LIGO-P1100103
AEI-2011-060

arXiv:1109.0442v1 [physics.data-an] 2 Sep 2011

# A Student-t based filter for robust signal detection

Christian Röver[*]
*Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut)*
*and Leibniz Universität Hannover, 30167 Hannover, Germany*

The search for gravitational-wave signals in detector data is often hampered by the fact that many data analysis methods are based on the theory of stationary Gaussian noise, while actual measurement data frequently exhibit clear departures from these assumptions. Deriving methods from models more closely reflecting the data's properties promises to yield more sensitive procedures. The commonly used *matched filter* is such a detection method that may be derived via a Gaussian model. In this paper we propose a generalized matched filtering technique based on a *Student-t distribution* that is able to account for heavier-tailed noise and is robust against outliers in the data. On the technical side, it generalizes the matched-filter's least-squares method to an iterative, or adaptive, variation. In a simplified Monte Carlo study we show that when applied to simulated signals buried in actual interferometer noise it leads to a higher detection rate than the usual (Gaussian) matched filter.

PACS numbers: 02.50.-r, 04.80.Nn, 05.45.Tp, 95.75.Wx

## I. INTRODUCTION

Since the existence of gravitational radiation has been established as a consequence from general relativity theory, a great amount of effort has gone into the development of instruments and methods to detect gravitational waves directly [1, 2]. Gravitational waves (GWs) are notoriously weak compared to the sources of noise in today's ground-based gravitational wave detectors, and so it takes both extraordinarily sensitive instruments as well as sophisticated data analysis techniques to measure them. The output of an interferometric GW detector is essentially a time series of non-white noise, and — potentially — a superimposed signal whose exact waveform is determined by several parameters. Data analysis aiming for GW detection hence requires filtering of time-series data for rare, weak signals that are often of a known, parameterized shape. Many commonly used approaches are based on *matched filtering* the data. The matched filter may be derived as a maximum-likelihood (ML) detection method in the framework of a Gaussian noise model, but more generally will actually be ML procedure for a wider class of models. While the method works remarkably well and is able to discriminate weak signals from the noise, it commonly runs into problems due to non-Gaussian or non-stationary behaviour of the actual instrument noise. For example, the matched filter often is sensitive to outliers or loud transient noise events in the data, which, although showing little similarity with the signal sought for, also do not look like plain noise either. A lot of effort needs to go into identifying such false alarms.

We propose a more robust procedure that is based on a Student-$t$ distribution for the noise, as introduced in Ref. [3]. Several motivations may be used for introducing the Student-$t$ model; most obviously it exhibits "heavier tails" and non-spherical probability density contours, allowing to accommodate outliers in the noise. Alternatively, the model may also be seen as incorporating imperfect prior knowledge of the noise spectrum, either because it is only estimated to limited accuracy, or because it is varying over time. Models of this kind are commonly used for robust *parameter estimation*, but, as we will show in the following, the model also exhibits a better performance for *detection* purposes when the assumption of stationary Gaussian noise is violated. We expect the proposed filtering method to be useful in other signal processing contexts as well.

In the following section II we will first derive the usual matched filter from a Gaussian noise model. In section III we introduce the Student-$t$ model, elaborate on the motivation for its use as well as point out the differences to the Gaussian model, and derive the analogous filtering procedure. In section IV we report on a case study using real detector data and simulated signals to show that here the Student-$t$ based filter indeed yields a better detection rate. We close with some concluding remarks.

## II. GAUSSIAN MATCHED-FILTERING

### A. General

A *matched filter* may be derived in different ways, for example based on considerations of the residual sum-of-squares (or *power*) decomposition, without reference to a more specific noise model [4]; however, here we will concentrate on a derivation via the assumption of stationary Gaussian noise and the Whittle likelihood. This will allow to easily generalize the usual matched filtering method to the case of Student-$t$ distributed noise in the following.

[*] christian.roever@aei.mpg.de

### B. The Gaussian noise model

In order to implement the assumption of stationary, Gaussian noise residuals, the *Whittle likelihood* approximation is commonly utilized [3, 5, 6]. In the Whittle approximation, signal and noise time series are treated in their Fourier-domain representation. The explicit assumption being made on the noise $n(t)$ is that its discrete Fourier transform $\tilde{n}(f)$ is independently Gaussian distributed with zero mean and variance proportional to the power spectral density (PSD):

$$\mathrm{Var}\Big(\mathrm{Re}\big(\tilde{n}(f_j)\big)\Big) = \mathrm{Var}\Big(\mathrm{Im}\big(\tilde{n}(f_j)\big)\Big) = \tfrac{N}{4\Delta_t} S_1(f_j), \quad (1)$$

where $f_j$ is the $j$th Fourier frequency, $S_1(f_j)$ is the corresponding 1-sided power spectral density, and $j = 0, \ldots, N/2$ indexes the Fourier frequency bins. An explicit definition of the Fourier transform conventions used here is given in the appendix.

For some measured data $d(t)$ one then commonly assumes a parameterized signal $s_\theta(t)$ with parameter vector $\theta$ and additive Gaussian noise with a known 1-sided power spectral density $S_1(f)$:

$$d(t) = s_\theta(t) + n(t) \quad \Leftrightarrow \quad \tilde{d}(f) = \tilde{s}_\theta(f) + \tilde{n}(f) \qquad (2)$$

(i.e., additivity holds in both time and Fourier domains). The corresponding likelihood function then is given by

$$p\big(d|\theta\big) \propto \exp\left(-\tfrac{1}{2}\sum_j \frac{|\tilde{d}(f_j) - \tilde{s}_\theta(f_j)|^2}{\frac{N}{4\Delta_t}\, S_1(f_j)}\right) \qquad (3)$$

[3].

### C. Likelihood maximisation

#### 1. ML detection and the profile likelihood

If there were no unknown signal parameters to the signal model (like time-of-arrival, amplitude, phase,...), then, according to the *Neyman-Pearson lemma* [7], the optimal detection statistic would be the likelihood ratio between the “signal” and “no-signal” models. Once there are unknowns in the signal model, a common approach is to use a *generalized Neyman-Pearson test statistic*, that is, the *maximized* likelihood ratio, where maximization is carried out over the unknown parameters [7]. While this is in general not an optimal detection statistic, this ad hoc approach is often efficient and effective. Such a maximum likelihood (ML) detection approach is closely related to ML estimation, as either way the parameter values maximizing the likelihood will need to be derived. In case of a Gaussian noise model as in (3), maximization of the likelihood is equivalent to minimizing a weighted sum-of-squares, i.e., a weighted least-squares approach.

It should be noted that in a Bayesian reasoning framework, the detection problem would be approached via the *marginal likelihood* rather than the maximized likelihood [8, 9]. The marginal likelihood is the expectation of the likelihood function with respect to the prior distribution, and both marginal and maximized likelihood may be equivalent for a certain choice of the prior distribution. One can show the marginal likelihood to be optimal for any particular choice of prior distribution, while the maximized likelihood in general is not (see e.g. [10, 11]). Maximization of the likelihood on the other hand is commonly much easier computationally.

As will be seen in the following, it is often convenient to divide the parameter vector into subsets, as it may be possible to analytically maximize the likelihood for fixed values of some parameters over the remaining parameters. This maximized *conditional* likelihood as a function of a subset of parameters is also called the *profile likelihood*. If a profile likelihood is given, likelihood maximization may be reduced to maximizing over the remaining lower-dimensional parameter subspace. As an example, consider a signal having 3 free parameters: amplitude, phase and time of arrival. If likelihood maximization can be done analytically over amplitude and phase for any given arrival time, this results in a profile likelihood that is a function of time. The likelihood's overall maximum then may be computed via a numerical brute-force search of the profile likelihood over the time parameter.

#### 2. Why care about linear models?

In signal processing in general, and in GW data analysis in particular, the signals of interest are commonly parameterized (among other additional parameters) in terms of an amplitude and a phase parameter. Consider e.g. a simple sinusoidal signal of the form

$$\begin{aligned} s_{A,\phi,f}(t) &= A\,\sin(2\pi f t + \phi) &(4)\\ &= \beta_{\mathrm{s}}\,\sin(2\pi f t) + \beta_{\mathrm{c}}\,\cos(2\pi f t) &(5)\end{aligned}$$

which instead of amplitude $A$ and phase $\phi$ may equivalently be parameterized in terms of sine- and cosine-amplitudes $\beta_{\mathrm{s}}$ and $\beta_{\mathrm{c}}$. Other examples of signal models given in terms of linear combinations are the singular value decomposition approach used e.g. in [12, 13] or the transformation of antennae pattern effects into four amplitude parameters in the derivation of the $F$-statistic [14]. A linear model formulation will turn out convenient in the following, as a linear (or conditionally linear) model will allow to perform (conditional) likelihood maximization analytically.

