the A, E, and T TDI variables [8] and can bring an improvement in signal-to-noise ratio (SNR) when the levels of the noises affecting the single-link measurements are unequal. Although the implementation details are still under debate [17–19], the matrix representation of TDI is a powerful tool for data analysis, as it allows one to express the model likelihood directly in terms of single-link measurements and readily account for data gaps.

TDI and the extensions cited above rely on accurate modeling of the dependence of the measurements on the various laser noise sources. TDI stems from a set of equations modeling every interferometer, from which one derives the sequence of combination and delaying operations needed to precisely cancel laser noise terms. Equivalently, TDI-∞ assumes perfect knowledge of the design matrix, while PCI assumes a model of the noise covariance. In addition, these methods need precise estimates of the time-delays, which can be obtained independently from the telemetry [20] or directly estimated from the data using noise power minimization [21] or, as it was

---

recently shown, by Bayesian inference [16, 22]. However, LISA is a complex apparatus involving many control laws, couplings, and frequency planning inducing correlations among lasers. Inaccurate modeling can lead to imperfect cancellation or even no laser noise suppression at all [9, 10]. In this work, we present a fully data-driven approach to cancel laser noise in space interferometry, with no prior assumption on the noise structure. The method (that we name aPCI for automatic PCI) relies on the principal component analysis (PCA) of a matrix directly constructed from the measurements.

PCA is usually used to reduce the dimensionality of a problem by decomposing the data into an orthogonal basis whose vectors are called principal components. Then, one keeps only the vectors that contain the most information about the data by restricting the analysis to the components having the largest variance. However, when the noise overwhelmingly dominates the signal of interest, as this is the case for gravitational waves (GW) by about a factor 10^8, we can use PCA with the opposite objective: to retain the components having the lowest variance.

In most problems, the decomposition is performed on a collection of n different realizations of the same p-dimensional measurement, forming a matrix of size $n \times p$. However, in a gravitational-wave observatory like LISA, the data is instead made of different channels affected by different noises and measuring different projections of the same signal. To build a data matrix that we can process with PCA, we can draw inspiration from the technique used to form the TDI combinations [23], which relies on interpolators like Lagrange filters to produce delayed versions of the measured data. The measurement estimate at a given point in time is interpolated by linearly combining nearby data samples. Thus, we build a $n \times p$ data matrix whose rows are versions of the same data shifted by an integer number of samples. Then, we compute the PCA of this matrix to obtain components organized in decreasing order of variance. The lowest variance components correspond to combinations where laser noise is significantly reduced. Projecting the data onto these components yields data streams that are readily usable for gravitational-wave searches.

We first detail the principle of the aPCI method in Section II, and present the theoretical framework. In Section III, we provide a means to evaluate the method performance by deriving the sensitivity of aPCI combinations to monochromatic GWs in the case of LISA. In Section IV, we use a numerical simulation of LISA data to demonstrate laser noise suppression, and also compare aPCI with classic TDI by computing their sensitivities averaged over sky location and orientation. Finally, in Section V we discuss important aspects of this approach and pave the way for future studies.

II. PRINCIPAL COMPONENT ANALYSIS THEORETICAL FRAMEWORK

A. Construction of the data matrix

Space-based gravitational-wave detector data include the outputs of several interferometers, which can be expressed in relative frequency deviations. We denote the corresponding measurements as $s_{ij}(t)$ to conform to the official convention (see, for example, Ref. [11]), where i is the index of the satellite hosting the optical bench, and j is the index of the sending spacecraft. In practice, we measure a discrete version of $s_{ij}(t)$ with N samples spaced every τ_s seconds. Therefore, we can form a vector $\boldsymbol{s}_{ij}$ with N entries given by $s_{ij}(t_n)$, where the time stamps are $t_n = n\tau_s$, $n \in [0, N-1]$. Hence, we have $\boldsymbol{s}_{ij} \equiv (s_{ij}(t_0), \ldots, s_{ij}(t_{N-1}))^{\dagger}$. We will stick to the LISA example, where the detector is made of a constellation of 3 satellites, each of them carrying 2 optical benches on board. If we consider the science measurements only, then we have 6 measurements that we can stack into a $N \times 6$ matrix $\boldsymbol{Y}$ as

$$\boldsymbol{Y} \equiv (\boldsymbol{s}_{12},\, \boldsymbol{s}_{23},\, \boldsymbol{s}_{31},\, \boldsymbol{s}_{13},\, \boldsymbol{s}_{21},\, \boldsymbol{s}_{32}). \tag{1}$$

Classical TDI algorithms usually operate in two steps. First, they compute delayed versions of discrete signals, interpolated at specific times depending on the light travel time delays along the constellation arms. In a second step, they combine these interpolated time series in such a way that laser noise terms vanish.

We would like to form laser-noise canceling combinations of the data without knowing the delays nor the way they enter the measurements. We know that the first step of TDI is usually performed through fractional-delay filter techniques which use linear combinations of samples at times before and after the target date [23]. In this process, one must choose the number of samples n_h to use before and after the desired interpolation time. In principle, the larger n_h, the better the interpolation accuracy, but the longer the computation time. To provide a fully data-driven algorithm with the ingredients necessary to automatically find the correlations among the data, we can therefore start from shifted versions of the measurement vector $\boldsymbol{s}$. Let us assume that we have measured data over a time span from $t = 0$ to $t = T_{\text{obs}}$, sampled every τ_s. We can build a matrix $\boldsymbol{X}$ containing $\boldsymbol{Y}$ itself, along with n_h backward-shifted versions of it and n_h forward-shifted versions, resulting in a matrix of size $N \times 6(2n_h + 1)$ such that:

$$\boldsymbol{X} \equiv (\boldsymbol{D}_{-n_h}\boldsymbol{Y},\, \boldsymbol{D}_{-n_h+1}\boldsymbol{Y},\, \ldots,\, \boldsymbol{D}_{n_h}\boldsymbol{Y}), \tag{2}$$

where we defined the integer delay matrix $\boldsymbol{D}_m$ of size $N \times N$ which acts on any vector $\boldsymbol{s}$ as

$$\boldsymbol{D}_m \boldsymbol{s}_n = \begin{cases} \boldsymbol{s}_{n-m} & \text{if } n \in [n_{\min}, n_{\max}] \\ 0 & \text{otherwise,} \end{cases}$$

with $n_{\min} \equiv \max(0, m)$ and $n_{\max} \equiv \min(N - 1, N - 1 + m)$. This way, the matrix $\boldsymbol{X}$ is formed of sub-vectors including 6 interferometric measurements, where a sub-vector is shifted by one time sample with respect to the previous one. In this form, the data is ready to be processed through PCA. In the following, we will set $p \equiv 2n_h + 1$ as the number of subchannels.

B. PCA decomposition of the data matrix

PCA aims at finding a transformation of the data matrix into a new set of variables which are as independent as possible with respect of the information they carry, measured in terms of variance. This is done by computing a set of vectors called principal components (PCs) representing orthogonal directions that are ordered from the one that best fits the data, to the one that is less representative. In the problem we study, the data matrix variance overwhelmingly represents laser noise. The first PCs will therefore better fit the laser noise. On the contrary, the last PCs should be more independent of laser noise, representing the remaining information present in the data.

We can compute the PCs through singular value decomposition (SVD):

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\intercal}, \qquad (3)$$

where $\boldsymbol{U}$ is a $N \times N$ unitary matrix of left-singular vectors, $\boldsymbol{S}$ is a $N \times 6p$ rectangular diagonal matrix whose non-zero elements are the singular values, $\boldsymbol{V}$ is a $6p \times 6p$ unitary matrix of right-singular vectors, and $\intercal$ is the matrix transpose operator. The PCs are given by the vectors forming the columns of $\boldsymbol{V}$.

Note that, in regular PCA, the columns of the matrix we decompose correspond to different features (like different sensor outputs) and the rows correspond to different realizations of the measurement. In the way we apply it to laser noise processing, the matrix columns include not only different interferometer outputs, but also different shifts, while the rows correspond to particular times at which the data is measured. Therefore, to draw a comparison between TDI and what we do here, the elements of particular PC vector could be interpreted as the kernels of several fractional delay filters.

The principal component decomposition of $\boldsymbol{X}$ is therefore the matrix $\boldsymbol{T}$ resulting in the transformation

$$\boldsymbol{T} \equiv \boldsymbol{X}\boldsymbol{V}. \qquad (4)$$

The rows of $\boldsymbol{T}$ are usually called the scores, but we will refer to them as the PCA (or aPCI) combinations in the following.

