PHYSICAL REVIEW D **83,** 122006 (2011)

Blind source separation and Wigner-Ville transform as tools for the extraction of the gravitational wave signal

L. A. Forte, F. Garufi, and L. Milano*

Dipartimento di Scienze Fisiche, Università di Napoli Federico II and INFN, Sezione di Napoli

R. P. Croce, V. Pierro, and I. Pinto

Università del Sannio, Benevento

(Received 30 March 2011; published 9 June 2011)

Coalescing binaries are credited as being relatively abundant sources of gravitational radiation, with a rich content of physical information. Their signals, apart from (important) complications due to higher-order post-Newtonian corrections, spin-orbit and spin-spin couplings, etc., are so-called chirp signals, i.e. a signal modulated both in amplitude and in frequency. The rate at which the frequency changes depends basically on the chirp mass, a particular combination of the masses of the two objets. It is known that the Wigner-Ville transform is an optimal time-frequency distribution in detecting chirping signals whose instantaneous frequency grows linearly in time. We compare the performance of the plain Wigner-Ville transform and of blind source separation-augmented Wigner-Ville transform. We consider a typical chirp of interest for ground-based gravitational wave (GW) detectors and inject it at a SNR $= 12$ into two independent time series of white Gaussian noise. We show that the blind source separation preprocessing acts as a powerful denoising tool, yielding a significant enhancement in the detection capability of the Wigner-Ville transform alone. We report preliminary results, focused on detection performances, which appear to be very promising; the improvement in parameters estimation will be discussed in a forthcoming paper. The possibility to apply our analysis to a network of GW interferometers is briefly discussed. Finally, we stress the fact the our methods are completely independent on the shape of the signal, and thus have broader applications besides chirp gravitational wave signals.

DOI: 10.1103/PhysRevD.83.122006 PACS numbers: 04.80.Nn, 07.05.Kf

I. INTRODUCTION

Modern interferometric GWs detectors [1,2] have now reached design sensitivities and in the next few years they will enter a phase of technological up-grading (advanced detectors) towards better sensitivities. This will increase the effective volume of the Universe (horizon) seen by each detector. Current estimates [3,4] on the rate of compact binaries coalescence within this volume indicate that a few detections per year should be a realistic figure for advanced detectors sensitivity (although the rates are still pretty uncertain). The expected gravitational signal for compact binaries (neutron stars-neutron stars, black hole-black hole or neutron star-black hole) is a so-called chirp signal, i.e. a signal with a frequency varying in time at a rate basically determined by the chirp mass of the system (for more details see, for example, [5]). The derivation of this result is based on perturbation theory in general relativity, mostly on post-Newtonian and post-Minkowskian approximations. The waveform is also confirmed by the simulations made in Numerical Relativity. Major complications are due to the spins of the two objects, their couplings, etc.; furthermore, for a neutron star (NS), there are many models for the equation of state, and this latter affects especially the merging phase of the coalescence (which contains the most energetic part of the signal), leading to different merging signals. Data analysis techniques for the detection of chirps from compact coalescing binaries are based on templates search, i.e. one builds a grid of templates (each template is a waveform of the type predicted by perturbation theory, with physical parameters ranging in an appropriate set) and filters the signals coming out of the interferometers through this grid (a filter bank) in order to apply the maximum likelihood detection/estimation principle. The more similar the GW signal is to the template used in the search, the higher will be the detection probability. Clearly, if the signal is, for some reason, different from the expected ones, a search based on templates will be useless. This is much more true when one considers that for some astrophysical sources like supernovae events, a realistic model is very complicated and thus the shape of the sought signal is basically unknown. Until first detections will be available (including consistency among the waveforms retrieved by several detectors), one cannot be sure that the signal is the one predicted by perturbative calculations

The Wigner-Ville transform (WVT) is a reliable tool for detecting nonstationary unmodelled waveforms. In particular, it works well with chirps. On the other hand, its detection performance is not as good as those of a template-bank-based matched-filter detector. In this paper we use, for the first time in this field, the concept of blind

[*] forte@na.infn.it

source separation (BSS) as a denoising tool. The BSS technique is also independent from the shape of the sought signals, and we prove that is capable of boosting the performance of a WVT based detector considerably.

Since BSS concepts are likely new to the GW community, we discuss shortly its formulation and possible implementations in Sec. II (more details on the used algorithms are described in the Appendixes); then we present a brief review of the WVT properties in Sec. III and finally describe our simulations in Sec. IV. Future applications and refinements are discussed in Sec. V, Conclusions.

II. INDEPENDENT COMPONENT ANALYSIS

In this section we review the concepts of Blind Source Separation (BSS) and Independent Component Analysis (ICA). Note that these two terms mean slightly different things [6]. Independent Component Analysis is a just possible way to solve the BSS problem.

In BSS one has a certain number, n, of observations $x_i(t)$. Each observation is a linear, memoryless combination of n independent signals $s_i(t)$ produced by different sources, which are not directly observable. The problem is to retrieve (separate) the n signals (known as the independent components) from the n observations. The matrix A relating the observations to the signals is typically unknown. Note that all or part of the signals may be in fact pure stochastic processes, i.e. noise [7].

A classical example of BSS is the so-called cocktail party problem. Suppose you are in a room where there are several people speaking and an equal number of microphones recording their voices. Each speaker is a source originating a signal $s_i(t)$, whereas each microphone produces a time series $x_i(t)$ which contains a (linear) mixture of all signals. The problem is to separate the original waveforms using only the recorded data. Note that this is also a problem of noise reduction. In fact, if all signals (sources) are noise except one, and we are interested in retrieving that one, we face a problem of noise reduction.

Figure 1 shows a typical example (see caption for details). Note that in order to perform separation, as further discussed below, only the independence of the components and their non-Gaussianity is important; the spectral properties of the signals do not matter, and have not been used. Indeed, the results would not change if the signals were all non-Gaussian stochastic processes.

Now, let us formalize the BSS through the ICA procedure (see [8,9] for more details and precise definitions) and explain the key underlying hypotheses, namely, the statistical independence and non-Gaussianity. Let our data $x_1, \ldots, x_n$ be obtained by linearly mixing n independent random variables s_i via a $n \times n$ matrix A, i.e.