#### 3. The general linear model

Consider a *linear* model for the data, i.e.,

$$y = X\beta + \epsilon \qquad (6)$$

where $y$ is a $N$-dimensional data vector, $X$ is a $(N \times k)$-matrix, $\beta$ is a $k$-dimensional parameter vector, and $\epsilon$ is an $N$-dimensional vector of error terms. The errors $\epsilon$ are assumed to be Gaussian distributed with mean zero and some covariance matrix $\Sigma$.

In the above signal processing context, $y$ and $\epsilon$ are the $N$-dimensional vectors of re-arranged real and imaginary parts of Fourier-domain data ($\tilde{d}$) and noise ($\tilde{n}$), the signal $s_\theta$ is given by a linear combination of the columns of a matrix $X$ according to the parameter vector $\beta$, and the noise covariance $\Sigma$ is a diagonal matrix defined through (1).

The Gaussian likelihood function is characterized by

$$p(y|\beta) \propto -(y - X\beta)' \, \Sigma^{-1} \, (y - X\beta). \tag{7}$$

In the linear model, the likelihood may be maximized analytically, and the ML estimator for the unknown parameter vector $\beta$ is given by

$$\hat{\beta} \;=\; (X'\Sigma^{-1}X)^{-1} X'\Sigma^{-1} y \tag{8}$$

[8, 15].

In the models of concern here, estimation is simplified by the fact that the noise covariance $\Sigma$ is a diagonal matrix (1) so that its inverse again is diagonal. In addition, here we add the common assumption that the vectors spanning the signal manifold, the columns of $X$, are orthogonal. A non-orthogonal basis $X$ would complicate the procedure slightly; see e.g. [14]. Under these conditions, the pivotal quantities for ML estimation and detection are

$$b_j = X'_{\cdot,j} \, \Sigma^{-1} \, y \;=\; \sum_{i=1}^{N} \frac{x_{i,j} \, y_i}{\sigma_i^2} \qquad \text{and} \tag{9}$$

$$c_j = X'_{\cdot,j} \, \Sigma^{-1} \, X_{\cdot,j} \;=\; \sum_{i=1}^{N} \frac{x_{i,j}^2}{\sigma_i^2}, \tag{10}$$

i.e., the quadratic forms, or inner products, involving the $j$th basis vector ($j$th column of $X$) with the data vector $y$, and with itself. The elements of the parameter vector's ML estimate $\hat{\beta}$ are then given by

$$\hat{\beta}_j = \frac{b_j}{c_j}, \tag{11}$$

the maximized likelihood ratio vs. the no-signal model is given by

$$\log\left(\frac{p(y|\hat{\beta})}{p(y|\vec{0})}\right) = \sum_{j=1}^{k} \frac{b_j^2}{2\,c_j}, \tag{12}$$

and the fitted values are given by

$$\hat{y} = X\hat{\beta} \;=\; \sum_{j=1}^{k} \hat{\beta}_j X_{\cdot,j} \;=\; \sum_{j=1}^{k} \frac{b_j}{c_j} X_{\cdot,j}. \tag{13}$$

### *4. The detection statistic and its distribution*

We define the statistic

$$H_k \;=\; \sum_{j=1}^{k} \frac{\left(\sum_{i=1}^{N} \frac{x_{i,j}\, y_i}{\sigma_i^2}\right)^2}{\sum_{i=1}^{N} \frac{x_{i,j}^2}{\sigma_i^2}} \;=\; 2 \times \log\left(\frac{p(y|\hat{\beta})}{p(y|\vec{0})}\right) \tag{14}$$

(see also (12)) which, under the null hypothesis of the data $y$ being purely noise, is $\chi^2$-distributed with $k$ degrees of freedom. Under the signal hypothesis, when a signal $s_{\beta^\star} = X\beta^\star$ is present in the data, the corresponding figure *evaluated at the true parameter values* $\beta^\star$,

$$2 \times \log\left(\frac{p(y|\beta^\star)}{p(y|\vec{0})}\right), \tag{15}$$

will be Gaussian distributed with mean $\varrho^2$ and variance $4\varrho^2$, where

$$\varrho^2 = \sum_{i=1}^{N} \frac{\left(\sum_{j=1}^{k} \beta_j^\star x_{i,j}\right)^2}{\sigma_i^2} \;=\; \sum_{i=1}^{N} \frac{(X\beta^\star)_i^2}{\mathrm{E}\left[\epsilon_i^2\right]} \tag{16}$$

$$= (X\beta^\star)' \, \Sigma^{-1} \, (X\beta^\star) \tag{17}$$

is the true signal's *signal to noise ratio (SNR)*. Consequently, for a signal of given SNR $\varrho^2$, the expected logarithmic likelihood ratio evaluated at the true parameters is $\mathrm{E}\Big[\log\big(\frac{p(y|\beta^\star)}{p(y|\vec{0})}\big)\Big] \;=\; \frac{1}{2}\varrho^2$, while the likelihood ratio $\frac{p(y|\beta^\star)}{p(y|\vec{0})}$ follows a log-Normal distribution with median $\exp(\frac{1}{2}\varrho^2)$ and expectation $\mathrm{E}\Big[\frac{p(y|\beta^\star)}{p(y|\vec{0})}\Big] \;=\; \exp(\varrho^2)$. The *maximized* likelihood ratio will be larger than that; the statistic $H_k$ follows a noncentral $\chi_k^2(\varrho^2)$-distribution with noncentrality parameter $\varrho^2$, its expectation is $\varrho^2 + k$ , so that $\mathrm{E}\Big[\log\big(\frac{p(y|\hat{\beta})}{p(y|\vec{0})}\big)\Big] \;=\; \frac{1}{2}\big(\varrho^2 + k\big)$. Note that the GW and signal processing literature is sometimes confusing, as both $\varrho^2$ and $H_k$, or their square roots, are commonly referred to as the *SNR*.

In common signal detection problems, the signal model is usually only *partially* linear, as suggested in Sec. II C 2, so that analytical maximization over the "linear" parameters only yields a maximized *conditional* likelihood, or profile likelihood. The statistic $H_k$ then is proportional to the profile likelihood, and (since the likelihood under the "noise only" null hypothesis, $p(y|\vec{0})$, is a constant) constitutes a generalized Neyman-Pearson test statistic. This statistic, or its maximum over additional parameters, is commonly referred to as a *detection statistic*, as it is used to find the signal fitting the data best, and to determine its significance. The detection statistic's distributions under null and alternative hypotheses as stated above only apply for a single (conditional) likelihood maximization, i.e., for a given data set $y$ and a given model matrix $X$. When maximizing the profile likelihood over additional parameters (or pieces of data), the testing problem turns in to a *multiple testing* problem, and the statistic's distribution will be an an extreme value statistic [7, 16]. Since the particular statistic $H_k$ only comes up in the context of the Gaussian model, we will in the following be mostly referring to the more universal corresponding likelihood ratio figure $\frac{p(y|\hat{\vec{\beta}})}{p(y|\vec{0})} = \exp(\frac{1}{2} H_k)$.