In this problem, we are interested in the q lowest-variance components of the decomposition, below some threshold where the laser noise is significantly suppressed. Let us rewrite $\boldsymbol{V}$ in two parts: a part $\boldsymbol{V}_q = \boldsymbol{V}\boldsymbol{\Pi}_q$, where $\boldsymbol{\Pi}_q \equiv (\boldsymbol{0}, \boldsymbol{I}_q)^{\intercal}$ is the $6p \times q$ matrix that

projects the last q low-variance PCs, and the complementary part $\boldsymbol{V}_{\backslash q}$ containing the first $2p - q$ PCs related to large variance:

$$\begin{pmatrix} \boldsymbol{T}_{\backslash q} \\ \boldsymbol{T}_q \end{pmatrix} \equiv \boldsymbol{X} \begin{pmatrix} \boldsymbol{V}_{\backslash q} & \boldsymbol{V}_q \end{pmatrix}, \qquad (5)$$

where we split $\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{\backslash q}, & \boldsymbol{V}_q \end{pmatrix}$. To a good approximation, the low-variance transformation $\boldsymbol{T}_q$ should be all we need to search for the signals of interest, i.e. gravitational waves.

A key effect of the PCA is that the transformation from $\boldsymbol{X}$ to $\boldsymbol{T}$ diagonalizes the data matrix sample covariance $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}} = \frac{1}{N}\boldsymbol{X}^{\intercal}\boldsymbol{X}$ and transforms it to the diagonal matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{T}} = \frac{1}{N}\boldsymbol{T}^{\intercal}\boldsymbol{T} = \frac{1}{N}\boldsymbol{S}^{\intercal}\boldsymbol{S}$. Here we are primarily interested in the *inverse* covariance, which the low-variance decomposition allows us to write as:

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{T}}^{-1} &= \begin{pmatrix} \boldsymbol{V}_{\backslash q}^{\intercal}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}^{-1}\boldsymbol{V}_{\backslash q} & \boldsymbol{V}_{\backslash q}^{\intercal}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}^{-1}\boldsymbol{V}_q \\ \boldsymbol{V}_q^{\intercal}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}^{-1}\boldsymbol{V}_{\backslash q} & \boldsymbol{V}_q^{\intercal}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}^{-1}\boldsymbol{V}_q \end{pmatrix} \\ &\approx \begin{pmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{V}_q^{\intercal}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{X}}^{-1}\boldsymbol{V}_q \end{pmatrix} \\ &= \boldsymbol{\Pi}_q^{\intercal}\boldsymbol{\Pi}_q\hat{\boldsymbol{\Sigma}}_{\boldsymbol{T}}^{-1}\boldsymbol{\Pi}_q^{\intercal}\boldsymbol{\Pi}_q \qquad (6) \end{aligned}$$

where in the last line we used the diagonality of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{T}}$ and neglected the first $6p - q$ inverse variances compared to the q last ones. Here our reverse-ordered preference for the nominally least-significant principal components has the effect of selecting the most significant parts of the inverse covariance matrix.

C. Relation of the PCA process with the model covariance

As is common in gravitational-wave data analysis, we assume our noise is Gaussian and stationary (over the relatively short timescales relevant for laser noise.) Stationarity allows us to assume that frequency bins are approximately uncorrelated in the Fourier domain. For any vector $\boldsymbol{z}$, we will label as $\tilde{\boldsymbol{z}}(f)$ its discrete Fourier transform (DFT) element at frequency f. Similarly, we will denote by $\tilde{\boldsymbol{\Sigma}}_z(f)$ its Fourier-domain covariance, which fully characterizes the (zero-mean) noise.

The construction of the data matrix in Eq. (2) can be rewritten in the frequency domain as

$$\tilde{\boldsymbol{X}}(f) = \tilde{\boldsymbol{y}}^{\intercal}(f) \begin{pmatrix} \tilde{\boldsymbol{D}}_{-n_h}(f), & \dots, & \tilde{\boldsymbol{D}}_{+n_h}(f) \end{pmatrix}, \qquad (7)$$

where $\tilde{\boldsymbol{y}}(f)$ is now a 6×1 column vector, and $\tilde{\boldsymbol{X}}(f)$ is a $1 \times 6p$ row vector. Here we approximate the mth integer delay operator involved in Eq. (2) by a multiplicative complex exponential phase in the Fourier domain, encoded by a 6×6 diagonal matrix $\tilde{\boldsymbol{D}}_m(f)$ with entries

$$\tilde{D}_{m,pq}(f) = \delta_{pq}e^{-2i\pi f m \tau_s}, \qquad (8)$$

where $i = \sqrt{-1}$ is the complex number and δ_{pq} is the Kronecker symbol ($\delta_{pq} = 1$ if $p = q$ and $\delta_{pq} = 0$ otherwise).

Similarly, we can write a frequency-domain version of the projection in Eq. (4) by defining $\tilde{\boldsymbol{t}}(f)$ as the $6p \times 1$ vector formed by stacking the $6p$ PCA combinations at frequency f. Then, we have

$$\tilde{\boldsymbol{t}}(f) = \boldsymbol{V}^{\mathsf{T}} \tilde{\boldsymbol{X}}^{\mathsf{T}}(f) = \boldsymbol{W}(f)\tilde{\boldsymbol{y}}(f), \tag{9}$$

where we set $\tilde{\boldsymbol{W}}(f) \equiv \boldsymbol{V}^{\mathsf{T}}\big(\tilde{\boldsymbol{D}}_{-n_h}(f), \ldots, \tilde{\boldsymbol{D}}_{+n_h}(f)\big)^{\mathsf{T}}$ the $6p \times 6$ matrix transformation which encodes, in the Fourier domain, the construction of the data matrix and its PCA transformation.

The transformation matrix $\tilde{\boldsymbol{W}}$ has useful properties:

$$\tilde{\boldsymbol{W}}^{\dagger}\tilde{\boldsymbol{W}} = \begin{pmatrix} \tilde{\boldsymbol{D}}_{-n_h}^{\dagger} & \cdots & \tilde{\boldsymbol{D}}_{+n_h}^{\dagger} \end{pmatrix} \boldsymbol{V}\boldsymbol{V}^{\mathsf{T}} \begin{pmatrix} \tilde{\boldsymbol{D}}_{-n_h} \\ \vdots \\ \tilde{\boldsymbol{D}}_{+n_h} \end{pmatrix}$$

$$= \sum_{m=-n_h}^{+n_h} \tilde{\boldsymbol{D}}_m^{\dagger}\tilde{\boldsymbol{D}}_m = p\boldsymbol{I}_6, \tag{10}$$

where we used the fact that $\boldsymbol{V}$ is unitary. Note that here we took the Hermitian transpose $\dagger$, because $\tilde{\boldsymbol{W}}$ is a complex-valued matrix.

We denote by $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \equiv \mathrm{E}\left[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^{\dagger}\right]$ the covariance of any zero-mean complex random vector $\tilde{\boldsymbol{x}}$. Then from Eq. (9) we can write the $6p \times 6p$ covariance of $\tilde{\boldsymbol{t}}$ as

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}} = \tilde{\boldsymbol{W}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{y}}\tilde{\boldsymbol{W}}^{\dagger}, \tag{11}$$

where we dropped the dependence on f to lighten the expressions.

We seek a similar expression relating the inverse covariances. Because the transformation $\tilde{\boldsymbol{t}} = (1/p)\tilde{\boldsymbol{W}}^{\dagger}\tilde{\boldsymbol{y}}$ involves a projection from $6p$ dimensions down to 6, we do not have a useful inverse identity involving $\tilde{\boldsymbol{W}}\tilde{\boldsymbol{W}}^{\dagger}$. Consequently it is not straightforward to find an expression for the inverse covariance by solving $\boldsymbol{A}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}} = \boldsymbol{I}_{6p}$. However, if we set $\boldsymbol{A} = (1/p^2)\tilde{\boldsymbol{W}}\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\tilde{\boldsymbol{W}}^{\dagger}$, we can verify that $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}}\boldsymbol{A}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}}$ and $\boldsymbol{A}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}}\boldsymbol{A} = \boldsymbol{A}$ and that both $\boldsymbol{A}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}}\boldsymbol{A}$ are Hermitian, thus establishing that A is the pseudoinverse of $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}}$, which we simply denote as

$$\boldsymbol{\Sigma}_{\boldsymbol{t}}^{+} = (1/p^2)\tilde{\boldsymbol{W}}\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1}\tilde{\boldsymbol{W}}^{\dagger}. \tag{12}$$

Using the explicit properties of $\tilde{\boldsymbol{W}}$, we can then invert this expression to find

$$\boldsymbol{\Sigma}_{\boldsymbol{y}}^{-1} = \tilde{\boldsymbol{W}}^{\dagger}\boldsymbol{\Sigma}_{\boldsymbol{t}}^{+}\tilde{\boldsymbol{W}}. \tag{13}$$

If the sample variance from the PCA data matrix provides a sufficient estimate of the (generally distinct) data variance as regards the laser noise, then as in Section II B we can estimate $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{+} \approx \boldsymbol{\Pi}_q^{\mathsf{T}}\boldsymbol{\Pi}_q\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{+}\boldsymbol{\Pi}_q^{\mathsf{T}}\boldsymbol{\Pi}_q$. Thus

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{y}}^{-1} \approx \tilde{\boldsymbol{W}}_q^{\dagger}\boldsymbol{\Pi}_q\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{+}\boldsymbol{\Pi}_q^{\mathsf{T}}\tilde{\boldsymbol{W}}_q. \tag{14}$$

where we set $\tilde{\boldsymbol{W}}_q \equiv \boldsymbol{\Pi}_q\tilde{\boldsymbol{W}}$.