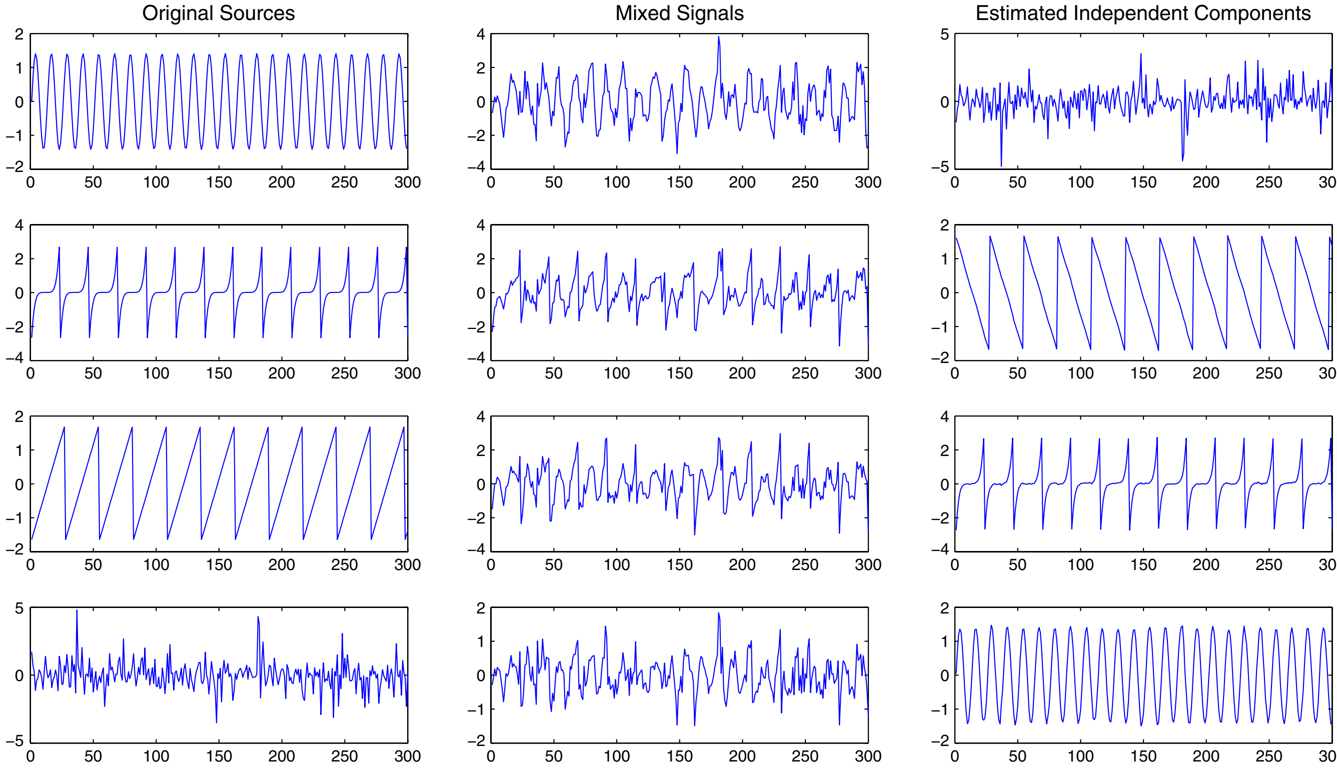

FIG. 1 (color online). Left: three deterministic signals and one random (non-Gaussian) noise are artificially generated and mixed via a random matrix (arbitrary units for both axes). Middle: the resulting observations. Right: the estimated original sources. Note that some of them are recovered up to a sign reversal, which is an ambiguity of the method.

$$\mathbf{x} = \mathbf{A}\mathbf{s}. \tag{1}$$

This is called the noise-free model, although a more realistic model would require an additive noise term **n** in the right-hand side of Eq. (1). The sources s_i are the independent components, and they are latent variables, which means that cannot be observed directly. All we observe is the vector **x**; the mixing matrix (assumed full rank) is also unknown. Thus we must estimate at the same time **A** and **s** using only the observations. This is possible (see e.g [8]) under the assumptions that the sources are statistically independent and have (possibly unknown) non-Gaussian distributions [10]: we will see why these assumptions are enough to separate the original components given only their mixtures. Once one can estimate **A**, by evaluating its inverse **W**, it is easy to estimate the sources:

$$\mathbf{s} = \mathbf{W}\mathbf{x} \tag{2}$$

The ICA model with noise is more complicated and also its well-posedness (for example the identifiability of the mixing matrix) is delicate. For time being, we omit explicitly the noise term in the determined case (number of sources = number of observations). This term will however be introduced in the under-determined case that we will study with the help of the Wigner-Ville transform (see below).

By its own nature, ICA presents some ambiguities. The first one is that one cannot recover the variances of the independent components. In fact, since both **s** and **A** are not known, any multiplicative factor in front of some s_i can be canceled by dividing by the same factor the corresponding column of **A**. This also implies that each independent component can only be determined up to a sign. Consequently, one assumes that all components have unit variance $E\{s_i^2\} = 1$; we also assume that all variables have zero mean. The second ambiguity is that in ICA there is no notion of ordering among the components. In fact, multiplying a solution by a permutation matrix would still give a solution to the same problem [11].

Let us stress that ICA is only a possible way to solve the BSS problem and the two algorithms should not be confused. Historically, there exist in literature three possible routes to the problem of BSS: non-Gaussianity, spectral diversity and nonstationarity. ICA was born to separate statistically independent, non-Gaussian components. However, different techniques allow us to separate also Gaussian components. More on this and the relative algorithms are described in the following sections.

A. Non-Gaussianity

Standard ICA works if the components are non-Gaussian (separation is still possible if there is at most one Gaussian component). In fact, for Gaussian components the mixing matrix (or equivalently its inverse) can be estimated only up to an orthogonal transformation, see [8]. We face thus the problem of measuring non-Gaussianity.

As a first measure of non-Gaussianity we can consider the kurtosis. If x is a random variable, then

$$\mathrm{kurt}\,(x) = E\{x^4\} - 3(E\{x^2\})^2, \tag{3}$$

which can be simplified if we assume x to have zero-mean and unit variance, $\mathrm{kurt}(x) = E\{x^4\} - 3$. Kurtosis is basically a normalized version of the fourth moment. For Gaussian variables, the kurtosis is zero, whereas it is nonzero for (almost) all the other distributions. From its definition, kurtosis can be positive or negative: this corresponds to super-Gaussian or sub-Gaussian variables, respectively. The former have typically some spikes in the probability density function (PDF) and heavy tails (for example the Laplace distribution), the latter have usually a flat PDF (for example the uniform distribution). Thus, non-Gaussianity can be measured by the absolute value of the kurtosis or its square. Note that to evaluate the kurtosis, it is enough to estimate the fourth moment from the observed data, so it is computationally easy. Furthermore, from its definition, these properties follow

$$\mathrm{kurt}\,(x+y) = \mathrm{kurt}(x) + \mathrm{kurt}(y) \tag{4}$$

if x, y are independent, and

$$\mathrm{kurt}\,(ax) = a^4\mathrm{kurt}(x), \tag{5}$$

where a is a scalar. To illustrate how maximization or minimization of kurtosis could be used to solve the ICA problem, let us assume that we have two original sources s_1, s_2 and the observed data $\mathbf{x} = \mathbf{A}\mathbf{s}$. We look for one of the independent components as $y = \mathbf{w}^t\mathbf{x}$ (^{t} denotes the transpose matrix). If we put $\mathbf{z} = \mathbf{A}^t\mathbf{w}$, then $y = \mathbf{z}^t\mathbf{s} = z_1 s_1 + z_2 s_2$. From the properties of kurtosis, $\mathrm{kurt}(y) = z_1^4\mathrm{kurt}(s_1) + z_2^4\mathrm{kurt}(s_2)$ and since we assume that the original sources have unit variance, it must also be $E\{y^2\} = 1$, which is the constraint $z_1^2 + z_1^2 = 1$. So the optimization problem is to find the maxima or minima of the function $|\mathrm{kurt}(y)| = |z_1^4\mathrm{kurt}(s_1) + z_2^4\mathrm{kurt}(s_2)|$ on the unit circle. It can be shown that the maxima are exactly at the points where one of the elements of the **z** vector is zero and since we are on the unit circle the other element must be 1 or -1. In other words, y must be equal to one of the components $\pm s_i$. This shows that, theoretically, maximization of kurtosis solves the ICA problem. In practice, one starts with a weight vector **w** and looks for a direction in which the kurtosis of $\mathbf{w}^t\mathbf{x}$ grows or decreases. Then one uses a gradient method or another method to find another direction.

Although kurtosis may be used as an optimization criterion for the determined ICA, it has the big drawback that it is very sensitive to outliers, i.e. its value may depend only on some values on the tail of the distribution, values which could be physically irrelevant. In other words, kurtosis is not a robust indicator of non-Gaussianity and we need a more reliable measure. Note that searching for non-Gaussianity/independence in order to solve the ICA problem, we are using statistical methods which go beyond classical second-order methods [12]; we have seen that kurtosis is related to the fourth-order moment. The need for higher-order methods is even more manifest in the following approach to a robust measure of non-Gaussianity, based on entropy concepts borrowed from information theory.