### D. Common implementation and terminology

In the GW data analysis literature, likelihoods and matched filters are commonly expressed in terms of the *inner product* $\langle a, b\rangle$ of real-valued functions (signal templates or data) $a$ and $b$, technically defined in terms of analytical Fourier transforms:

$$\langle a, b\rangle = \int_{-\infty}^{\infty} \frac{\tilde{a}(f)\,\tilde{b}(f)^*}{S_1(f)}\,\mathrm{d}f \tag{18}$$

[6, 14], which in practice is implemented (analogously to the Whittle likelihood) in terms of discrete Fourier transforms:

$$\begin{aligned}&\langle a, b\rangle \qquad (19)\\ &= 2\sum_{j=0}^{N/2} \frac{\frac{\Delta_t}{N}\Big[\mathrm{Re}\big(\tilde{a}(f_j)\big)\mathrm{Re}\big(\tilde{b}(f_j)\big) + \mathrm{Im}\big(\tilde{a}(f_j)\big)\mathrm{Im}\big(\tilde{b}(f_j)\big)\Big]}{S_1(f_j)}.\end{aligned}$$

In terms of the linear models discussed in the previous section, this is equivalent to a quadratic form

$$\vec{a}'\,\Sigma^{-1}\,\vec{b} \tag{20}$$

as in equations (9), (10) above. Note that especially in the context of the Student-$t$ model discussed below, expression (18) may be hard to motivate, as it is continuous in frequency, but the corresponding discrete expression (19) may readily be related to expressions derived above. In this terminology, the signal-to-noise ratio of a signal $s_\theta$ (16) turns out as

$$\varrho^2 = \sum_j \frac{|\tilde{s}_\theta(f_j)|^2}{\frac{N}{4\Delta_t} S_1(f_j)} \;=\; 2\,\langle s_\theta, s_\theta\rangle, \tag{21}$$

the correlation of some data $d$ with a template $s_\theta$ (as in (9)) simplifies to

$$\begin{aligned}&\sum_j \frac{\Big[\mathrm{Re}\big(\tilde{d}(f_j)\big)\mathrm{Re}\big(\tilde{s}_\theta(f_j)\big) + \mathrm{Im}\big(\tilde{d}(f_j)\big)\mathrm{Im}\big(\tilde{s}_\theta(f_j)\big)\Big]}{\frac{N}{4\Delta_t} S_1(f_j)}\\ &= 2\,\langle d, s_\theta\rangle, \qquad (22)\end{aligned}$$

the likelihood ratio of some signal template $s$ for given data $d$ is

$$\begin{aligned}\log\Big(\frac{p(d|s_\theta)}{p(d|\vec{0})}\Big) &= \frac{\sum_j \frac{|\tilde{d}(f_j) - \tilde{s}_\theta(f_j)|^2}{S_1(f_j)}}{\sum_j \frac{|\tilde{d}(f_j)|^2}{S_1(f_j)}} \qquad (23)\\ &= 2\,\langle d, s_\theta\rangle - \langle s_\theta, s_\theta\rangle \qquad (24)\end{aligned}$$

[3, 6, 14], and the *maximized* likelihood ratio for a signal that is a linear combination of waveforms ($d = \sum_j \beta_j s_j + n$, see also (12)) then is

$$\log\Big(\frac{p(d|\hat{\beta})}{p(d|\vec{0})}\Big) = \sum_j \frac{\langle d, s_j\rangle^2}{\langle s_j, s_j\rangle}. \tag{25}$$

An implementation of a matched filter in the GW context is concisely described e.g. in [17, 18]. The signal searched for is a "chirping" binary inspiral waveform of increasing frequency and amplitude, which is characterized by 5 parameters, namely two mass parameters determining the phase/amplitude evolution, and amplitude, phase and arrival time. The signal waveform $s$ for given mass parameters $\vartheta = (m_1, m_2)$ is (in analogy to (4)) given in terms of "sine" and "cosine" components $s_{\mathrm{s},\vartheta}$ and $s_{\mathrm{c},\vartheta}$:

$$s_\vartheta(t) \;=\; \beta_{\mathrm{s}}\, s_{\mathrm{s},\vartheta}(t - t_0) + \beta_{\mathrm{c}}\, s_{\mathrm{c},\vartheta}(t - t_0) \tag{26}$$

[17] where $\beta_{\mathrm{s}}$ and $\beta_{\mathrm{c}}$ are determined by the orbital phase and orientation of the binary system, and $t_0$ defines the signal arrival time. The sine and cosine waveforms here constitute the signal manifold's orthogonal "basis vectors". The actual matched filter detection statistic is defined as $\rho(t_0) = \sqrt{X^{\mathrm{s}}(t_0)^2 + X^{\mathrm{c}}(t_0)^2}$, where

$$X^{\mathrm{s/c}}(t_0) \;\propto\; \int \frac{\tilde{d}(f)\,(\tilde{s}_{\mathrm{s/c},\vartheta}(f))^*\,\exp(-2\pi\mathrm{i} f t_0)}{S_y(|f|)}\,\mathrm{d}f \tag{27}$$

[17], and where the exponential term does the time-shifting of data and template against each other. For any given time shift $t_0$, this filter corresponds to (the square root of) the detection statistic $H_k$ above (14). Computing the matched filter (27) across time points $t_0$ yields the profile likelihood, the conditional likelihood (conditional on time $t_0$ and waveforms $s_{\mathrm{s},\vartheta}$, $s_{\mathrm{c},\vartheta}$) maximized over phase and amplitude. The "overall" maximum likelihood then is determined via a brute-force search over $t_0$ and over additional alternative signal waveforms corresponding to different mass parameters $\vartheta$. Note that the search over arrival time $t_0$ in (27) may be efficently implemented via another Fourier transform [18]. The matched filtering algorithm is also described in more detail in Appendix A 3.

In order to claim the *detection* of a signal, one needs to determine a threshold for the detection statistic (the maximized likelihood), with respect to some pre-specified false alarm rate. The detection statistic's distributions derived in Sec. II C 4 are likely not to be of much practical relevance, due to common non-Gaussian or nonstationary features in the data. Critical values for the detection statistic instead are commonly computed using bootstrapping methods (see e.g. [19, 20]).

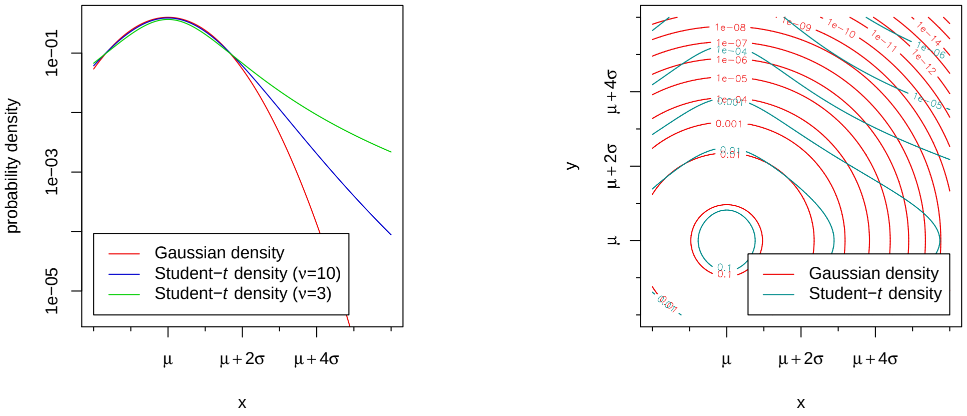

FIG. 1. Density functions of Gaussian and Student-$t$ distributions. The left panel shows univariate densities on the logarithmic scale. The right panel shows density contours of the joint distribution of two independent Gaussian random variables in contrast with two independent Student-$t$ distributed variables of the same location ($\mu$) and scale ($\sigma$). The two Student-$t$ variables have differing degrees-of-freedom; the one corresponding to the $x$-axis has $\nu = 3$, while the one along the $y$-axis has $\nu = 10$.

## III. THE STUDENT-T FILTER

### A. The Student-t noise model

The Student-$t$ model for time series analysis was introduced in [3] as a generalisation of the commonly used Gaussian model described in the previous section. The Student-$t$ distribution has an additional *degrees-of-freedom* parameter, essentially controlling the distribution's heavy-tailedness, i.e., the allowance for large outliers. The Student-$t$ likelihood function is given by

$$p(\vec{d}|\theta) \propto \prod_j \left(1 + \frac{1}{\nu_j}\,\frac{\bigl|\tilde{d}(f_j) - \tilde{s}_\theta(f_j)\bigr|^2}{\frac{N}{4\Delta_t}\,S_1(f_j)}\right)^{-\frac{\nu_j+2}{2}} \quad (28)$$

$$= \exp\left(-\sum_j \frac{\nu_j+2}{2}\log\left[1+\frac{1}{\nu_j}\,\frac{\bigl|\tilde{d}(f_j)-\tilde{s}_\theta(f_j)\bigr|^2}{\frac{N}{4\Delta_t}\,S_1(f_j)}\right]\right) \quad (29)$$

[3]. According to this model, the residuals ($\mathrm{Re}(\tilde{n}(f_j))$, $\mathrm{Im}(\tilde{n}(f_j))$) within each Fourier frequency bin $j$ follow a *bivariate Student-t distribution* [8] with location $\mu = \vec{0}$, scale matrix $\Sigma = \frac{N}{4\Delta_t}\begin{pmatrix} S_1(f_j) & 0 \\ 0 & S_1(f_j)\end{pmatrix}$, degrees-of-freedom $\nu_j > 0$ and implicit dimension 2. This implies that (i) residuals in different frequency bins are independent, (ii) residuals within the same bin are uncorrelated, but dependent, and (iii) the marginal distribution of each individual residual is a Student-$t$ distribution with scale proportional to $S_1(f_j)$ and degrees-of-freedom $\nu_j$. Decreasing values of the degrees-of-freedom parameters $\nu_j$ imply a heavier-tailed distribution, and in the limit of $\nu_j \rightarrow \infty$ the model again reduces to the Gaussian model.