This approximation amounts to ignoring the negligible terms depending on laser noise when computing the inverse covariance, through a suitable transformation that allows us to restrict the analysis to the terms that are relevant for the GW parameter inference.

D. Bayesian inference

Gravitational-wave signals are usually analyzed using Bayesian inference. We consider the frequency-domain data $\tilde{\boldsymbol{y}}$ comprising six channels evaluated at frequency bin f. We drop the explicit dependence on f for more clarity. For a model parameterized by $\boldsymbol{\theta}$, the posterior distribution given $\tilde{\boldsymbol{y}}$ is

$$p\left(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}\right) = \frac{p\left(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)}{p\left(\tilde{\boldsymbol{y}}\right)}, \tag{15}$$

where $p\left(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}\right)$ is the likelihood, $p\left(\boldsymbol{\theta}\right)$ is the prior distribution of the parameters, and the evidence $p\left(\tilde{\boldsymbol{y}}\right)$ provides normalization.

For zero-mean Gaussian noise, the likelihood has the form:

$$p\left(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}\right) = \frac{\exp\left\{-\frac{1}{2}\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{h}}\right)^{\dagger}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{y}}^{-1}\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{h}}\right)\right\}}{\sqrt{(2\pi)^6\left|\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{y}}\right|}}, \tag{16}$$

where $\dagger$ denotes the Hermitian conjugate, $\tilde{\boldsymbol{h}}$ is the contribution to the data $\tilde{\boldsymbol{y}}$ expected from the gravitational wave signal implied by parameters $\boldsymbol{\theta}$, and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{y}}$ is the $6N \times 6N$ covariance. At this level the covariance includes all the noise processes including dominant contributions from the laser noise.

Bayesian analysis is clearest, when the data represented in the analysis comes directly from the basic measurements. In practice though some pre-treatment or approximating assumptions can be important to make the analysis practical. For LISA, the challenge is that the basic problem represented in Eq. (16) is dominated by the uninteresting laser noise. TDI processing is one approach to recast the problem without this dominant part.

Here we propose to apply our PCA treatment instead. To see how this works, we first plug the approximation for the inverse covariance derived in Eq. (14) into Eq. (16) and take the logarithm to obtain:

$$\ln p\left(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}\right) \approx -\frac{1}{2}\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{h}}\right)^{\dagger}\tilde{\boldsymbol{W}}_q^{\dagger}\boldsymbol{\Pi}_q\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{+}\boldsymbol{\Pi}_q^{\dagger}\tilde{\boldsymbol{W}}_q\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{h}}\right)$$
$$-\frac{1}{2}\ln\left((2\pi)^6\left|\boldsymbol{\Pi}_q^{\dagger}\tilde{\boldsymbol{W}}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{y}}\tilde{\boldsymbol{W}}^{\dagger}\boldsymbol{\Pi}_q\right|\right). \tag{17}$$

Then, we use the transformation in Eq. (9) to write

the log-likelihood as a function of $\tilde{\boldsymbol{t}}$:

$$\ln p\left(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}\right) = -\frac{1}{2}\left(\tilde{\boldsymbol{t}}_q - \tilde{\boldsymbol{t}}_{hq}\right)^{\dagger}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}_q\boldsymbol{t}_q}^{-1}\left(\tilde{\boldsymbol{t}}_q - \tilde{\boldsymbol{t}}_{hq}\right)$$
$$-\frac{1}{2}\ln\left((2\pi)^6\left|\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}_q\boldsymbol{t}_q}\right|\right), \quad (18)$$

where we have also defined $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}_q\boldsymbol{t}_q}^{-1} \equiv \boldsymbol{\Pi}_p\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}}^{+}\boldsymbol{\Pi}_p$, $\tilde{\boldsymbol{t}}_q \equiv \tilde{\boldsymbol{W}}_q\tilde{\boldsymbol{y}}$ and $\tilde{\boldsymbol{t}}_{hq} \equiv \tilde{\boldsymbol{W}}_q\boldsymbol{h}$. As a result, decomposing the data with PCA and keeping only the lowest variance components is equivalent to restricting the likelihood to its most significant terms (i.e., the terms that are the most relevant for GW parameters inference). The data-driven PCA process outlined here is similar to the TDI analysis, which has already been shown to be equivalent to an approximation of the likelihood [13, 15, 16]. Depending on the choice of q, Eq. (18) can be considered as a family of approximations for Eq. (16). We have argued that we expect that small q will provide a sufficient approximation, but the result become exact in the limit that $q \to 6p$.

E. Orthogonalization of aPCI streams with respect to secondary noises

To further simplify the computation of the log-likelihood in Eq (17), it is more convenient to manipulate streams which are orthogonal with respect to the noise, including its non-laser part. Therefore, we need a process equivalent to the generation of A, E, T combinations [8] in TDI. We already have generalized this operation in a previous work [16], where we simply compute the eigenvectors of the covariance matrix:

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{t}}(f) = \boldsymbol{\Phi}(f)\boldsymbol{\Lambda}(f)\boldsymbol{\Phi}(f)^{\dagger}, \quad (19)$$

where $\boldsymbol{\Phi}(f)$ is the matrix of eigenvectors, and $\boldsymbol{\Lambda}(f)$ is the diagonal matrix of eigenvalues. Note that we perform this decomposition for each frequency. Ultimately, the orthogonal PCA combinations are given by

$$\tilde{\boldsymbol{t}}_{\perp}(f) = \boldsymbol{\Phi}(f)^{\dagger}\tilde{\boldsymbol{t}}(f), \quad (20)$$

and their PSDs are given by the diaogonal elements of $\boldsymbol{\Lambda}(f)$. We have now a process to compute the noise levels necessary to derive the gravitational strain sensitivities.

III. EVALUATION OF APCI PERFORMANCE

The right quantity to assess the performance of the PCA process to analyse gravitational-wave measurements is the sensitivity of the PC projections, which we derive in this section.

A. Orbit assumptions

We assume that the LISA spacecraft follows Keplerian orbits approximated to first order in eccentricity, with

$e = 4.8237 \times 10^{-3}$. We assume that the spacecraft form an equilateral triangle whose center follows a circular orbit around the Sun, such that its position at any time t is given by

$$\boldsymbol{r}_0 = (R\cos\phi_T, \, R\sin\phi_T, \, 0)^{\mathsf{T}}, \quad (21)$$

where $R = 1$ au is the distance between the Sun and the center of the constellation, and $\phi_T(t)$ is the flight path angle.

B. Derivation of gravitational-wave responses

Here we derive the aPCI variables response to gravitational-wave signals, which is needed to compute their sensitivities. At the time of reception t, the response of gravitational waves in a single link ij is given in fractional frequency by [24]:

$$y_{ij}^{\text{GW}}(t) = \frac{H_{ij}\left(t_j - \frac{\boldsymbol{k}\cdot\boldsymbol{r}_j(t_j)}{c}\right) - H_{ij}\left(t - \frac{\boldsymbol{k}\cdot\boldsymbol{r}_i(t)}{c}\right)}{2\left(1 - \boldsymbol{k}\cdot\hat{\boldsymbol{n}}_{ij}\right)}, \quad (22)$$

where we labeled the GW projection onto arm ij as H_{ij}, the spacecraft i position vector as $\boldsymbol{r}_i$, the unit vector along arm ij as $\hat{\boldsymbol{n}}_{ij} = (\boldsymbol{r}_i - \boldsymbol{r}_j)/\|\boldsymbol{r}_i - \boldsymbol{r}_j\|$ and the time of emission as $t_j = t - L_{ij}/c$, which depends on L_{ij}, the effective distance between the sending spacecraft j and the receiving spacecraft i. We compute the vectors $\boldsymbol{r}_i$ from the orbital motion postulated in Section III A. In this equation, we also defined $\boldsymbol{k}$ as the direction of GW propagation, which can be expressed as a function of the source's ecliptic latitude β and longitude λ:

$$\boldsymbol{k} = -\left(\cos\lambda\cos\beta, \, \sin\lambda\cos\beta, \, \sin\beta\right)^{\mathsf{T}}. \quad (23)$$