The entropy of a random variable can be thought of as the associated amount of information. The larger the entropy, the more random or unstructured the variable is. For a discrete random variable y, the entropy H is defined according to Shannon as

$$H = -\sum_i p(y = a_i) \log p(y = a_i), \tag{6}$$

where the a_i are the possible values of y. The generalization to continuous-valued random vectors $\mathbf{y}$ is

$$H(\mathbf{y}) = -\int f(\mathbf{y}) \log f(\mathbf{y}) d\mathbf{y}, \tag{7}$$

where $f(\mathbf{y})$ is the related first-order probability density. A fundamental result in information theory states that, among all random variables of equal variances, the Gaussian distribution has the largest entropy. This means that the Gaussian distribution is the most random, unstructured of all distributions. This leads to the definition of negentropy $J(\mathbf{y})$

$$J(\mathbf{y}) = H(\mathbf{y}_{\mathrm{Gauss}}) - H(\mathbf{y}), \tag{8}$$

where $\mathbf{y}_{\mathrm{Gauss}}$ is a Gaussian variable having the same covariance matrix (or variance if we deal with scalar random variables) of $\mathbf{y}$. Negentropy is always non-negative and is zero if and only if $\mathbf{y}$ has a Gaussian distribution. Thus, negentropy is a very good estimator of non-Gaussianity; the only problem with it is that using negentropy one would need to estimate the PDF via suitable approximations.

One way to approximate negentropy is via higher-order momenta, in fact it can be shown that

$$J(y) \approx \frac{1}{12} E\{y^3\}^2 + \frac{1}{48} \mathrm{kurt}(y)^2. \tag{9}$$

This approximation suffers anyway from the nonrobustness of kurtosis due to outliers. In [8] it is shown that other approximations are possible, in particular

$$J(y) \approx \sum_i^p k_i [E\{G_i(y)\} - E\{G_i(\nu)\}]^2, \tag{10}$$

where k_i are some constants, ν is a standardized Gaussian variable and the G_i are nonquadratic functions. The important thing to note is that although Eq. (10) may not be so good as a negentropy approximant, the expression in the right-hand side is still a good measure of non-Gaussianity. One can simplify further Eq. (10), and keep only one nonquadratic function G

$$J(y) \propto [E\{G(y)\} - E\{G(\nu)\}]^2, \tag{11}$$

the requests on G being very mild. Taking $G(y) = y^4$ one has the kurtosis-based approximation. The choices

$$G(t) = \frac{1}{a} \log \cosh \quad at, \qquad G(t) = -\exp(-t^2/2) \tag{12}$$

have been experimentally proven to be [8] effective. This approximation of negentropy is a good compromise between the two classical measures of non-Gaussianity, i.e. kurtosis and negentropy. Moreover, the previous expression is easy to compute, is robust and thus is a practical contrast function. It is the one used in the FASTICA algorithm, see Appendix A.

Another approach, also inspired by information theory, is via minimization of mutual information. If we have m scalar random variables y_i, their mutual information is defined as follows

$$I(y_1, \ldots, y_m) = \sum_i^m H(y_i) - H(\mathbf{y}). \tag{13}$$

Note that $H(\mathbf{y})$ contains the joint density $f(\mathbf{y})$ while $H(y_i)$ contains the i-th marginal density. It is clear that mutual information is a natural measure of dependence among the variables, being zero only if the variables are statistically independent and otherwise always non-negative. If the y_i are (zero-mean, unit variance) uncorrelated variables, one can show that mutual information and negentropy may differ only by a constant and by a sign

$$I(y_1, \ldots, y_m) = \mathrm{const} - \sum_i J(y_i). \tag{14}$$

Thus finding an invertible transformation $\mathbf{W}$ that minimizes the mutual information is equivalent to finding directions in which the negentropy is maximized. More precisely, solving ICA by minimization of mutual information is equivalent to maximizing the sum of the non-Gaussianities of the estimates, when the estimates are constrained to be uncorrelated.

B. Under-determined BSS/ICA

We have up to now made the hypothesis of having the same number (n) of sources and sensors. In many practical situations, and indeed in the case of interferometric detectors for gravitational waves, this is not the case. Classical BSS/ICA methods can-not be directly applied to this situation and one has to face the so-called under-determined case, which is much more difficult to solve. In fact, in the determined case the lack of *a priori* knowledge about the mixing process is compensated by the statistical assumption that the original components are independent, an assumption which is often physically reasonable. In the under-determined case, this is not enough; separation is still possible but one has to make further assumptions about the sources, for example, considering sparse sources, etc. (see [13] for applications of the under-determined case to audio signals) [14]. Note that even if one could know the mixing matrix **A**, the separation process in the strict sense is not obvious, since **A** is a rectangular matrix and thus is not invertible.

We circumvent this problem with the help of the Wigner-Ville transform. As we will see, the joint use of BSS/ICA and the WVT enhances the localization properties of the plain WVT.

III. THE WVT AND THE CROSS-WVT

Given an (analytical) signal $x(t)$, the Wigner-Ville transform (see [15] for more details) is a time-frequency distribution defined as

$$W_x(t,\nu) = \int x(t+\tau/2)x^*(t-\tau/2)e^{-2\pi\iota\nu t}d\tau. \qquad (15)$$

For comparison, we remind also the definition of the short-time Fourier transform

$$F_x(t,\nu;h) = \int x(u)h^*(u-t)e^{-2\pi\iota\nu t}du \qquad (16)$$

where $h(t)$ is a short-time analysis window. Note first of all that the WVT is a quadratic distribution (in the signal), whereas the short-time Fourier transform is a linear one. The WVT belongs to the so-called Cohen's class of time-frequency distributions (i.e. those which are covariant by translations in time and frequency) and has some interesting mathematical and physical properties (unitarity, etc.). Like the spectrogram, it preserves the energy condition, i.e.

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} W(t,\nu)dtd\nu = E_x = \int |x(t)|^2 dt \qquad (17)$$

The WVT is real-valued and has the property of perfect localization for linear chirp signals

$$x(t) = e^{2\pi\iota\nu(t)t}, \qquad \nu(t) = \nu_0 + 2\beta t \qquad (18)$$

$$\Rightarrow W_x(t,\nu) = \delta(\nu - (\nu_0 + \beta t)) \qquad (19)$$

being the only distribution in the Cohen's class with this property. Note that this result is true only for chirp signals with a linear frequency law. For example, for power-law chirps, the best identification of the frequency line is achieved by the Bertrand distribution (see [16]), which belongs to the affine time-frequency distributions. In this paper, we use the WVT for its simplicity.

The WVT given by Eq. (15) is also known as an auto-WVT, since the integral is evaluated on the same copy of the signal shifted in time. It is possible also to introduce a cross Wigner-Ville transform (XWVT) between two signals $x_i(t)$ and $x_j(t)$

$$W_{x_ix_j}(t,\nu) = \int x_i(t+\tau/2)x_j^*(t-\tau/2)e^{-2\pi\iota\nu t}d\tau. \qquad (20)$$

In general, we use the term WVT to mean both the auto-WVT and the cross-WVT, and differentiate the two only when we want to emphasize that the transform is made on a copy of the signal or another signal, respectively. The utility of this will be clear from our simulations.

Finally, in our simulations we have used for the WVT a moving window of 4096 points and the result of the transform has been normalized (equalized) by an empirical fit as obtained and described in [17]. This enhances the detection performance.