Besides simply constituting a heavier-tailed noise model, the Student-$t$ model arises as a generalisation of the Gaussian model when the power spectral density $S(f_j)$ is treated as uncertain, where the degrees-of-freedom parameter $\nu_j$ denotes the (prior) precision [3]. So the model is not only applicable in contexts where the noise itself is in fact $t$-distributed, but also in cases where it is Gaussian, but the noise spectrum is a priori only known to a certain accuracy. Alternatively, the same model would result when the noise spectrum itself was assumed to be randomly deviating from the scale parameter $S_1(f)$, according to a $\chi^2$ distribution, e.g. because it is only estimated with some uncertainty, which in fact resembles the original motivation for introducing Student's $t$-distribution in the context of the $t$-test and related procedures [7, 21]. Both randomness or uncertainty of the noise PSD technically lead to the same likelihood expression here [3]. In general, the interpretation of the scale parameter $S_1(f)$ in the contexts of the Gaussian and the Student-$t$ model is not necessarily exactly the same. For the Gaussian model, it may be defined via the expected power $S_1(f_j) = \mathrm{E}\bigl[2\frac{\Delta_t}{N}\,|\tilde{n}(f_j)|^2\bigr]$, while for the Student-$t$ model this only holds in the limiting case of great certainty ($\nu \rightarrow \infty$). Within the Student-$t$ model, the $S(f)$ term specifies the scale of the uncertain PSD parameter, the expected power is in fact given by $\mathrm{E}\bigl[2\frac{\Delta_t}{N}\,|\tilde{n}(f_j)|^2\bigr] = \frac{\nu_j}{\nu_j-2}\,S_1(f_j)$. The choice of the degrees-of-freedom parameter $\nu_j$ as well as the spectrum parameter $S_1(f_j)$ may be approached in different ways and may for the filtering purpose eventually be considered a matter of tuning [3]. In the example in Sec. IV below, we simply kept the scale parameter $S_1(f_j)$ to be the estimated noise spectrum as in the Gaussian case, and fitted a common d.f. parameter $\nu_j = \nu$ for all frequency bins to the empirical data.

### B. Comparison to the Gaussian model

When comparing to the Gaussian distribution, first of all the Student-$t$ distribution exhibits *heavier tails*, i.e., the probability for obtaining "large" values (relative to the distribution's scale) is much greater. While the density functions are very similar within the range of $\mu \pm 2\sigma$, where the bulk of probability is concentrated, the densities' ratio will grow indefinitely towards the distributions' tails (see Fig. 1). The degrees-of-freedom parameter $\nu$ controls the distribution's heavy-tailedness; a setting of $\nu = 1$ yields the "pathological" Cauchy distribution, for $\nu > 2$ the variance is finite and in the limit of $\nu \to \infty$ it again approaches the Gaussian distribution.

Another discriminating feature is the shape of the density contours. While a Gaussian density will always have elliptical contours, the Student-$t$ distribution is different in that its contours are rather diamond-shaped, with elongations pointing along the principal axes (see Fig. 1). This way the Student-$t$ model does not only allow for larger outliers, but it also considers outliers more likely to occur only in individual variables rather than jointly in all variables. Note that this latter effect follows from the fact the different frequency bins are *stochastically independent* and not merely *uncorrelated* [22, 23]. Since the two (real and imaginary) residuals within each Fourier frequency bin follow a joint, bivariate, $t$-distribution, the density contours *within* bins will still be spherical—otherwise a strange phase/amplitude dependence would be implied for the Fourier-domain model. The effect of independent Student-$t$ variables only comes to bear *between* frequency bins.

An important difference to note between Gaussian and Student-$t$ model is that the least-squares fitting that results from the Gaussian model will actually be a ML procedure for any model within the wider class of "elliptically symmetric" models for the noise residuals (including e.g. a Student-$t$ model with merely *uncorrelated*, but not *independent* residuals) [22, 23]. The Student-$t$ model described here hence advances into a fundamentally different class of models.

Student-$t$ or similar models are commonly used in parameter estimation contexts as robust alternatives to the Gaussian model that are less sensitive to outliers in the data [24–27]. Such models may be motivated in a "top-down" manner by the observation that the data do not actually fit the Gaussianity assumption, or also in "bottom-up" way by pointing out that the resulting least-squares procedures are very sensitive to occasional outliers in the data. In the spirit of the latter viewpoint, the concept of *M-estimation* was introduced, which aims at "fixing" outlier-sensitive least-squares procedures by replacing them by more robust statistics corresponding to more favourable *influence functions* [28, 29]. Similar approaches, namely down-weighting or ignorance of outliers in the data, have been proposed in the context of gravitational-wave detection before [30, 31], and the Student-$t$ assumption may in fact be considered a special case of M-estimation [26, 27].

Another fix that is commonly applied in GW data analysis is the $\chi^2$ *veto* [32], which is a figure computed along with a detection statistic that is supposed to discriminate actual signals from noise bursts. Such noise events may show little similarity with the signal template, but may often, due to non-negligible correlation with the template and very large power, still seem to indicate the presence of a signal. The $\chi^2$ veto then essentially checks for excess power that is inconsistent with the shape of the signals aimed for and that way will rule out such alleged detections. The consideration of excess residual power is also implicitly happening in the Student-$t$ model. From the different likelihood formulations ((3), (29)) one can write down the corresponding likelihood ratios for some data $d$ and a signal template $s_\theta$:

$$\log\left(\frac{p(d|\theta, \text{Gauss})}{p(d|\vec{0}, \text{Gauss})}\right) = \sum_j \frac{1}{2}\left(\frac{\big|\tilde{d}(f_j)\big|^2}{\frac{N}{4\Delta_t}\, S_1(f_j)} - \frac{\big|\tilde{d}(f_j) - \tilde{s}_\theta(f_j)\big|^2}{\frac{N}{4\Delta_t}\, S_1(f_j)}\right), \qquad (30)$$

$$\log\left(\frac{p(d|\theta, \text{Student})}{p(d|\vec{0}, \text{Student})}\right) = \sum_j \tfrac{\nu_j+2}{2} \log\left(\frac{1 + \frac{1}{\nu_j}\frac{\left|\tilde{d}(f_j)\right|^2}{\frac{N}{4\Delta_t}\, S_1(f_j)}}{1 + \frac{1}{\nu_j}\frac{\left|\tilde{d}(f_j) - \tilde{s}_\theta(f_j)\right|^2}{\frac{N}{4\Delta_t}\, S_1(f_j)}}\right). \qquad (31)$$

In both of the above cases the likelihood ratio is a function of the "data power" $\frac{\left|\tilde{d}(f_j)\right|^2}{\frac{N}{4\Delta_t} S_1(f_j)}$ and the "residual power" $\frac{\left|\tilde{d}(f_j) - \tilde{s}_\theta(f_j)\right|^2}{\frac{N}{4\Delta_t} S_1(f_j)}$, i.e., the data's normalized sum-of-squares in each frequency bin $j$ before and after subtracting the signal $s_\theta$. For the Gaussian case, a "data power" of 10 and a "residual power" of 1 in the $j$th bin would have the same effect on the likelihood ratio as if the numbers were, say, 1010 and 1001 instead; the only relevant figure is their difference. In the Student-$t$ model, the latter case would lead to a lower likelihood ratio; here not only the amount by which the signal $s_\theta$ is able to reduce the sum-of-squares is relevant, but also its magnitude relative to the remaining residual term. The additional feature of the ML fit that is intrinsically considered in the Student-$t$ likelihood ratio (31) is essentially the corresponding *coefficient of determination* ($R^2$) [15]. As will become obvious in the following, when the actual implementation is described, the generalisation to the Student-$t$ model will on the technical side essentially replace the least-squares procedure by an adaptive version. The adaptation step again ensures that excess residual noise power will downweight the supposed significance of a signal.