The GW projection can be written as a function of antenna pattern functions F_+ and $F_\times$ such that

$$H_{ij}(\tau) = h_+(\tau)F_+\left(\psi, \boldsymbol{k}, \hat{\boldsymbol{n}}_{ij}\right) + h_\times(\tau)F_\times\left(\psi, \boldsymbol{k}, \hat{\boldsymbol{n}}_{ij}\right), \quad (24)$$

where h_α is the strain deformation in the source frame for the polarization mode α. The antenna pattern functions depend on the GW polarization angle ψ, the wave propagation vector $\boldsymbol{k}$ and the link vector $\hat{\boldsymbol{n}}$ as

$$F_+\left(\psi, \boldsymbol{k}, \hat{\boldsymbol{n}}\right) = \cos(2\psi)\xi_+\left(\boldsymbol{k}, \hat{\boldsymbol{n}}\right) - \sin(2\psi)\xi_\times\left(\boldsymbol{k}, \hat{\boldsymbol{n}}\right);$$
$$F_\times\left(\psi, \boldsymbol{k}, \hat{\boldsymbol{n}}\right) = \sin(2\psi)\xi_+\left(\boldsymbol{k}, \hat{\boldsymbol{n}}\right) + \cos(2\psi)\xi_\times\left(\boldsymbol{k}, \hat{\boldsymbol{n}}\right), \quad (25)$$

where ξ_+ and $\xi_\times$ are modulation functions defined as

$$\xi_+\left(\boldsymbol{k}, \hat{\boldsymbol{n}}\right) = \left(\boldsymbol{u}\cdot\hat{\boldsymbol{n}}\right)^2 - \left(\boldsymbol{v}\cdot\hat{\boldsymbol{n}}\right)^2;$$
$$\xi_\times\left(\boldsymbol{k}, \hat{\boldsymbol{n}}\right) = 2\left(\boldsymbol{u}\cdot\hat{\boldsymbol{n}}\right)\left(\boldsymbol{v}\cdot\hat{\boldsymbol{n}}\right). \quad (26)$$

The vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ in the above equation form the source localization basis $(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{k})$, with

$$\boldsymbol{u} = (\sin\lambda, \, -\cos\beta, \, 0)^{\mathsf{T}};$$
$$\boldsymbol{v} = (-\sin\beta\cos\lambda, \, -\sin\beta\sin\lambda, \, \cos\beta)^{\mathsf{T}}. \quad (27)$$

To write the response as a function of the source frequency, it will be useful to express the GW strain in terms of its Fourier transform $\tilde{h}_\alpha(f)$:

$$h_\alpha(t) = \int_{-\infty}^{+\infty} \tilde{h}_\alpha(f')e^{2i\pi f't}df'. \tag{28}$$

We plug Eq. (28) into Eq. (22) to get

$$y_{ij}^{\text{GW}}(t) = \sum_{\alpha=+,\times} \int_{-\infty}^{+\infty} \tilde{h}_\alpha(f')e^{2i\pi f't}df' \frac{F_\alpha(\psi, \boldsymbol{k}, \hat{\boldsymbol{n}}_{ij})}{2(1-\boldsymbol{k}\cdot\hat{\boldsymbol{n}}_{ij}))} \\ \times \left[e^{-2i\pi f'(L_{ij}+\boldsymbol{k}\cdot\boldsymbol{r}_j(t_j))/c} - e^{-2i\pi f'\boldsymbol{k}\cdot\boldsymbol{r}_i(t)/c} \right]. \tag{29}$$

In the following, we assume that the gravitational wave is monochromatic, with frequency f. Choosing a null phase, the Fourier transform of $h_\alpha(t)$ can then be written as $\tilde{h}_\alpha(f') = \frac{1}{2}A_\alpha\delta(f'-f)$ for positive frequencies. Inserting this into Eq. (30) gives

$$y_{ij}^{\text{GW}}(t,f) = \sum_{\alpha=+,\times} \frac{1}{2}A_\alpha e^{2i\pi ft}\frac{F_\alpha(\psi, \boldsymbol{k}, \hat{\boldsymbol{n}}_{ij})}{2(1-\boldsymbol{k}\cdot\hat{\boldsymbol{n}}_{ij}))} \\ \times \left[e^{-2i\pi f(L_{ij}+\boldsymbol{k}\cdot\boldsymbol{r}_j(t_j))/c} - e^{-2i\pi f\boldsymbol{k}\cdot\boldsymbol{r}_i(t)/c} \right]. \tag{30}$$

The modes amplitudes A_α can be expressed as a function of the inclination ι of the system with respect to the line of sight:

$$A_+ = A\left(1+\cos^2\iota\right); \\ A_\times = -2A\cos\iota, \tag{31}$$

where A is the amplitude of the source, that we consider constant in our monochromatic assumption.

We can now derive the response of PCA projections to GWs using the time-domain equivalent of Eqs. (9):

$$\boldsymbol{t}^{\text{GW}}(f) = \boldsymbol{V}^\dagger \begin{pmatrix} \boldsymbol{y}^{\text{GW}}(t+n_h\tau_s, f) \\ \vdots \\ \boldsymbol{y}^{\text{GW}}(t-n_h\tau_s, f) \end{pmatrix}^\dagger, \tag{32}$$

where $\boldsymbol{y}^{\text{GW}}(t,f)$ is simply the vector of single-link GW responses

$$\boldsymbol{y}^{\text{GW}}(t,f) \equiv \begin{pmatrix} y_{12}^{\text{GW}}(t,f) \\ y_{23}^{\text{GW}}(t,f) \\ y_{31}^{\text{GW}}(t,f) \\ y_{13}^{\text{GW}}(t,f) \\ y_{21}^{\text{GW}}(t,f) \\ y_{32}^{\text{GW}}(t,f) \end{pmatrix}, \tag{33}$$

whose entries are given by Eq. (30).

C. Computation of sky-averaged sensitivity

In Sec. II C we derived the frequency-domain covariance of the aPCI combinations as a function of the single-link data covariance $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{y}}$, while in Sec. III B we derived their response to gravitational waves. Following Babak et al. [25], we define the instantaneous strain sensitivity of a given component m as its noise PSD divided by its response averaged over all sky locations and orientations:

$$S_{h,m}(f) = A^2 \frac{S_{\perp,m}(f)}{\langle|t_{\perp,m}^{\text{GW}}(f)|^2\rangle}, \tag{34}$$

where $\boldsymbol{t}_\perp^{\text{GW}} \equiv \boldsymbol{\Phi}(f)^\dagger \boldsymbol{t}^{\text{GW}}$ is the GW response in the orthogonalized aPCI combinations, $S_{\perp,m}(f) \equiv \Lambda_{mm}(f)$ is the noise PSD given by the diagonal elements of matrix $\Lambda(f)$ in Eq. (19), and the brackets denote the averaging over sky location angles, polarization and inclination:

$$\langle|t_{\perp,m}^{\text{GW}}|^2\rangle = \frac{1}{4\pi}\int_{\boldsymbol{k}}\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{2}\int_{-1}^{+1}|t_{\perp,m}^{\text{GW}}|^2 d^3\boldsymbol{k}d\psi d\cos\iota. \tag{35}$$

In this definition, the sensitivity corresponds to the level of strain noise within an infinitesimal frequency band df (as stated in, e.g., [26]). Note that we include the factor A^2 in front of Eq. (34) to cancel the one present in the denominator via Eqs. (31), making the sensitivity independent of the amplitude, as it precisely determines the variance of A. We can sum the sensitivity over a subset of components between $m_{\min}$ and $m_{\max}$:

$$S_h(f) = \left[\sum_{m=m_{\min}}^{m_{\max}} S_{h,m}^{-1}(f) \right]^{-1}, \tag{36}$$

where the terms in the sum are the inverse sensitivities of each component, which can be viewed as their SNR per unit time and unit amplitude. In the following, we compute this sensitivity on a given example and compare it with its TDI analog.

To evaluate the sensitivity in Eq. (34), we need an analytical model for the PSD of the single-link data, at least under some assumptions. This will then allow us to derive an expression for the PSD of the PCA combinations $S_{\perp,m}(f)$. We assume that LISA's single-link measurements include two main sources of noise: a large term due to laser frequency fluctuations and a much smaller term due to other instrumental noises. We assume that the signals measured by the two optical benches on-board spacecraft 1 are given by the equations

$$s_{12}(t) = p_{21}(t-c^{-1}L_{12}) - p_{12}(t) + n_{12}(t) \\ s_{13}(t) = p_{31}(t-c^{-1}L_{13}) - p_{13}(t) + n_{13}(t), \tag{37}$$

where p_{ij} are laser noises, n_{ij} are secondary noises, and c is the speed of light, labeling by the convention noted in Section II A. We obtain the other measurements on-board spacecraft 2 and 3 by permuting the indices (modulo 3)

in Eq. (37). To simplify the analysis, we assume that the two lasers onboard each spacecraft i are identical, so that $p_{ij} = p_{ik}$. This amounts to assuming that the s_{ij} are equal to the vectors of intermediate variables η_{ij} in the TDI pre-processing (see, e.g., Ref. [27]). In this model, the measured laser noises can be seen as the result of the mixing (and delaying) of a vector of laser noise sources $\boldsymbol{p} = (\boldsymbol{p}_{12}^{\mathsf{T}}, \boldsymbol{p}_{23}^{\mathsf{T}}, \boldsymbol{p}_{31}^{\mathsf{T}})^{\mathsf{T}}$. We consider the $6N$ column-vector version of the $N \times 6$ matrix $\boldsymbol{Y}$ in Eq. (1), which stacks all the measurements as

$$\boldsymbol{y} \equiv (\boldsymbol{s}_{12}^{\mathsf{T}}, \boldsymbol{s}_{23}^{\mathsf{T}}, \boldsymbol{s}_{31}^{\mathsf{T}}, \boldsymbol{s}_{13}^{\mathsf{T}}, \boldsymbol{s}_{21}^{\mathsf{T}}, \boldsymbol{s}_{32}^{\mathsf{T}})^{\mathsf{T}}. \tag{38}$$