IV. SIMULATIONS

As we saw in the previous sections, standard ICA or other BSS methods allow us to separate independent components under the assumption that the number of sources and the number of sensors is the same. Now if we have n GW interferometers (each one with its own noise) and a GW signal impinging on all of them, it is evident that we have $n+1$ sources but only n sensors. Thus, determined ICA or BSS cannot be applied in a straightforward fashion. One has to use under-determined techniques. Yet, in the case of chirping signals [18], we know that the Wigner-Ville transform has optimal properties in detecting chirping waveforms. Since any BSS method separates on the basis of the independence and other factors, the idea is to use these techniques to lower the level of noise in each time series at the output of the algorithm, i.e. reduce the contamination among different processes and then use the WVT or the XWVT to enhance the detection performances. We illustrate these ideas by numerical simulations. All simulations were made in MATLAB; in particular, for the BSS algorithms, we capitalize the toolbox BSSGUI [19], which implements EFICA, WASOBI and BARBI (as well as other BSS algorithms, details for the algorithms of our interest are given in the Appendices A, B, and C), for the WVT we used the time-frequency toolbox [20] and the empirical equalization factor derived in [17]. The rest of the code was also written in MATLAB.

For the sake of clarity, let us consider a typical chirp signal s, see Figs. 2 and 3. This signal has been generated by the LAL software [21] using random parameters for the values of the masses, inclination of the orbit, etc. This means that the chirp could originate from a NS-NS system or a BH-BH system or a mixed system, it does not make any difference for our analysis. Then we consider two noise time series n_1, n_2, with MATLAB we have produced two independent realizations of white Gaussian noise. Finally, we inject the signal s into n_1, n_2 with a linear signal to noise ratio (SNR) of 12. More precisely, we identify the SNR with the deflection d as defined in [22] in relation to the matched filter

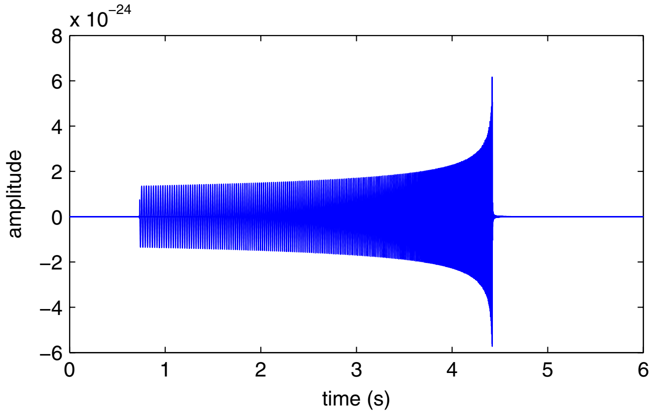

FIG. 2 (color online). Time plot of a typical chirp signal generated by LAL.

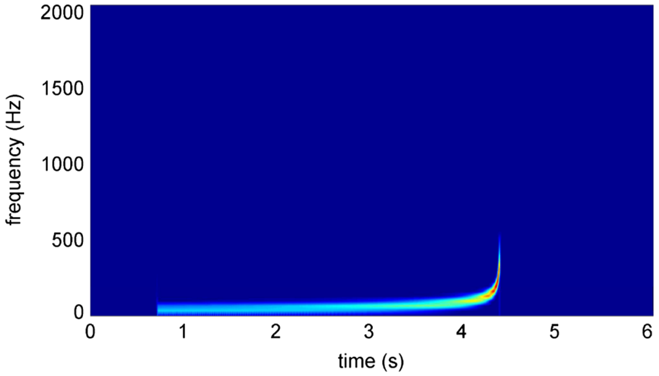

FIG. 3 (color online). Spectrogram of the chirp.

$$d = \frac{\sigma_{\mathrm{signal}}}{\sigma_{\mathrm{noise}}}\sqrt{N_{\mathrm{samples}}}. \tag{21}$$

Thus the observed data [23] $\mathbf{x} = \mathbf{A}\mathbf{s}$ take the form

$$x_1(t) = as(t) + n_1(t) \tag{22}$$

$$x_2(t) = as(t) + n_2(t) \tag{23}$$

$$\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \begin{pmatrix} a & 1 & 0 \\ a & 0 & 1 \end{pmatrix}\begin{pmatrix} s(t) \\ n_1(t) \\ n_2(t) \end{pmatrix}, \tag{24}$$

where $a = d/\sqrt{N_{\mathrm{samples}}}$, $d = 12$ is a constant which fixes the SNR and N_{samples} is the number of samples. We consider six seconds of data sampled at $f_s = 4096$ Hz. This is clearly a particular situation of under-determined BSS (with noise) where the mixing is made via the rectangular matrix

$$A = \begin{pmatrix} a & 1 & 0 \\ a & 0 & 1 \end{pmatrix}$$

(to control the SNR) rather than a random matrix. The signals s, n_1, n_2 have been normalized to unit variance before mixing.

Each time we apply a BSS algorithm to $\mathbf{x}$, we obtain two time series at the output $y_1(t)$, $y_2(t)$. For the purposes of our analysis, we distinguish two cases: the only noise case (H_0) where we do not inject the signal s (i.e. $a = 0$) and the signal plus noise case (H_1). Note that in the H_0 hypothesis, it does make sense only to apply the XWVT since there is no mixing between the components (in other words the observed data x_1, x_2 are already independent and separated). The utility of the XWVT on the noises n_1, n_2 will be clarified later. Our analysis is sketched in the flow chart in Fig. 4. Note that since BSS separates the input time series $\mathbf{x}$, the output time series $y_1(t)$, $y_2(t)$ will be as much independent as possible, so it is more convenient to run an auto-WVT on each estimated component rather than run a single cross-WVT between the two estimated components. In this way, after any BSS algorithm, we have two

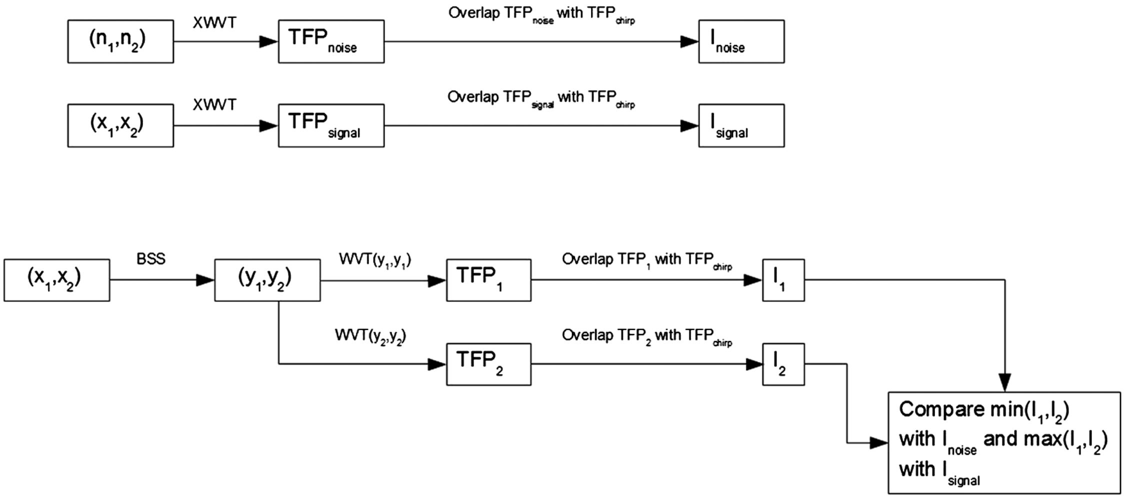
(n1,n2)
XWVT
TFPnoise
Overlap TFPnoise with TFPchirp
Inoise
(x1,x2)
XWVT
TFPsignal
Overlap TFPsignal with TFPchirp
Isignal
(x1,x2)
BSS
(y1,y2)
WVT(y1,y1)
TFP1
Overlap TFP1 with TFPchirp
I1
WVT(y2,y2)
TFP2
Overlap TFP2 with TFPchirp
I2
Compare min(I1,I2) with Inoise and max(I1,I2) with Isignal