### C. Likelihood maximization: the EM-algorithm

While likelihood maximization in the Gaussian model boils down to least-squares fitting, the maximization step is not quite as simple for the Student-$t$ model. However, due to the structure of the problem, the *Expectation-Maximization (EM) algorithm* may be used to efficiently maximize the likelihood function [8, 33]. In order to apply the EM algorithm, the likelihood expression needs to be reformulated. The Student-$t$ likelihood may be viewed as a *marginal* likelihood, averaging out a set of unknown variance parameters $\vec{\sigma}^2$ [3]. Each of the variance parameters $\sigma_j^2$ then corresponds to the power spectral density at the $j$th Fourier frequency bin. The EM algorithm's details as applied to the present problem are derived in detail in appendix A 2 below. It turns out that maximization of the Student-$t$ likelihood may be done in an iterative manner, where each iteration again requires a weighted least-squares fit as in the Gaussian matched filter. The EM algorithm requires a starting value $\theta_0$ for the signal parameters. Given $\theta_0$, the expression

$$\mathcal{E}(\theta_0, \theta) - \frac{1}{2} \sum_j \frac{\left|\tilde{d}(f_j) - \tilde{s}_\theta(f_j)\right|^2}{\frac{N}{4\Delta_t}\left(\frac{\nu_j}{\nu_j+2} S_1(f_j) + \frac{2}{\nu_j+2}\frac{2\Delta_t}{N}\left|\tilde{d}(f_j) - \tilde{s}_{\theta_0}(f_j)\right|^2\right)} \tag{32}$$

is maximized with respect to the parameter vector $\theta$. The parameter value maximizing the above expression then consitutes the new $\theta_0$ value, for which the expression again is maximized, and so forth. The resulting sequence of parameter values then converges to the maximum likelihood estimate $\hat{\theta}$ [8].

Maximizing the above expression (32) again amounts to a weighted least-squares fit, exactly as in the case of the Gaussian matched filter (see also the corresponding likelihood expression (3)). The Student-$t$ filter will therefore generalize the Gaussian matched filter by replacing the least-squares procedure by an iterative, or adaptive, least-squares fit. Note that the denominator in (32) simply is a weighted average of the noise spectrum (as in (3)) and the previous iteration's residual noise power, where the degrees-of-freedom parameter $\nu$ defines the relative weighting. Instead of the "plain" weighted least-squares match that is done in the Gaussian filter, the EM-iterations adapt the weights (the denominator in (32), which in the Gaussian model was the a priori known, fixed noise spectrum) to the residual noise power as found in the data, and the level of adaptation is regulated by the degrees-of-freedom parameter $\nu$.

The (ML) detection statistic does not follow a simple distributional form as in the Gaussian model, but in the example below one can already see that both statistics still behave similarly. The generalized likelihood ratio statistic will, by Wilks' theorem, in fact still approximately follow a $\chi^2$-distribution [7, 34].

### D. The filter implementation

As for the Gaussian matched filter, the aim again is to maximize the likelihood (29), i.e., find best-fitting parameter values $\hat{\theta}$ in parameter space. Again, it is advantageous if the signal model can (at least partly) be formulated as a linear model.

There are two obvious points in the matched-filtering procedure at which one could insert the EM-iterations in order to generalize it to the Student-$t$ case: either at the level of each (originally analytical) maximization over linear model coefficients (usually corresponding to amplitude and phase), or at a higher level, iterating over linear coefficients as well as the signal arrival time parameter. It is not obvious whether one implementation is more sensitive than the other, but there definitely are differences in the implied computational costs. Both approaches are described and discussed in more detail in appendix A 3. An implementation of the latter algorithm, together with analogous matched filter, is available in [35]. In case of a brute-force search over additional signal parameters (i.e., a "template bank"), one could in fact consider moving the EM-level yet another stage higher.

As a starting parameter value ($\theta_0$) for the algorithm, one could for example use the null vector or an initial least-squares fit. As a stopping criterion, one could terminate the algorithm once the improvement in logarithmic likelihood from the previous iteration falls below some threshold, or when some maximum number of iterations is reached. Note that—unlike for the Gaussian linear least-squares fit—the (conditional) likelihood might actually be multimodal [36], so that different starting values might lead to different results. It is not obvious whether this occurs frequently in practice, or rather requires particularly rare pathological circumstances; however, it does not appear to pose a problem in the example below.

## IV. FILTERING EXPERIMENT ON ACTUAL DATA

### A. General

Besides any theoretical or heuristic arguments why a Student-$t$ based filter may improve detection, the figure of eventual relevance is going to be the resulting improvement in detection efficiency when applied to actual data — keeping in mind the additional complication and computational cost. In the following, we will demonstrate the filter's performance in a minimalistic, yet realistic toy problem. To that end, we will set up a filter for a certain kind of parameterized signal, and then test it against a conventional matched filter using injections of simulated signals. For the additive noise, we will use both simulated Gaussian noise as well as actual gravitational-wave detector instrument noise. Detection efficiency is going to be measured via the *receiver operating characteristic (ROC)* curve, allowing to compare detection probabilities for given false alarm probabilities, or vice versa.

In order to make the example realistic, we require a nontrivial signal waveform to be searched for; in particular the waveform should not be monochromatic, but should instead span a wider range of Fourier frequencies. There should be parameters to be maximized over ana-

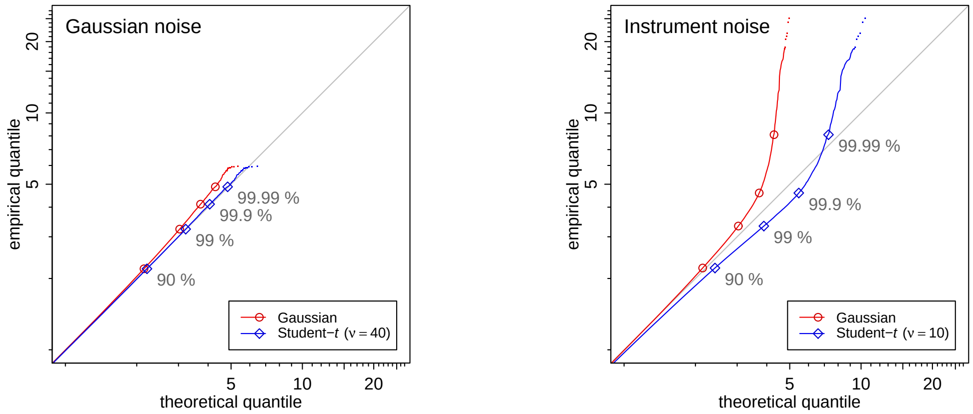

FIG. 2. Quantile-quantile plots (Q-Q plots) of the empirically found normalized residual noise power (33) versus its theoretical values assuming Gaussian and Student-$t$ models. The marks indicate particular quantiles corresponding to powers of 10 in tail probability. The 10 largest empirical samples are shown as individual dots, the remaining quantiles are connected by a line.

lytically as well as numerically, and we should use noise that is non-Gaussian or non-stationary. The example described in the following mimics the setup of a search for binary inspiral signals in interferometric gravitational-wave detector data (see e.g. [17]). The noise data are taken from an actual detector, and, for comparison, a second data set of simulated, Gaussian noise of a realistic noise spectrum is used in parallel. The "search" being performed however is much simplified and not intended to be exhaustive or to span an astrophysically sensible parameter range.

### B. The data

The data used in the following examples is going to be either simulated Gaussian noise with a power spectral density corresponding to LIGO's initial design sensitivity [37], or real instrument noise from LIGO's Livingston interferometer, taken during LIGO's 5th science run ("S5") in late 2005 [38]. The data will be considered in chunks of 8 seconds length, downsampled to a sampling rate of 1024 Hz, and windowed using a Tukey window tapering 10% of the data (5% at each end). The noise's power spectral density $S_1(f)$ is estimated essentially using Welch's method [39], by considering the empirical power in the 32 preceding data segments, and taking the median as a robust estimator. The figures shown in the following are each based on 100 000 such data chunks.