This mixing is represented by a $6N \times 3N$ design matrix $\boldsymbol{M}$, such that the model for $\boldsymbol{y}$ is

$$\boldsymbol{y} = \boldsymbol{M}\boldsymbol{p} + \boldsymbol{n}, \tag{39}$$

where we set $\boldsymbol{n} \equiv (\boldsymbol{n}_{12}^{\mathsf{T}}, \ldots, \boldsymbol{n}_{32}^{\mathsf{T}})^{\mathsf{T}}$, using the same ordering as in Eq. (38). Using this assumption, we have

$$\boldsymbol{M} = \begin{pmatrix} -\mathbf{1} & \boldsymbol{D}_{12} & \mathbf{0} \\ \mathbf{0} & -\mathbf{1} & \boldsymbol{D}_{23} \\ \boldsymbol{D}_{31} & \mathbf{0} & -\mathbf{1} \\ -\mathbf{1} & \mathbf{0} & \boldsymbol{D}_{13} \\ \boldsymbol{D}_{21} & -\mathbf{1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{D}_{32} & -\mathbf{1} \end{pmatrix}, \tag{40}$$

where $\boldsymbol{D}_{lm}$ represents the fractional delay operator acting on any discretized signal $x(t_n)$ as $\boldsymbol{D}_{lm}\boldsymbol{x}|_n \approx x(t - L_{lm}/c)$.

In the following, we assume that $\boldsymbol{p}_{ij}$ and $\boldsymbol{n}_{ij}$ are zero-mean stationary Gaussian noises vectors of size N described by PSDs $S_{p,ij}(f)$ and $S_{n,ij}(f)$, respectively. We further assume that these entries are uncorrelated. Since the laser noise level is much larger than the other noises, we have $S_{p,ij}(f) \gg S_{n,ij}(f) \, \forall i, j$. For this reason, we will refer to $\boldsymbol{n}$ as the vector of secondary noises.

D. Practical implementation

For a number of aPCI variables, we compute the orientation and sky-averaged sensitivity following Eqs. (36). We perform the integration with respect to the polarization angle analytically, which is equivalent to setting $\psi_{\text{eq}} = \pi/4$ (as it corresponds to equal contributions between the $+$ and $\times$ modes). We also use an analytical result for the integration over inclination, as it amounts to multiplying the squared response by $4/5$ [28]. Finally, we average over sky locations numerically by randomly drawing 3000 values of the source ecliptic latitude β and longitude λ from a uniform distribution in the interval $[0, \pi] \times [0, 2\pi]$.

Suppose we have computed the SVD decomposition. Among the $6p$ PCs provided by the algorithm, let us consider only q of them. For each sky location (λ, β) and each GW frequency f, we compute the response by following 3 steps:

1. Compute the arm responses y_{ij}^{GW} using Eq. (30);

2. Form the q PCA projections $\boldsymbol{t}_m^{\text{GW}}$ with Eq. (32);

3. Orthogonalize the projections using Eq. (20);

To get a baseline for comparison, we also compute the sensitivity of first-generation TDI, as defined in, e.g., Ref. [29]. The computation steps follow a similar scheme as above. The first step is the same: we calculate the arm response to harmonic GWs exactly as before. Step 2 is replaced by the computation of the Michelson TDI combinations X, Y and Z. This is done by applying the TDI 1.5 transfer function to the single-link response vector $\boldsymbol{t}_{\text{TDI}}^{\text{GW}} = \boldsymbol{V}_{\text{TDI}}^{\dagger} \boldsymbol{y}^{\text{GW}}$, where $\boldsymbol{V}_{\text{TDI}}$ can be written as

$$\boldsymbol{V}_{\text{TDI}} \equiv \begin{pmatrix} D_{13}D_{31} - 1 & (1 - D_{23}D_{32})D_{21} & 0 \\ 0 & D_{21}D_{12} - 1 & (1 - D_{31}D_{13})D_{32} \\ (1 - D_{12}D_{21})D_{13} & 0 & D_{32}D_{23} - 1 \\ 1 - D_{12}D_{21} & 0 & (D_{32}D_{23} - 1)D_{31} \\ (D_{13}D_{31} - 1)D_{12} & 1 - D_{23}D_{32} & 0 \\ 0 & (D_{21}D_{12} - 1)D_{23} & 1 - D_{31}D_{13} \end{pmatrix}. \tag{41}$$

Then, Step 3 is replaced by forming the A, E and T combinations from X, Y and Z, as defined in Ref. [4]. Finally, we compute the TDI sensitivity over all channels with a calculation similar to Eq. (36).

IV. NUMERICAL SIMULATIONS AND RESULTS

In this section, we verify that the aPCI algorithm yields a significant noise subtraction. Then we check the PSD model derived in Section II C against realizations of the noise. Finally, we compute averaged sensitivities to gravitational waves, and compare them with Sagnac TDI observables A, E and T.

A. Laser noise subtraction

We test the aPCI algorithm on simulated LISA data streams. We adopt the same assumptions as in Ref. [16]. To generate the noises, we assume that the laser sources of two optical benches belonging to the same spacecraft are the same, and that the light travel time delays between spacecraft are constant and equal to L/c, where $L = 2.5 \times 10^9$ m is the mean inter-spacecraft distance.

We simulate the science measurement time series s_{ij} from Eq. (37). We generate noises from zero-mean Gaussian stationary processes in the frequency domain described by PSDs $S_p(f)$ and $S_n(f)$ given by:

$$S_p(f) = \left(\frac{28.2 \, \text{Hz.Hz}^{-1/2}}{\nu_0} \right)^2 ;$$

$$S_n(f) = \frac{S_a(f)}{(2\pi f c)^2} + \left(\frac{2\pi f}{c} \right)^2 S_{\text{oms}}(f), \qquad (42)$$

where $\nu_0 = c \cdot (1064 \times 10^{-9}\text{m})^{-1}$ is the central frequency of the lasers, $S_a(f)$ is the test-mass acceleration noise PSD (in $\text{m}^2\text{s}^{-4}\text{Hz}^{-1}$) and $S_{\text{oms}}(f)$ is the PSD representing all other pathlength noises in the measurement (in m^2Hz^{-1}). Their expressions are given in Appendix A, where we adopted assumptions made for the LISAN-ODE simulator [30]. Likewise, we apply light travel time delays in the frequency domain. To prevent any artificial periodicity, we generate $2N$ samples corresponding to Fourier frequencies $\pm k f_s/(2N)$ before transforming the generated noise to the time domain, and discarding the second half of the data to get a time series of size N.

We choose the duration of the time series to be 12 hours, with a sampling rate of $f_s = 2$ Hz, which allows us to cover the frequency band between 0.1 mHz and 1 Hz. We choose the half-stencil size n_h so that the stencil spans the duration of one complete turn around the constellation $n_h = \lfloor 3L f_s/c \rfloor = 25$. Note that we did not find significant changes in the results presented in the following when choosing larger values for N and n_h. A full assessment of the impact of data and stencil sizes is left for future work. We form the data matrix following Eq. (2) and compute its PCA as in Eq. (4) with the algorithm provided by the PYTHON package SCIKIT-LEARN [31]. In this example, the decomposition yields 306 PCA components.

We can analyze the relative amount of noise present in each component m by calculating the variance quantity $S_{m,m}^2/(N-1)$, where $\boldsymbol{S}$ is the singular value matrix in Eq. (3). We report these values for all m on Fig. (1). We see that the variance decreases with increasing values of m, with a drop spanning 15 orders of magnitudes. This plot is a good hint about the ability of the algorithm to separate large noise from low noise components.

Now let us order the aPCI variables from the lowest variance to the largest variance, so that index $m = 1$ corresponds to the rightmost end of Fig. 1, and $m = 6p$ corresponds to the leftmost end. We project the data vector
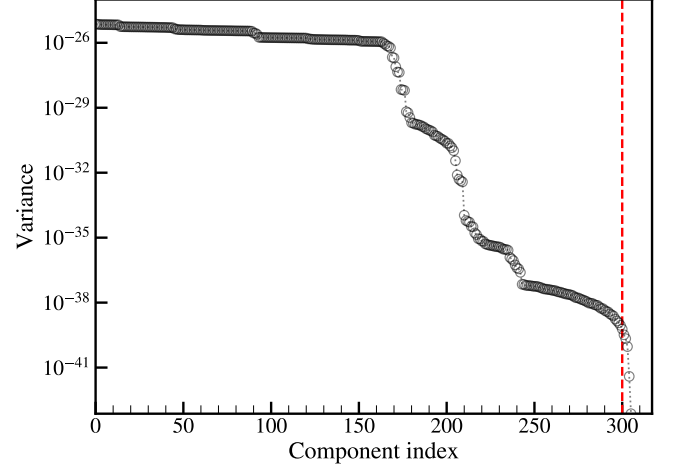

FIG. 1. Variance associated to PCA components in decreasing order. Instead of selecting the largest variance components like in classic PCA, here we are interested in the lower end of the plot, i.e. the less noisy components at the right-hand side of the red dashed line.