FIG. 4. The flow chart for our analysis (see text for more details). (Top) The XWVT is applied to two time series containing noise only and signal plus noise in a direct way. (Bottom) BSS algorithms are applied to the same time series before doing a WVT.

sets of points in the time-frequency plane (TFP) to compare with the TFP obtained by the XWVT on the data **x**. In order to understand the performance improvement due to the BSS algorithms, we overlap the TFP obtained by the XWVT with the TFP of the clean chirp (see Fig. 5) and the TFPs obtained by BSS + WVT with the TFP of the clean chirp. In this way, we can count the number of points, say $\mathcal{I}$, in the TFP recovered by a specific algorithm. In other words, we integrate each TFP after XVWT or BSS + WVT along the original time-frequency line (Fig. 5). It is evident that a higher value of $\mathcal{I}$ corresponds to a better identification of the chirp frequency line and consequently to a better estimation of the physical parameters. The goal is to reduce the level of noise in the observed data **x**. To be more precise, given a BSS algorithm, only in one of the output series $y_i(t)$ (and after performing an auto-WVT), the chirp will be evident. When overlapping with the TFP of the clean chirp, and for each simulation (i.e. each generation of

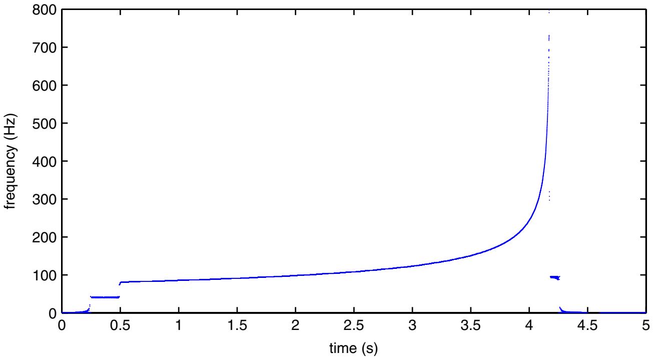

FIG. 5 (color online). Instantaneous time-frequency line of the chirp evaluated by the auto-WVT.

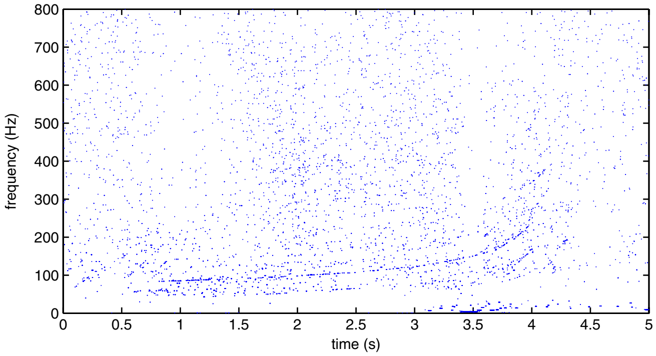

FIG. 8 (color online). Typical TFP for the EFICA + WVT case.

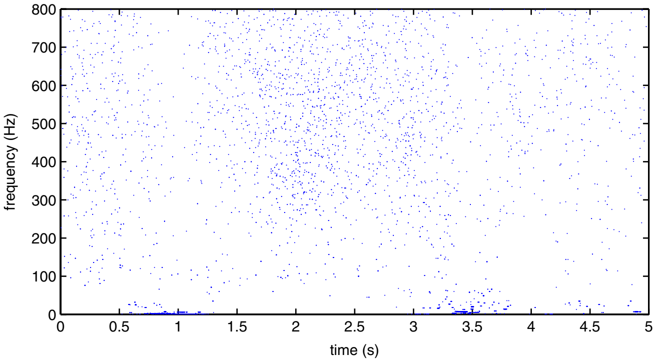

FIG. 6 (color online). Typical TFP for the XWVT in the only noise case (H_0).

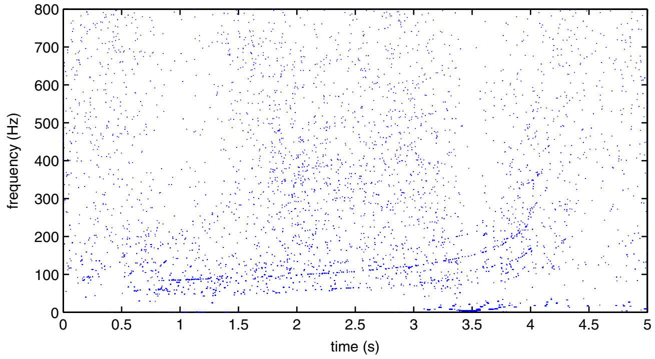

FIG. 9 (color online). Typical TFP for the WASOBI + WVT case.

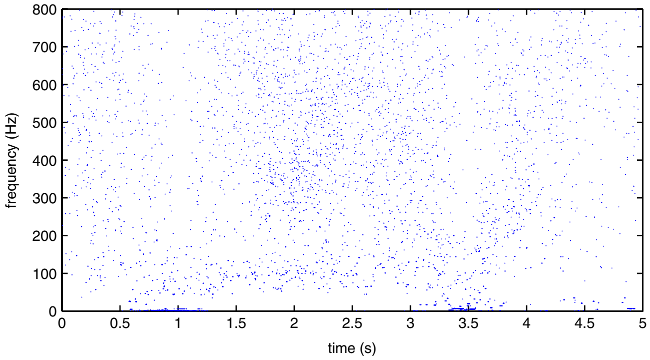

FIG. 7 (color online). Typical TFP for the XWVT in the H_1 case.

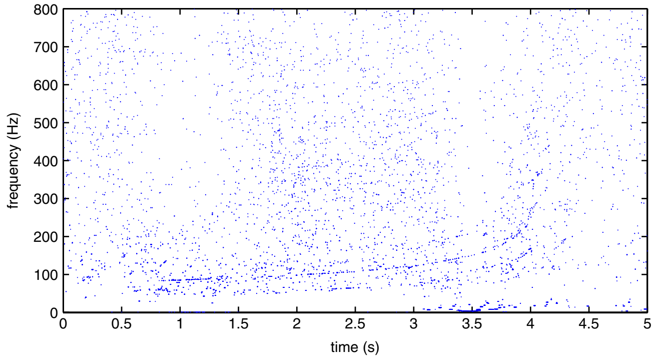

FIG. 10 (color online). Typical TFP for the BARBI + WVT case.