The signal waveform searched for here is taken to be a binary inspiral waveform approximated to the 2.0 post-Newtonian order [40]. The same waveform family is used for both injections as well as in the detection stage, and it has 5 free parameters: chirp mass ($m_c$), mass ratio ($\eta$), coalescence time ($t_c$), coalescence phase ($\phi_c$), and amplitude ($A$). The signal waveforms injected into the data were all done at the same mass parameters ($m_c = 4.5$, $\eta = 0.25$), and the amplitude is set such that the signal's SNR (as computed based on the current PSD estimate) is $\varrho = \sqrt{\varrho^2} = 5.257$ so that $\mathrm{E}\Big[\log\big(\frac{p(y|\beta^\star)}{p(y|\vec{0})}\big)\Big] = \frac{1}{2}\varrho^2 = \log(10^6)$ and $\mathrm{E}\Big[\frac{p(y|\beta^\star)}{p(y|\vec{0})}\Big] = \exp\big(\varrho^2\big) = 10^{12}$ (see also Sec. II C 4). Each 8-second chunk of data is eventually analyzed twice, with and without a signal injection.

### C. Setting the degrees-of-freedom parameter

In order to determine a suitable degrees-of-freedom parameter $\nu$ for the Student-$t$ model, we considered the tail behaviour of the noise. If the Gaussian ("Whittle") model was accurate, then the normalized Fourier-domain noise power at the $j$th frequency bin,

$$\frac{\big|\tilde{n}(f_j)\big|}{\sqrt{\frac{N}{4\Delta_t} S_1(f_j)}}, \tag{33}$$

being the square root of the sum of two independent standard Gaussian random variables (see Sec. II B), should follow a *Rayleigh distribution*. The residuals' normalisation here is done — in analogy to the computations done in an actual search — via the *estimated* noise spectrum, as described in the previous subsection. We are only considering the binned noise *power* here (and not the individual real and imaginary components) as this is the relevant figure entering both the Gaussian as well as the Student-$t$ likelihoods ((3), (29), (30), (31)). Under the Student-$t$ model, instead of being Rayleigh-distributed, the power (33) would instead follow a similar, more heavy-tailed distribution. We will refer to the Student-$t$ power's distribution as the "*Student-Rayleigh*" distribution here; more details on this distribution's particular form are given in appendix A 4.

We investigated the empirical distribution of actual noise residuals, for both simulated and actual instrumental data. For the simulated data, this will account for effects of finite sample size, windowing and PSD estimation, and for actual data it will in addition give some in-

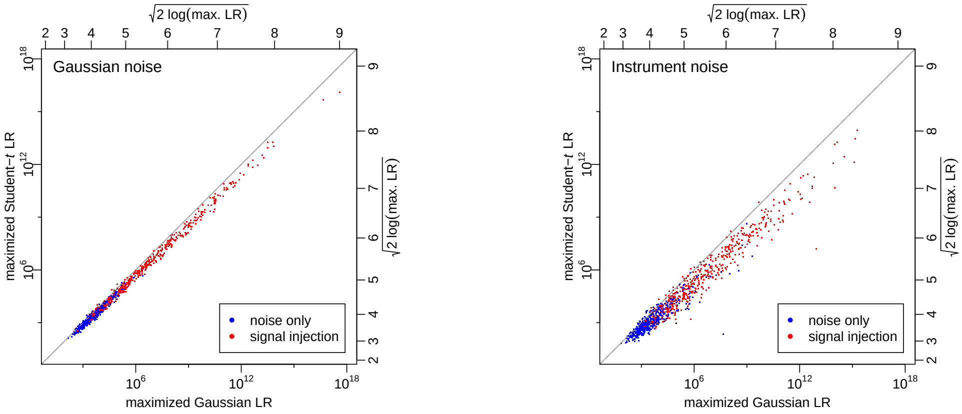

FIG. 3. Detection statistics (maximized likelihood ratios) based on Gaussian and Student-$t$ models for simulated Gaussian data (left panel) and actual interferometer noise (right panel). Injections were of SNR $\varrho = 5.257$.

sight into the effects of realistic nonstationarities or non-Gaussianities in actual measurement noise. The noise samples are based on the residuals from 200 eight-second noise realisations of either simulated Gaussian noise, or actual instrument noise from LIGO's Livingston interferometer. The residuals (33) are each normalized via a PSD estimate from 32 preceding noise samples, as described in the previous section, yielding a total of 800 000 residuals. The data used here did not overlap with the data used in the following detection experiment.

Fig. 2 shows quantile-quantile plots (Q-Q plots) illustrating how well the models fit the actual data. The axes indicate theoretical (Rayleigh or "Student-Rayleigh") quantiles, and the empirical quantiles as found in the data. If a model fits the data well, both theoretical and empirical quantiles should coincide, so that the quantiles follow a straight, diagonal line. A mismatch between model and data results in a differently shaped curve; in particular, if the data are more heavy-tailed than predicted by the model, the curve will have an upward slope [41].

One can see that the actual data exhibit heavier tails in both cases of simulated, Gaussian noise as well as the instrument noise. In the case of Gaussian noise this is due to the estimation uncertainty in the noise spectrum. If we had been using the mean instead of the median to estimate the noise PSD, then the distribution of normalized noise residuals should be *exactly* Student-$t$ with degrees of freedom equal to twice the number of noise samples averaged over (here: $32 \times 2 = 64$) [7, 21]. For the median estimation method, this is only approximately true, but apparently still roughly accurate; a maximum-likelihood fit for $\nu$ suggests a value of $\nu \approx 40$ here. For the case of Gaussian data, the mismatch between assumed and observed quantiles is minimal anyway.

For the real interferometer noise, the discrepancy between Gaussian model and actual data is more dramatic; in the distribution's tails, the empirical quantiles are significantly larger than the assumed quantiles. For example, according to the Gaussian model, 99.99 % of the samples should be $\leq 4.3$, while empirically the 99.99 % quantile lies at 8.1 for actual instrument noise (see the right panel of Fig. 2). A Student-$t$ model seems to fit the data better, especially in the distributions' tails, although discrepancies in the extreme outliers are still large. Trying to estimate the degrees-of-freedom parameter $\nu$ from different subsets of the empirical data yields ML estimates roughly in the range from 5 to 50; in the following we simply fixed the parameter at $\nu = 10$ for the simulations involving actual data. A value of $> 40$ would not seem to make sense here (even if the data were perfectly Gaussian) and in the simulation results below we found that detection performance seemed to depend only weakly on $\nu$ as long as it was roughly in the range 5–20. While the Student-$t$ distribution does not fit the data perfectly, it seems to fit better than the Gaussian model. Instead of only fitting the d.f. parameter, one could actually in addition also adapt the $t$-distribution's scale to the data (see also Sec. III A, or [3]).

### D. Filtering setup

For each piece of data, the likelihood ratio is maximized over phase and amplitude for given combinations of time and mass parameter values, where the evaluated time points were $t_c \in \{6.50, 6.55, \ldots, 7.50\}$ and the considered masses were $\eta = 0.25$, $m_c \in \{3.0, 3.1, \ldots, 6.0\}$. The injected signal's parameter values always were among the grid points maximized over, so that signal/template mismatch considerations are not of concern here. On the technical side, this is implemented in a loop over template waveforms (corresponding to different mass parameters) and time points. At each mass/time combination, computation of the conditionally maximized Gaussian likelihood ratio amounts to computing an inner product / quadratic form (see Sec. II C), while maximizing the conditional Student-$t$ likelihood requires iterating over several such least-squares fits within the EM-algorithm (see Sec. III C). The EM-iterations were

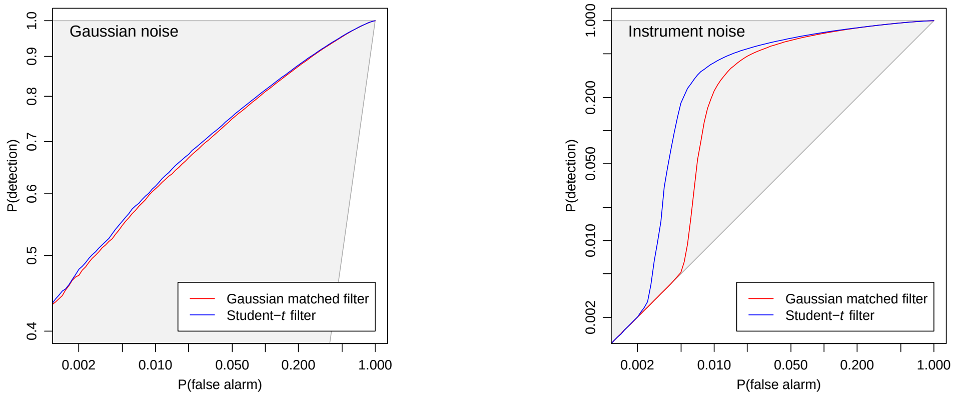

FIG. 4. ROC curves for the Gaussian and the Student-$t$ detection statistics in both data scenarios. The shaded area marks the region where any sensible detection statistic (one that is not worse than mere guessing) should lie.

terminated whenever the improvement in logarithmic likelihood over the previous iteration fell below $10^{-6}$. In this example setting, this lead to an average number of 4 EM iterations for each conditional likelihood maximization in both noise scenarios. The eventual maximized likelihood then is given by the overall maximum over the conditional maxima, and as the detection statistic we use the maximized likelihood ratio $\frac{p(d|\hat{\theta})}{p(d|\vec{0})}$. The algorithm used was essentially the one described in Appendix A 3 d.