$\boldsymbol{y}$ onto the last 3 singular vectors using Eq. (4), obtaining 3 data projections whose periodograms are plotted in Fig. 2 in red for $m = 1$, in orange for $m = 2$, in blue for $m = 3$. We define the cross-periodogram of two time series x and y of size N as

$$P_{xy}(f) = \tilde{x}(f)\tilde{y}^*(f), \qquad (43)$$

where $\tilde{x}$ refers to the windowed discrete Fourier transform of x normalized as

$$\tilde{x}(f) \equiv \sqrt{\frac{2\tau_s}{N}} \sum_{n=0}^{N-1} w_n x_n e^{-2i\pi f n \tau_s}. \qquad (44)$$

As a comparison, we plot the periodogram of link 12 (gray curve). The difference between the gray and colored curves shows that the noise is significantly suppressed by several orders of magnitude. In addition, we verify that the noise model derived in Eq. (11) accurately predicts the noise levels in the periodograms. To do so, we include the added contributions of laser noise and secondary noises to the aPCI residuals.

Besides, we analyze the correlations among the aPCI variables by comparing the cross-periodograms $P_{t_m t_l}(f)$ with the cross spectral density (CSD) model provided by the non-diagonal terms of the covariance in Eq. (11). As an example, we plot in Fig. 3 the CSD of the pairs $(m, l) = (2, 3)$ in blue, $(3, 1)$ in brown, $(2, 1)$ in turquoise. They are all consistent with the cross-periodogram (lighter colors) averaged over 25 realizations of the noise. We see that the aPCI variables are strongly correlated, with CSD levels comparable to the PSDs. There is no reason why this should not be the case, as the PCA decomposition is built mainly from the dominant laser noise statistics. In addition, the matrix
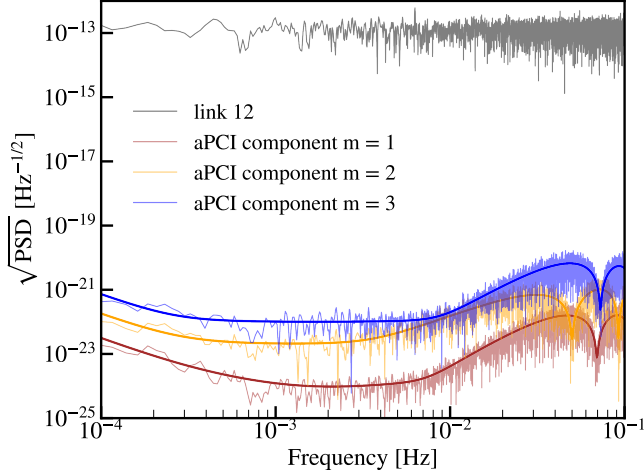
FIG. 2. Periodograms (light colors) of the three least noisy PCA projections as a function of frequency. We also plot the theoretical PSDs (dark colors) obtained with Eq. (11), expressed in units of relative frequency deviation. To compare with the input noise, we plot in gray the periodogram of the raw measurement from laser link 12, which is overwhelmed by laser noise.

$\boldsymbol{X}$ we decompose is formed from the data itself, not from the actual covariance of the data $\boldsymbol{\Sigma}_y$. As a result, the process does not give exactly orthogonal data streams by construction. That is why an extra diagonalization step can be applied, as described in Section II E, to account for secondary noises.
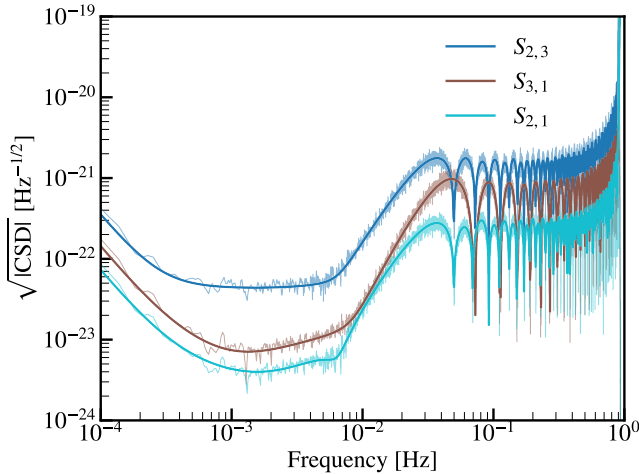
FIG. 3. Averaged cross-periodograms of the 3 aPCI variables of lowest variance (thin light lines), along with the corresponding analytical CSD model (darker solid lines). Correlations among components are comparable to their individual PSDs (plotted in Fig. 2).

While the noise levels in the aPCI streams $\boldsymbol{t}_m$ give us an indication about the amount of laser noise that is suppressed, there is little sense in comparing them directly

to the TDI PSDs. It is rather preferable to compute GW sensitivities, which we do in the next section.

B. Comparison of aPCI and TDI sensitivities

In this section, we arrange the aPCI variables in increasing order of variance, i.e. we start from the lower end of the curve in Fig. 1. We define the sensitivity of the q variables of lowest variance as the sensitivity obtained by accumulating the SNRs of their orthogonal transformations, as outlined in the steps of Section III D.

We plot the sensitivities for a number of components ranging from $q = 1$ to $q = 6$ in Fig. 4. Each blue dashed line corresponds to a value of q, with darker colors for larger q. To better distinguish the levels, we fill the areas in between two levels with related shades of blue. For example, the uppermost, lightest curve is the sensitivity obtained with a single aPCI variable ($q = 1$), i.e. the one with lowest variance. The curve that comes just below ($q = 2$) is the sensitivity obtained by using the two variables of lowest variance, and so on. As expected, we can see that the larger the number of components we include, the better the sensitivity.
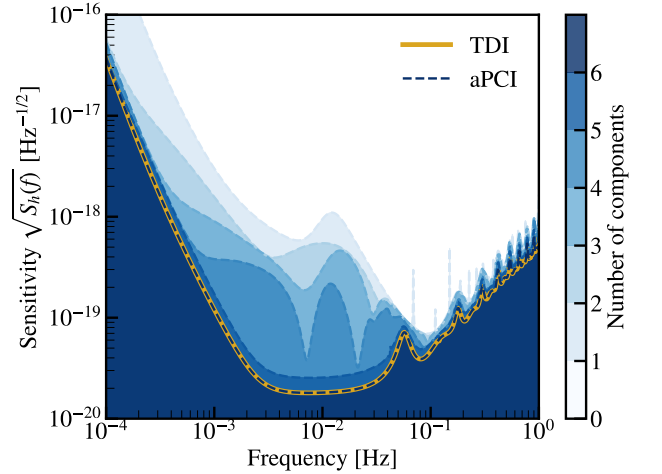

FIG. 4. Comparison of aPCI (blue dashed lines) and TDI (thick orange line) sensitivities summed over combinations. The lightest, uppermost blue dashed line corresponds to the sensitivity obtained with a single aPCI component, chosen to be the one with the smallest singular value. The darker curve just below corresponds to the accumulated sensitivity obtained by including an additional component (corresponding to the next smallest singular value). Darker blue dashed curves beneath are constructed similarly, by adding more components. The lowermost dashed blue curve is the sensitivity reached with 6 components, which almost perfectly fits the orange line representing the accumulated AET TDI sensitivity.

We also plot in Fig. 4 the summed sensitivity of TDI channels A, E and T with a thick, solid orange line. Remarkably, the figure shows that we reach the same sen-

sitivity as orthogonal TDI with 6 aPCI variables, as we can see from the superimposition of the lowest dashed blue line and the orange solid line. We have checked that adding more aPCI variables does not improve the sensitivity any further, as they are too much dominated by laser noise.

C. SNR distribution and statistical behavior

Fig. 4 suggests that we need 6 aPCI variables to reach the exact same sensitivity as the 3 orthogonal TDI combinations. To analyse the extent to which this is true, we can assess the fraction of sensitivity, or rather equivalent squared SNR, contained in each component. We do so by defining the squared SNR as simply the inverse sensitivity integrated over the frequency band:

$$\overline{\mathrm{SNR}}_m^2 \equiv \int_{f_{\min}}^{f_{\max}} S_{h,m}^{-1}(f) df, \tag{45}$$

where $S_{h,m}(f)$ is given by Eq. (34), $f_{\min} = 10^{-4}$ Hz and $f_{\max} = 1$ Hz. We are not using the term "SNR" in the usual way here, as the inverse sensitivity is just a proxy for the SNR of a monochromatic source. It can be thought of as a SNR per unit time and unit strain. Then, the squared SNR fraction of component m relative to TDI is simply $\overline{\mathrm{SNR}}_m^2$ divided by the total SNR^2 achieved by variables A, E, T.