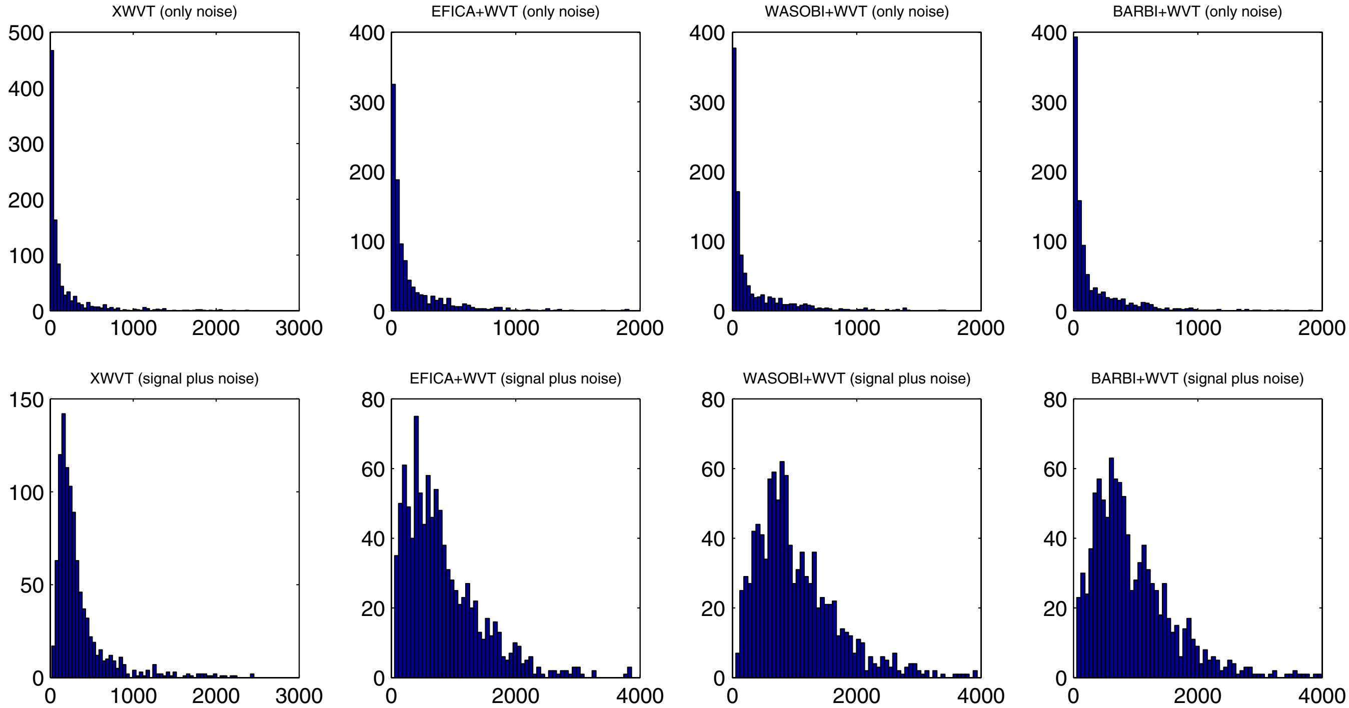

FIG. 11 (color online). Probability distribution functions for our statistical observable $\mathcal{I}$, under both H_0 and H_1 (only noise/signal plus noise). Note the presence of longer tails in the signal plus noise case analyzed by the combined algorithms BSS + WVT.

white Gaussian noise), we take the value of $\mathcal{I}$ which has the maximum value and compare this latter with the corresponding value of $\mathcal{I}$ given by the XWVT on the data $\mathbf{x}$. Finally, we compare the minimum value of $\mathcal{I}$ with the value obtained by XWVT on the noises n_1, n_2. The results of these simulations are shown in Figs. 6–10, where we show some typical TFPs obtained by the different algorithms. In Figs. 11–14, we show the

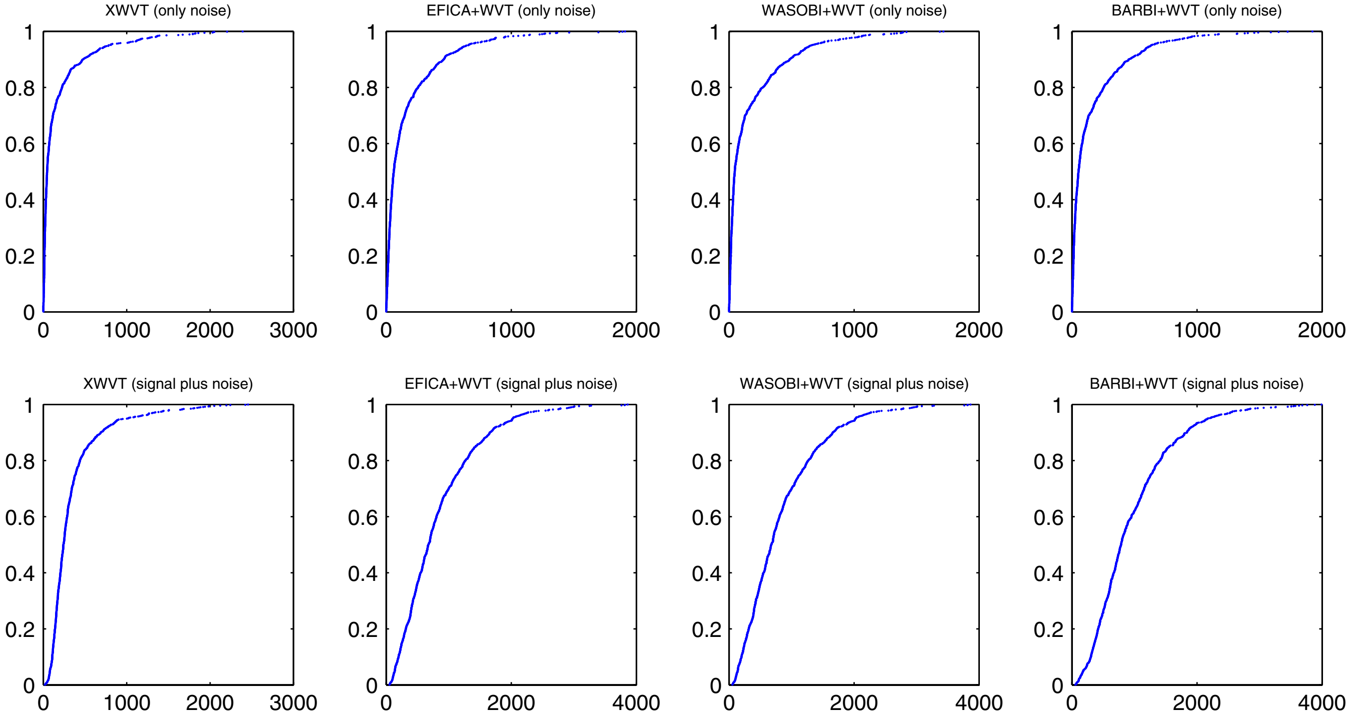

FIG. 12 (color online). Cumulative distribution functions of $\mathcal{I}$.

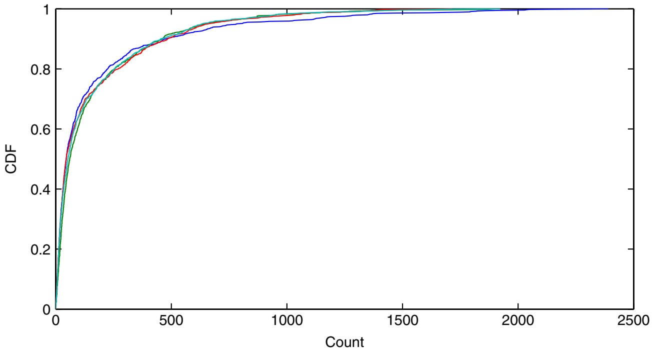

FIG. 13 (color online). CDF of $\mathcal{I}$ in the H_0 case. The curves are nearly coincident.

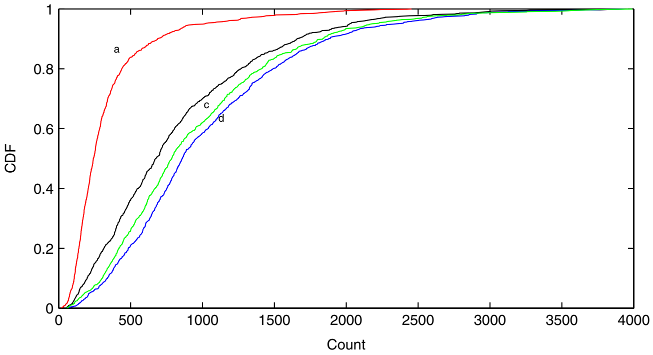

FIG. 14 (color online). CDF of $\mathcal{I}$ in the H_1 (signal + noise) case. (a) XWVT, (b) EFICA + WVT, (c) WASOBI + WVT, (d) BARBI + WVT.