### E. Simulation results

Fig. 3 shows resulting detection statistic values (maximized likelihood ratios) under the Gaussian and the Student-$t$ models both when a signal is injected as well as when he data are noise only. The signal injections here were all done at the same amplitude relative to the noise spectrum (SNR $\varrho = 5.257$). In general, both detection statistics are very similar; the Student-$t$ likelihood ratio tends to turn out slightly lower than the Gaussian one, in particular in the case of real interferometer noise.

The question of to what extent these differences affect the ability to discriminate signals from noise will be approached by considering the receiver operating characteristic (ROC) curves. ROC curves are based on the detection statistics' (here: empirical) distributions. Placing different detection thresholds on a detection statistic yields a corresponding false alarm probability (based on the distribution under the noise-only hypothesis) as well as a detection probability (based on the distribution under the particular signal hypothesis). The ROC curve illustrates these combinations over varying threshold values [42].

Fig. 4 shows ROC curves for the Gaussian and the Student-$t$ filter for both noise cases. In the case of simulated Gaussian noise, both detection statistics perform almost identically. For real instrument noise on the other hand, the Student-$t$ model is able to provide a significantly greater detection probability especially at low false alarm probabilities. A remarkable feature of the ROC curve for instrumental noise is that for very low false alarm probabilities both filters eventually perform as poorly as mere guessing. The Student-$t$ filter is able to sustain its discriminating power for lower false alarm rates, though. This effect is connected to the frequency of noise outliers ("glitches") in the data, leading to very large detection statistic values even in the absence of a signal. Fig. 5 shows the corresponding detection thresholds as a function of false alarm probabilities. The point where the detection threshold reaches the injected signals' SNR is where the corresponding detection probability is ≈ 50%. One can see that, due to the heavy tailed distribution of detection statistics in the case of actual instrument noise, the detection threshold necessary for low false alarm probabilities very quickly grows beyond values that could obviously be attributed to be due to the signal injections considered here; the rate of noise transients of "SNR" greater than the injections' SNR exceeds the false alarm rate (in a realistic search, some of these might actually be vetoed beforehand). This effect is very obvious here also because signal injections were done only at a single SNR, but it will of course persist for other SNR distributions—assuming other SNR distributions for injections will affect the detection probability, but not the detection threshold, i.e., the detection procedure itself.

The exact relative performance of both methods of course depends on the details of the particular detection problem, the kind of signal searched for, the parameter space, noise characteristics, data conditioning, and tuning parameters. The ROC curves shown above are based on a particular, artificial signal population, but their general features persist in a number of additional simulations not shown here, for a range of d.f. settings, injection SNRs, data from a different instrument, and data from a different time period.

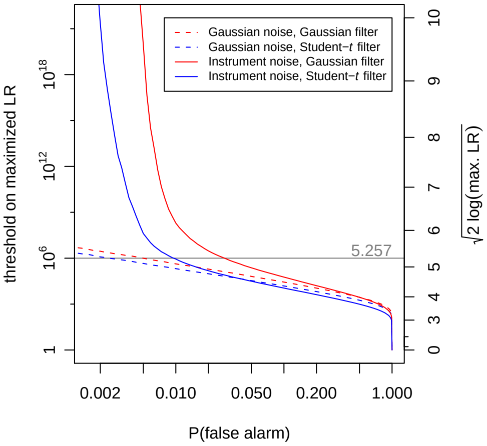

FIG. 5. Detection thresholds on the maximized likelihood ratio (the detection statistic), corresponding to certain false alarm probabilities. These thresholds are based on the detection statistic's distribution in the absence of a signal. The horizontal line indicates the injected signals' SNR (see also Fig. 4).

## V. CONCLUSIONS

We introduced a generalization of the matched filter that is commonly applied in signal detection problems. The Student-$t$ filter is derived as a maximum-likelihood detection method that is based on a Student-$t$ distribution for the noise, rather than a Gaussian distribution, which would again yield the common matched filter instead. On the technical side, it generalizes a least-squares method to an adaptive variety. While a "Gaussian" matched filter is certainly appropriate when the assumption of stationary Gaussian noise and a known spectrum is met, there are several ways to motivate the Student-$t$ filter as a robust alternative when these assumptions are violated: (i) "*theoretically*": the Student-$t$ model allows for uncertainty in the PSD, heavier-tailed noise and outliers, (ii) "*heuristically*": the resulting adaptive least-squares method is less outlier-sensitive, or (iii) "*pragmatically*": the filter may turn out more effective in practice, as in the realistic example shown above. Besides that, being a generalisation of the (Gaussian) matched filter, it should generally be able to perform as well or better. The question of course is whether the gain in detection efficiency is worth the additional implementation, tuning and computational effort. The difference in computational cost for deriving both detection statistics suggests that a combined, hierarchical search strategy may also be worth considering.

In the example shown above, the Student-$t$ model's degrees-of-freedom parameter was treated as a single constant. In the context of gravitational-wave interferometric data, this is an oversimplification; a study of actual instrument noise shows that the Fourier-domain data's tail behaviour clearly depends on the frequency [43]. Accounting for this effect in an actual search by fitting individual $\nu_j$ parameters for different frequency ranges may yield a significant improvement. It may also make sense to specify the degrees-of-freedom parameter dependent on additional information, like e.g. the data quality category [44].

It will be interesting to study the Student-$t$ filter's performance in a realistic search for gravitational-wave signals, in conjunction with the existing infrastructure (data quality flags, additional vetoes, etc.) and in comparison with the conventinal matched filter [18, 19]. We are also investigating the use of the Student-$t$ model in the context of Bayesian model selection [45]. Here it may again yield a more robust discriminator for actual signals against noise; on the computational side this problem is based on integration of the likelihood, rather than maximization, and we do not expect a difference in computational cost between Gaussian and Student-$t$ models. We expect the Student-$t$ filtering procedure to be also useful in many other signal-processing contexts, wherever robustness or uncertainty in the power spectrum is an issue.

## ACKNOWLEDGMENTS

The author wishes to thank Nelson Christensen, Drew Keppel, Karsten Lübke, Renate Meyer and Reinhard Prix for fruitful discussions at various stages of this work, and the LIGO Scientific Collaboration (LSC) for providing the data used here. The author gratefully acknowledges the support of the United States National Science Foundation for the construction and operation of the LIGO Laboratory and the Science and Technology Facilities Council of the United Kingdom, the Max-Planck-Society, and the State of Niedersachsen/Germany for support of the construction and operation of the GEO600 detector. The author also gratefully acknowledges the support of the research by these agencies and by the Australian Research Council, the International Science Linkages program of the Commonwealth of Australia, the Council of Scientific and Industrial Research of India, the Istituto Nazionale di Fisica Nucleare of Italy, the Spanish Ministerio de Educación y Ciencia, the Conselleria d'Economia, Hisenda i Innovació of the Govern de les Illes Balears, the Royal Society, the Scottish Funding Council, the Scottish Universities Physics Alliance, The National Aeronautics and Space Administration, the Carnegie Trust, the Leverhulme Trust, the David and Lucile Packard Foundation, the Research Corporation, and the Alfred P. Sloan Foundation.