Besides, since the PCA is performed on noisy data, singular vectors are random variables with a variability depending on the particular realization of the noise present in $\boldsymbol{y}$. Therefore, to study how the sensitivity is distributed between different components, we need to consider several noise realizations and examine this distribution on average.

For this purpose, we generate 100 realizations of the 12-hour long vector $\boldsymbol{y}$, and compute the aPCI variables for each of them, along with their theoretical PSDs. We gather the results in Fig. 5, where we plot the mean PSDs for the three variables with lowest singular values (the same as in Fig. 2), along with their 68% confidence intervals. This plot reveals the significant variability of the relative noise levels within each component.

For each noise realization, we compute the sensitivity of the 6 lowest variance aPCI variables, and the SNR^2 fraction (relative to TDI) carried by each variable. In Fig. 6, we plot the distribution of cumulative SNR^2 fraction as a function of the number of components in the form of box plots (blue), using the SEABORN package [32]. We range the components from largest to smallest SNR^2. As a reference, we plot the TDI combinations' cumulative SNR^2 when adding A, E and T in this order (horizontal yellow dashed lines).

We see that in average, most of the SNR^2 (about 94%) is concentrated in the 3 most sensitive aPCI variables (boxes on the left-hand side of the figure). While the first component always gather about 50% of the SNR^2,
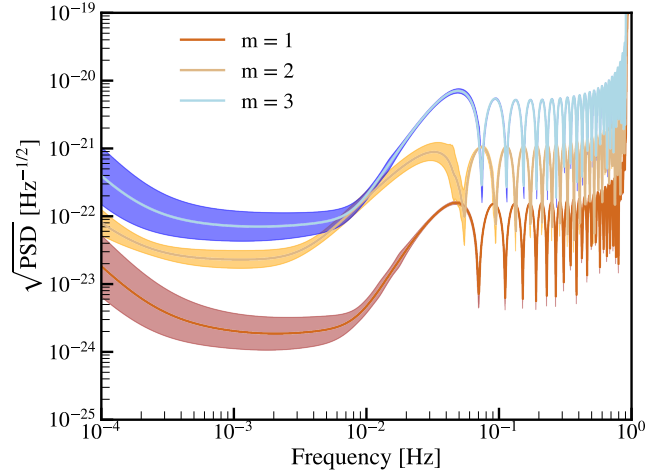

FIG. 5. Impact of noise realizations on the three least noisy aPCI variables noise PSDs. The solid lines correspond to the mean PSDs accross 100 noise realizations, while the colored areas around each line represent the 1-σ region. The indexing m is ordered along increasing variance, from $m = 1$ (lowest variance) to $m = 3$ (third lowest variance).
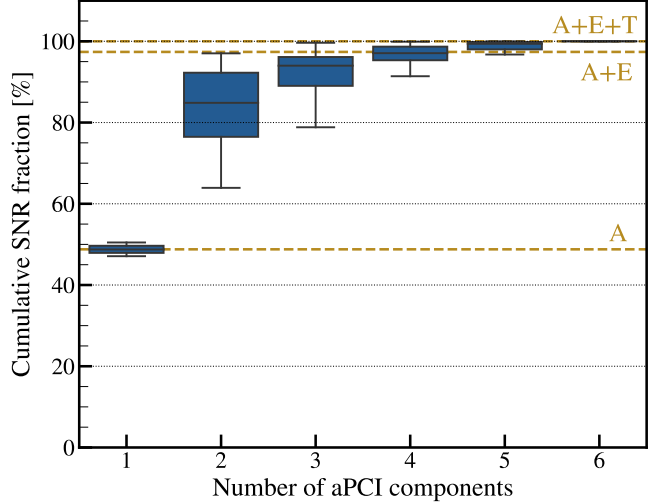

FIG. 6. Distributions of cumulative fraction of SNR^2 (in percent) as a function of the number of variables, ordered in decreasing SNR^2 fractions. The blue boxes correspond to aPCI variables and are made of 3 line showing the 3 quartiles, while the vertical lines on both sides extend to the extreme values of the distribution. The yellow horizontal lines show the cumulative SNR^2 fraction of TDI variables A, E, T. The distributions are obtained from 100 realizations of the data vector $\boldsymbol{y}$.

the heights of the boxes indicate a significant spread of the SNR^2 fraction around the median when adding components 2 and 3. The 3 last components (4, 5, 6) have marginal SNR concentrations, since their total fraction is about 6% on average. It should be noted that the last box is almost flat, which means that the total SNR

(or sensitivity) hardly changes from one realization to another. It fluctuates around 99.99751(7)% of the TDI total SNR. Therefore, the SNR contained in the last 6 aPCI variables is virtually equal to the cumulative SNR of the orthogonal TDI variables. The model independence results in negligible SNR loss, however the gravitational wave signal is diluted across more than the traditional 3 TDI variables.

D. Impact of gravitational waves

The PCA process outlined in this work acts directly on the measurements, hence the combinations that we derive depend on the content of the data. A natural question that arises is the impact of the presence of gravitational-wave signals onto the PC vectors, and thus on the overall sensitivity of the aPCI variables.

The core property which distinguishes the laser noise as we isolate it for cancellation is just that it is loud. If a sufficiently large gravitational-wave signal is present, we should expect that our process will cancel that as well, but how loud does the signal have to be for such losses to be significant?

Here, we take a first step in answering that question by simulating a quasi-monochromatic signal from a ultra-compact galactic binary, whose amplitude A is allowed to vary. For the purpose of this assessment, we consider arbitrary and unphysical amplitude values chosen to give the source particular SNRs, ranging from 1 to 10^9. We choose a fixed GW frequency equal to 1 mHz, as well as a fixed sky location. For each value of the amplitude, we create a data set by injecting the GW signal into the phasemeter measurements, in addition to a noise realization which remains the same for all sets. The duration of the time series is the same as in the previous sections, i.e. 12 hours.

For each measurement set containing the GW signal, we build the data matrix, compute its PCA decomposition, and derive the corresponding averaged sensitivity that we denote by S_h^{GW}. We compare these sensitivities to the signal-free sensitivity (derived with data containing only noise) denoted by S_h by computing their relative error maximized over the frequency range:

$$\gamma = \max_f \left\{ \frac{\left| S_h^{\mathrm{GW}}(f) - S_h(f) \right|}{S_h(f)} \right\}. \tag{46}$$

We report the results in Fig. 7 where we plot γ as a function of the source SNR. The relative error in sensitivity remains below 1 % for SNR values below or equal to 10^7. Beyond SNR = 10^8, the error becomes significant, reflecting the fact that the PCA decomposition starts to be affected by the presence of the injected signal. This demonstrates the robustness of the method in the presence of deterministic signals, as the chosen SNRs are extremely large relative to typical GW sources, especially given the short observation time we consider. As

a comparison, the total GW SNR accumulated over the mission duration should be of order $\sim 10^4$, which is way below the threshold where it starts to affect the PCA decomposition.
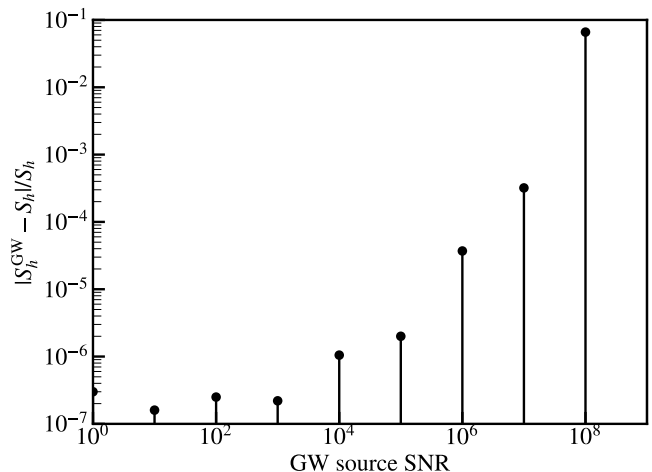

FIG. 7. Impact of a 1 mHz monochromatic GW source onto the aPCI sensitivity: relative difference of the sensitivity obtained with the presence of the source relative to the source-free sensitivity.

V. DISCUSSION

We outlined a method designed to process space-borne interferometer telemetry and minimize laser frequency noise for data analysis purposes. This method, aPCI, is completely blind: it does not rely on any model describing the measurements. The only implicit assumption is that there exists some unknown linear relationship over short times between the various measurements. We begin by building a data matrix by stacking integer-shifted versions of all single-link measurements; then we compute its PCA decomposition. Projecting the data onto the components with lowest variances yields data streams that are almost free from laser noise, and are directly usable for gravitational-wave searches. We demonstrate its equivalence with Bayesian inference based on TDI. This approach is a robust, complementary alternative to existing TDI techniques which depend on a specific model of the data (the laser noise equations) and on physical parameters (the time delays). This model-independent technique could be an important tool to ensure the robustness of the data analysis in missions such as LISA.