PDF and the cumulative distribution function (CDF) of $\mathcal{I}$. It is interesting to note that the CDF's of the XWVT and BSS + WVT in the case H_0 are very similar, which means that the BSS algorithms do not change the statistical properties of the noise. It is also evident that the performance of the combined algorithms are better, i.e. the detection performance of the WVT is enhanced by the preprocessing with BSS.

We have also verified that by increasing the SNR the differences between the approach with the only WVT and the combined one with the BSS-WVT tend to disappear.

V. CONCLUSIONS

The present analysis can be extended to a network of N interferometers. For the moment, we neglect time delays, but this is not a serious obstacle. For each couple of GW interferometers we can apply our method. The joint use of WVT and BSS gives one TFP. For N interferometers, we can produce N independent TFPs to compare. Indeed, a suitable combination (i.e. a simple superimposition) of the N TFP plots allows to increase the detection probability of the GW chirp and to improve the parameters estimation. We are currently investigating the best combinations of the TFP functions yielding the best performance.

Note also that, although the WVT has the property to localize optimally a chirp in the TFP, for other astrophysical sources it may not be the optimal choice. On the other hand, the denoising properties of BSS algorithms are completely independent of the shape of the sought signal, and thus can be very useful as a preprocessing step for any kind of gravitational signal.

We have made the simplifying assumptions of only one signal, two independent noises and no time-delays. Because of the finite speed of GW propagation, different interferometers detect the same signal with a certain (known) time delay. Furthermore, the output of a GW detector is

$$x(t) = D^{ij}h_{ij}(t) + n(t), \tag{25}$$

where D^{ij} describes the directional responde of the antenna and h_{ij} are the transverse, traceless components of the incoming GW. Thus, for any given astrophysical source, one has two signals, h_+, $h_\times$ and a number of independent observations, each corrupted by its (independent noise), equal to the number of interferometers in the network. The problem is clearly under-determined and current methods may allow us to extract both polarizations using sparse decomposition or correlation with rough templates, etc.

Yet, if the GW direction of arrival is known, e.g., from different (optical, radio) observations, one knows the time-delays and the antenna directional response tensors. The problem of under-determined BSS with known or partly known mixing matrix is easier than the general one, and we are confident that the methods may work with the real data provided the coordinates of the source and the positions of the interferometers are known, *even if* the GW waveform is unspecified. This important point will be further discussed in a future paper.

In a forthcoming paper, we will compare the performance of a classical matched filter on the observed data $\mathbf{x}$ with the performance of the matched filter after the preprocessing with BSS algorithms [24]. This will allow us to fully characterize our method (i.e. derive a Receiving Operator Characteristic). Note that BSS algorithms *per se* are not suitable for detection, since they estimate the original components but do not allow us to set up an hypothesis test. It is only in combination with pure detection algorithms (like matched filter, time-frequency methods, etc.) that BSS techniques can be used for detection.

APPENDIX A: FASTICA

Before describing the FASTICA algorithm [25], let us remind that usually some preprocessing steps are in order before applying any BSS technique. The first one is to center the data by removing the mean $\mathbf{m} = E\{\mathbf{x}\}$; after estimating the mixing matrix one can reintroduce the mean by adding $\mathbf{A}^{-1}\mathbf{m}$. A further useful preprocessing consists in finding a linear transformation $\mathbf{x} \longrightarrow \tilde{\mathbf{x}}$ such that

$$E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^t\} = \mathbf{I}. \tag{A1}$$

The resulting data are referred to as sphered. This is always possible using, for example, an eigenvalue decomposition of the covariance matrix $E\{\mathbf{x}\mathbf{x}^t\} = \mathbf{EDE}^t$, where $\mathbf{E}$ is the orthogonal matrix of eigenvectors of $E\{\mathbf{x}\mathbf{x}^t\}$ and $\mathbf{D}$ is the diagonal matrix of its eigenvalues, $D = \mathrm{diag}(d_1, \ldots, d_n)$. Then, sphered data are given by

$$\tilde{\mathbf{x}} = \mathbf{ED}^{-1/2}\mathbf{E}^t\mathbf{x} \tag{A2}$$

and a scatter plot would indicate a spherical symmetry which explains its name. The advantage in doing this preprocessing is that

$$\tilde{\mathbf{x}} = \tilde{\mathbf{A}}\mathbf{s}, \tag{A3}$$

where the mixing matrix $\tilde{\mathbf{A}}$ is now orthogonal, thus we are left with the problem to find an orthogonal demixing matrix, which has less free parameters than the original one.

Let us now describe the popular FASTICA algorithm. It is based on a learning rule that finds a direction $\mathbf{w}$ such that non-Gaussianity of $\mathbf{w}^t\mathbf{x}$ is maximized; non-Gaussianity is measured by negentropy, and in practice by one of its approximation, see Eq. (11) above. Since one assumes that all the original components have unit variance and the data $\mathbf{x}$ have been sphered, $\mathbf{w}$ must have unit norm. The main iteration of FASTICA are the following:

(i) take an initial vector $\mathbf{w}$, for example, random
(ii) put $\mathbf{w}^\dagger = E\{\mathbf{x}g(\mathbf{w}^t\mathbf{x})\} - E\{g'(\mathbf{w}^t\mathbf{x})\}\mathbf{w}$
(iii) $\mathbf{w} = \mathbf{w}^\dagger \backslash \parallel \mathbf{w}^\dagger \parallel$
(iv) continue until convergence

where g, g' are the first and the second derivative of the G function in Eq. (11). The convergence is reached when successive approximations to $\mathbf{w}$ become aligned, i.e. when the scalar product between an old value of $\mathbf{w}$ and an updated value of $\mathbf{w}$ is ± 1.

The algorithm described just finds one direction, i.e. one component; it is a one-unit algorithm. To estimate more independent components, one needs to run the algorithms many times with different weight vectors (units) $\mathbf{w}_1, \ldots, \mathbf{w}_n$ and one needs, obviously, to prevent different weight vectors converging to the same maxima of the contrast function, that is converging to the same independent component. In order to do that, one must decorrelate the quantitities $\mathbf{w}_1^t\mathbf{x}, \ldots, \mathbf{w}_n^t\mathbf{x}$ after each iteration. A simple way to achieve this is to use the Gram-Schmidt-like algorithm. One estimates the components one by one (deflation scheme). Once p components have been estimated, that is p weight vectors, one runs the algorithm for $\mathbf{w}_{p+1}$, subtracts from $\mathbf{w}_{p+1}$ the projections $\mathbf{w}_{p+1}^t\mathbf{w}_j\mathbf{w}_j$ of the previously estimated p vectors and then renormalizes $\mathbf{w}_{p+1}$

- $\mathbf{w}_{p+1} = \mathbf{w}_{p+1} - \sum_{j=1}^{p} \mathbf{w}_{p+1}^t\mathbf{w}_j\mathbf{w}_j$
- $\mathbf{w}_{p+1} = \mathbf{w}_{p+1} \backslash \sqrt{\mathbf{w}_{p+1}^t\mathbf{w}_{p+1}}$.

Another possibility is to use a symmetric scheme, in which no privileged weight vectors exist; in this scheme the components are estimated all together via, for example, a symmetric decorrelation, see [8]. Both schemes (deflation and symmetric) are implemented in the FASTICA algorithm as well as the choice of alternative nonlinearity g functions.

In our simulations, we have used an improved version of FASTICA known as EFICA developed in [26,27], which implements an adaptive choice of the nonlinearity function g and refines the estimate of the demixing matrix.

APPENDIX B: SPECTRAL DIVERSITY

As we have mentioned, ICA is a way to solve the blind source separation problem: it is a method that works in the time domain, based on the non-Gaussianity of the components one wants to recover and uses higher-order statistical momenta. On the other hand, BSS can be solved using different algorithms which take into account frequency information or spectral properties [28]. In this and the following section, we describe two such methods which are suitable for separating signals with different spectral contents or with nonstationary properties.

A classical BSS algorithm which allows to separate components with different spectra is known as SOBI (second-order blind identification) [29]. For our convenience, we have used a refinement of this algorithm known as WASOBI (weight adjusted SOBI), described in [30], which has shown to be asymptotically optimal for Gaussian autoregressive (AR) processes. As usual, one considers the determined, noiseless model

$$\mathbf{x}[n] = \mathbf{As}[n], \tag{B1}$$

where $n = 1, \ldots, N$ is a discrete time index and $\mathbf{A}$ is a $d \times d$ mixing matrix. We assume that there are d sources which are Gaussian AR processes of known order and let $p_{\max}$ be the maximal AR order of the original sources [31]. One can then show that the quantity

$$\frac{1}{2}(\hat{\mathbf{R}}_{\mathbf{x}}[\tau] + \hat{\mathbf{R}}_{\mathbf{x}}^t[\tau]) \tag{B2}$$

forms a sufficient statistics for the separation of the AR sources ($\tau = 0, \ldots, p_{\max}$). The correlation matrices are estimated via

$$\hat{\mathbf{R}}_{\mathbf{x}}[\tau] = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}[n]\mathbf{x}[n+\tau]. \tag{B3}$$

For more details see [30].

APPENDIX C: NON-STATIONARITY

An algorithm aimed to use both spectral diversity and nonstationarity has been proposed in [32], and it is known as BARBI (block auto-regressive blind identification). It is asymptotically optimal when the sources are piecewise stationary AR processes. One assumes that the received signals can be divided into M blocks, for simplicity of equal length, and that the sources are Gaussian AR sources of order less or equal than a maximum value $p_{\max}$. If N is the data length, $N_B = N/M$ the length of each block and $L = p_{\max} + 1$, one considers the following $M \cdot L$ matrices

$$\hat{\mathbf{R}}_{ml} = \frac{1}{2N_B}\{\mathbf{X}^{(m,0)}[\mathbf{X}^{(m,l)}]^t + \mathbf{X}^{(m,l)}[\mathbf{X}^{(m,0)}]^t\} \qquad \text{(C1)}$$

where $m = 1, \ldots, M$ (it is the block index) and $l = 0, \ldots, p_{\max}$. Finally the matrix

$$\mathbf{X}^{(m,l)} = [\mathbf{X}_{:,(m-1)N_B+l+1}, \ldots, \mathbf{X}_{:,mN_B+l}] \qquad \text{(C2)}$$

is the m-th signal block of the observed data, shifted to the right by l samples. The separation procedure is done by evaluating a demixing matrix $\hat{\mathbf{V}}$ such that the matrices $\hat{\mathbf{V}}\hat{\mathbf{R}}_{ml}\hat{\mathbf{V}}^t$ are all roughly diagonal. This is achieved by an approximate joint diagonalization of the matrices $\hat{\mathbf{R}}_{ml}$, see [32]. Note also that the unknown mixing matrix $\mathbf{A}$ is assumed to be the same in each block, i.e. at each instant of time. This is reasonable for short duration signals.

[1] F. Acernese *et al.*, Classical Quantum Gravity **25**, 114045 (2008).

[2] B. Abbott *et al.*, Rep. Prog. Phys. **72**, 076901 (2009).

[3] V. Kalogera *et al.*, Astrophys. J. **601**, L179 (2004).

[4] J. Abadie *et al.*, Classical Quantum Gravity **27**, 173001 (2010).

[5] M. Maggiore, *Gravitational Waves: Theory and Experiments* (OUP, Oxford, 2007), Vol. 1.

[6] One can have situations where one has to separate dependent components, etc.

[7] In the processing of acoustic signals, complications arise due to time delays, echoes, etc. Also in the case of GW interferometers we have time delays; in this paper, which represents a first study of BSS for GW data analysis, we neglect time delays.

[8] J. K. A. Hyvarinen and E. Oja, *Independent Component Analysis* (Wiley-Interscience, New York, 1998).

[9] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing* (Wiley, New York, 1998).

[10] The separation problem is easier if some information of the statistical properties of the sought components is available. This could be useful for detecting the stochastic GW background whose statistical properties are fiducially known.

[11] This is different from classical principal component analysis, where the directions w along which to project the data x are given by the eigenvectors of the covariance matrix $E\{\mathbf{x}\mathbf{x}^t\}$. The ordering of the components $s_i = \mathbf{w}_i^t\mathbf{x}$ is done according to the largest eigenvalues of the covariance matrix. Principal component analysis is useful in reducing the dimension of the data (and the level of noise), and it can be used as a preprocessing tool for ICA especially when one has to deal with too many time series.

[12] This is necessary if we want to identify non-Gaussian variables, since any Gaussian distribution can be identified by only two numbers, the mean and the variance, all the other moments being zero or related by simple formulae to the first two moments.

[13] T. L. S. Makino and H. Sawada, *Blind Speech Separation* (Springer, New York, 2007).

[14] The most general situation is even more complicated, for example, one may not know *a priori* the number of independent components underlying the physical process and the mixing could not be instantaneous (convolutive mixtures).

[15] P. Flandrin, *Time-Frequency/Time-Scale Analysis* (Academic Press, New York, 1998).

[16] E. Chassande-Mottin and P. Fladrin, Appl. Comput. Harmon. Anal. **6**, 252 (1999).

[17] R. P. Croce, Ph. D. thesis, Universita del Sannio, 2002.

[18] The idea is that the full process should be a preprocessing step for more refined, and less heavy, grid search based on templates, in other words a hierarchical search. The use of the WVT or XWVT (with a defined threshold for an excess of power in the time-frequency plane) is optimal for chirp hunting, but it is still independent of the parameters of the chirp in the sense that it does not make any use of the templates (so one doesnot have to worry about the post-Newtonian order, the spin contribution, etc., but only assuming a signal whose frequency grows in time).

[19] http://bssgui.wz.cz/.

[20] http://tftb.nongnu.org/.

[21] https://www.lsc-group.phys.uwm.edu/daswg/.

[22] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection* (Prentice Hall, Englewood Cliffs, NJ, 1998), Vol. 2.

[23] We have included the noise terms among the sources rather than consider the noisy-BSS problem with an explicit additive term.

[24] More precisely, we will inject a controlled chirp (with known parameters) at the same SNR, compare the correlation of the processed time series with our template with the correlation of the classical matched filter and estimate the physical parameters with their errors.

[25] http://www.cis.hut.fi/projects/ica/fastica/.

[26] P. Tichavsky, Z. Koldovsky and E. Oja, IEEE Transactions on Neural Networks **17**, 1265 (2006).

[27] P. Tichavsky, Z. Koldovsky and E. Oja, IEEE Trans. Signal Process. **54**, 1189 (2006).

[28] These methods can separate also Gaussian components, although for our purposes (identifying a chirp GW signal) the Gaussianity/non-Gaussianity request is not much important.

[29] J.-F. Cardoso, A. Belouchrani, K. Abed-Meraim, and E. Moulines, IEEE Trans. Signal Process. **45**, 434 (1997).

[30] A. Y. P. Tichavsk, E. Doron, and J. Nielsen, in *EUSIPCO 2006, Florence, Italy* (2006).

[31] Note that this last assumption is useful in deriving the asymptotical optimality, but the usage of algorithm is not limited to Gaussian sources.

[32] A. Y. P. Tichavsky and Z. Koldovsky, in *ICASSP 2009, Taipei, Taiwan* (2009), 3133.