We demonstrated the effectiveness of the method to remove laser noise using simulated LISA single-link measurements, assuming fixed armlengths. We showed that the accumulated GW sensitivity of the 6 lowest-variance PCA projections is always effectively equal to the sensitivity of first-generation Michelson TDI combinations, and that the major part of the SNR concentrates in the

2 most sensitive channels.

This work lays the foundation for future investigations to build a fully data-driven approach to space-based interferometer data-analysis. Many aspects remain to be studied in more detail. First, the ability of the method to process measurements obtained with time-varying arm-lengths and other realistic details should be tested. Second, even though we showed that the method is robust against loud monochromatic gravitational-wave signals, we should extend this assessment to the presence of various sources across the full sensitive bandwidth, as expected in LISA. Third, one should investigate how aPCI can be further improved towards a better compression of the GW information, using orthogonality with respect to the sensitivity. Fourth, similar to TDI, aPCI would have to face data artifacts like instrumental transients and data gaps expected in LISA-like detector measurements. We should study the impact of these perturbations on aPCI performance and test its combination with mitigation techniques such as imputation [33]. Finally, an important step to take is to fully develop and demonstrate Bayesian inference analysis using this approach.

Appendix A: Pathlength noises

We provide here the analytical forumlas for the secondary noise PSDs used in this study. The acceleration noise PSD is given by

$$S_a(f) = a_{\mathrm{TM}}^2 \left[1 + \left(\frac{f_1}{f}\right)^2\right]\left[1 + \left(\frac{f}{f_2}\right)^4\right] \quad \text{(A1)}$$

where $a_{\mathrm{TM}} = 3 \times 10^{-15}\,\mathrm{ms}^{-2}$, $f_1 = 4 \times 10^{-4}$ Hz and $f_2 = 8 \times 10^{-3}$ Hz. The OMS noise PSD is

$$S_{\mathrm{oms}}(f) = a_{\mathrm{oms}}^2 \left[1 + \left(\frac{f_3}{f}\right)^4\right], \quad \text{(A2)}$$

where $a_{\mathrm{oms}} = 15 \times 10^{-12}\,\mathrm{mHz}^{-1/2}$ and $f_3 = 2 \times 10^{-3}$ mHz.

ACKNOWLEDGMENTS

[1] Karsten Danzmann, Pau Amaro-Seoane, Heather Audley, Stanislav Babak, John Baker, et al. LISA laser interferometer space antenna. *arXiv*, pages 1–41, 2017.

[2] V. K. Milyukov. TianQin Space-Based Gravitational Wave Detector: Key Technologies and Current State of Implementation. *Astronomy Reports*, 64(12):1067–1077, 2020. Publisher: Pleiades journals.

[3] Massimo Tinto and J. W. Armstrong. Cancellation of laser noise in an unequal-arm interferometer detector of gravitational radiation. *Physical Review D - Particles, Fields, Gravitation and Cosmology*, 59(10):1–11, 1999. Publisher: American Institute of Physics Inc.

[4] Massimo Tinto and Sanjeev V. Dhurandhar. Time-Delay Interferometry. *Living Reviews in Relativity*, 17(1):6, 2014. Publisher: Springer International Publishing.

[5] F. B. Estabrook, Massimo Tinto, and J. W. Armstrong. Time-delay analysis of LISA gravitational wave data: Elimination of spacecraft motion effects. *Physical Review D - Particles, Fields, Gravitation and Cosmology*, 62(4):8, 2000. Publisher: American Physical Society.

[6] Neil J. Cornish and Ronald W. Hellings. The Effects of Orbital Motion on LISA Time Delay Interferometry. *Classical and Quantum Gravity*, 20(22):4851–4860, 2003.

[7] Daniel A. Shaddock, Massimo Tinto, Frank B. Estabrook, and J. W. Armstrong. Data combinations accounting for LISA spacecraft motion.

[8] Thomas A. Prince, J. W. Armstrong, Massimo Tinto, and Shane L. Larson. LISA optimal sensitivity. *Physical Review D - Particles, Fields, Gravitation and Cosmology*, 66(12), 2002.

[9] Jean Baptiste Bayle, Marc Lilley, Antoine Petiteau, and Hubert Halloin. Effect of filters on the time-delay interferometry residual laser noise for LISA. *Physical Review D*, 99(8), 2019.

[10] Olaf Hartwig and Jean Baptiste Bayle. Clock-jitter reduction in LISA time-delay interferometry combinations. *arXiv*, May 2020.

[11] Jean-Baptiste Bayle, Olaf Hartwig, and Martin Staab. Adapting time-delay interferometry for LISA data in frequency. 2021.

[12] Martina Muratore, Daniele Vetrugno, and Stefano Vitale. Revisitation of time delay interferometry combinations that suppress laser noise in LISA. *Classical and Quantum Gravity*, 37(18):185019, aug 2020.

[13] J. D. Romano and G. Woan. Principal component analysis for LISA: The time delay interferometry connection. *Physical Review D - Particles, Fields, Gravitation and Cosmology*, 73(10), 2006.

[14] Michele Dawn Leighton. A principal component approach to space-based gravitational wave astronomy. Technical report, 2016.

[15] Michele Vallisneri, Jean Baptiste Bayle, Stanislav Babak, and Antoine Petiteau. TDI-infinity: time-delay interferometry without delays. *arXiv*, 2020.

[16] Quentin Baghi, James Ira Thorpe, Jacob Slutsky, and John Baker. Statistical inference approach to time-delay

interferometry for gravitational-wave detection. *Physical Review D*, 103(4), 2021.

[17] Massimo Tinto, Sanjeev Dhurandhar, and Prasanna Joshi. Matrix Representation of Time-Delay Interferometry. 2021.

[18] Jean-Baptiste Bayle, Michele Vallisneri, Stanislav Babak, and Antoine Petiteau. On the matrix formulation of time-delay interferometry. 2021.

[19] Massimo Tinto, Sanjeev Dhurandhar, and Prasanna Joshi. Reply to the Bayle et al. gr-qc document dated June 7, 2021. 2021.

[20] Yan Wang, Gerhard Heinzel, and Karsten Danzmann. First stage of LISA data processing: Clock synchronization and arm-length determination via a hybrid-extended Kalman filter. *Physical Review D*, 90(6):064016–064016, September 2014. Publisher: American Physical Society.

[21] Massimo Tinto, Michele Vallisneri, and J. W. Armstrong. Time-delay interferometric ranging for space-borne gravitational-wave detectors. *Physical Review D*, 71(4):041101, February 2005. Publisher: American Physical Society.

[22] Jessica Page and Tyson Littenberg. Bayesian Time Delay Interferometry. *arXiv:2107.08174 [gr-qc]*, July 2021. arXiv: 2107.08174.

[23] D A Shaddock, B Ware, R E Spero, and M Vallisneri. Postprocessed time-delay interferometry for LISA. *Physical Review D - Particles, Fields, Gravitation and Cosmology*, 70(8):081101–081101, 2004.

[24] Frank B. Estabrook and Hugo D. Wahlquist. Response of Doppler spacecraft tracking to gravitational radiation. *General Relativity and Gravitation*, 6(5):439–447, October 1975.

[25] Stanislav Babak, Martin Hewitson, and Antoine Petiteau. LISA Sensitivity and SNR Calculations. *arXiv:2108.01167 [astro-ph, physics:gr-qc]*, August 2021. arXiv: 2108.01167.

[26] Michele Vallisneri and Chad R. Galley. Non-sky-averaged sensitivity curves for space-based gravitational-wave observatories. *Classical and Quantum Gravity*, 29(12):124015, June 2012. Publisher: IOP Publishing.

[27] Markus Otto, Gerhard Heinzel, and Karsten Danzmann. TDI and clock noise removal for the split interferometry configuration of LISA. *Classical and Quantum Gravity*, 29(20):205003–205003, October 2012. Publisher: IOP Publishing.

[28] Travis Robson, Neil J Cornish, and Chang Liug. The construction and use of LISA sensitivity curves. *Classical and Quantum Gravity*, 36(10):105011–105011, May 2019. Publisher: IOP Publishing.

[29] Massimo Tinto, Frank B. Estabrook, and adn J. W. Armstrong. Time Delay Interferometry with Moving Spacecraft Arrays. *Physical Review D*, 69(8):082001, April 2004. arXiv: gr-qc/0310017.

[30] Jean-Baptiste Bayle. *Simulation and Data Analysis for LISA (Instrumental Modeling, Time-Delay Interferometry, Noise-Reduction Performance Study, and Discrimination of Transient Gravitational Signals)*. Theses, Université de Paris ; Université Paris Diderot ; Laboratoire Astroparticules et Cosmologie, October 2019.

[31] Fabian Pedregosa, Gal Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[32] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[33] Quentin Baghi, James Ira Thorpe, Jacob Slutsky, John Baker, Tito Dal Canton, et al. Gravitational-wave parameter estimation with gaps in LISA: A Bayesian data augmentation method. *Physical Review D*, 100(2):022003–022003, July 2019. Publisher: American Physical Society.