## APPENDIX

### 1. Discrete Fourier transform

The Fourier transform convention used in this paper is specified below; it is defined for a real-valued function $h$

of time $t$, sampled at $N$ discrete time points, at a sampling rate of $\frac{1}{\Delta_t}$, and it maps from

$$\{h(t) \in \mathbb{R} : \ t = 0, \Delta_t, 2\Delta_t, \ldots, (N-1)\Delta_t\} \tag{A34}$$

to a function of frequency $f$

$$\{\tilde{h}(f) \in \mathbb{C} : \ f = 0, \Delta_f, 2\Delta_f, \ldots, (N-1)\Delta_f\}, \tag{A35}$$

where $\Delta_f = \frac{1}{N\Delta_t}$ and

$$\tilde{h}(f) \ = \ \sum_{j=0}^{N-1} h(j\Delta_t)\, \exp(-2\pi \mathrm{i} j \Delta_t f) \tag{A36}$$

[3].

### 2. Applying the EM-algorithm

#### *a. Preliminaries*

The *Expectation-Maximization (EM) algorithm* is required for maximizing the Student-*t* likelihood; see Sec. III C. What is desired is the maximum of the marginal likelihood $p(d|\theta)$, which is equivalent to the marginal density $p(\theta|d)$ when assuming a uniform prior distribution on $\theta$. What is required in order to apply the EM algorithm are expressions involving the marginalized $\sigma_j^2$ parameters, namely the conditional distribution $\mathrm{P}(\vec{\sigma}^2|\theta, d)$ and the joint density $p(\theta, \vec{\sigma}^2|d)$. The EM algorithm will then iteratively maximize the likelihood function by performing alternating "expectation" and "maximization" steps [8, 33].

The conditional posterior distribution $\mathrm{P}(\sigma_j^2|\theta, d)$ of the $j$th variance parameter $\sigma_j^2$ for given data and signal $s_\theta$ is a scaled inverse $\chi^2$-distribution:

$$\text{Inv-}\chi^2\left(\nu_j + 2, \frac{\nu_j S_1(f_j)_j + 4\frac{\Delta_t}{N}\left|\tilde{d}(f_j) - \tilde{s}_\theta(f_j)\right|^2}{\nu_j + 2}\right) \tag{A37}$$

[3] with probability density function

$$f(\sigma_j^2) \propto \left(\sigma_j^2\right)^{-\frac{\nu_j+4}{2}} \exp\left(-\frac{\frac{\nu_j}{2} S_1(f_j) + 4\frac{\Delta_t}{N}\left|\tilde{d}(f_j) - \tilde{s}_\theta(f_j)\right|^2}{2\sigma_j^2}\right) \tag{A38}$$

[3].

The conditional distribution of the data $d$ for given variances $\vec{\sigma}^2$ and signal parameters $\theta$, $\mathrm{P}(y|\theta, \vec{\sigma}^2)$, is Gaussian [3], and the variance parameters' prior, $\mathrm{P}(\vec{\sigma}^2)$, again was Inv-$\chi^2$ [3]. The joint conditional density of $\theta$ and $\vec{\sigma}^2$ for given data $d$ is given by

$$\log\big(p(\theta, \sigma^2|y)\big) \ \propto \ \log\big(p(y|\theta, \sigma^2) \times p(\theta, \sigma^2)\big) \tag{A39}$$

$$\propto -\sum_j \left(\log(\sigma_j^2) + \frac{4\frac{\Delta_t}{N}\left|\tilde{d}(f_j) - \tilde{s}_\theta(f_j)\right|^2}{2\sigma_j^2}\right)$$
$$-\sum_j \left((1 + \tfrac{\nu_j}{2}) \log(\sigma_j^2) + \frac{\nu_j S_1(f_j)}{2\sigma_j^2}\right) \tag{A40}$$

$$= -\sum_j \left((2 + \tfrac{\nu_j}{2}) \log(\sigma_j^2) + \frac{\nu_j S_1(f_j) + 4\frac{\Delta_t}{N}\left|\tilde{d}(f_j) - \tilde{s}_\theta(f_j)\right|^2}{2\sigma_j^2}\right) \tag{A41}$$

[3].

#### *b. The E-step*

For the EM algorithm's "expectation" step, one needs to evaluate the conditional posterior expectation

$$\mathrm{E}_{\mathrm{P}(\sigma^2|\theta=\theta_0, y)}\left[\log\big(p(\theta, \sigma^2|y)\big)\right]$$
$$= \int \log\big(p(\theta, \sigma^2|y)\big)\, p(\sigma^2|\theta{=}\theta_0, y)\, \mathrm{d}\sigma^2 \tag{A42}$$

as a function of $\theta$ for some given $\theta_0$ [8]. Here:

$$\int \log\big(p(\theta, \sigma^2|y)\big)\, p(\sigma^2|\theta{=}\theta_0, y)\, \mathrm{d}\sigma^2 \tag{A43}$$

$$\propto -\sum_j \int \left((2 + \tfrac{\nu_j}{2}) \log(\sigma^2) + \frac{\nu_j S_1(f_j) + 4\frac{\Delta_t}{N}\left|\tilde{d}(f_j) - \tilde{s}_\theta(f_j)\right|^2}{2\sigma_j^2}\right)$$
$$\times \left(\left(\sigma^2\right)^{-(2+\frac{\nu_j}{2})} \exp\left(\frac{\nu_j S_1(f_j) + 4\frac{\Delta_t}{N}\left|\tilde{d}(f_j) - \tilde{s}_{\theta_0}(f_j)\right|^2}{2\sigma^2}\right)\right) \mathrm{d}\sigma_j^2 \tag{A44}$$

$$\propto -\sum_j \frac{4\frac{\Delta_t}{N}\left|\tilde{d}(f_j) - \tilde{s}_\theta(f_j)\right|^2}{2} \times \int \frac{1}{\sigma_j^2} \left(\left(\sigma^2\right)^{-(2+\frac{\nu_j}{2})} \exp\left(\frac{\nu_j S_1(f_j) + 4\frac{\Delta_t}{N}\left|\tilde{d}(f_j) - \tilde{s}_{\theta_0}(f_j)\right|^2}{2\sigma^2}\right)\right) \mathrm{d}\sigma_j^2, \tag{A45}$$

$$\text{where} \quad \int \frac{1}{\sigma_j^2} \overbrace{\left(\left(\sigma^2\right)^{-(2+\frac{\nu_j}{2})} \exp\left(\frac{\nu_j S_1(f_j) + 4\frac{\Delta_t}{N}\left|\tilde{d}(f_j) - \tilde{s}_{\theta_0}(f_j)\right|^2}{2\sigma^2}\right)\right)}^{(*)} \mathrm{d}\sigma_j^2 = \frac{\nu_j + 2}{\nu_j S_1(f_j) + 4\frac{\Delta_t}{N}\left|\tilde{d}(f_j) - \tilde{s}_{\theta_0}(f_j)\right|^2}, \tag{A46}$$

since the term marked by the asterisk $(*)$ is the density function of an Inv-$\chi^2\big(\nu_j+2, \frac{\nu_j S_1(f_j)+4\frac{\Delta_t}{N}\left|\tilde{d}(f_j)-\tilde{s}_{\theta_0}(f_j)\right|^2}{\nu_j+2}\big)$ probability distribution, so that

$$\int \log\big(p(\beta,\sigma^2|y)\big)\, p(\sigma^2|\beta\!=\!\beta_0,y)\,\mathrm{d}\sigma^2$$

$$\propto\; -\tfrac{1}{2}\sum_j \frac{4\frac{\Delta_t}{N}\left|\tilde{d}(f_j)-\tilde{s}_\theta(f_j)\right|^2}{\frac{\nu_j}{\nu_j+2}S_1(f_j)+\frac{1}{\nu_j+2}\left(4\frac{\Delta_t}{N}\left|\tilde{d}(f_j)-\tilde{s}_{\theta_0}(f_j)\right|^2\right)} \tag{A47}$$

$$=\; -\tfrac{1}{2}\sum_j \frac{\left|\tilde{d}(f_j)-\tilde{s}_\theta(f_j)\right|^2}{\frac{N}{4\Delta_t}\left(\frac{\nu_j}{\nu_j+2}\,S_1(f_j)+\frac{2}{\nu_j+2}\,\frac{2\Delta_t}{N}\left|\tilde{d}(f_j)-\tilde{s}_{\theta_0}(f_j)\right|^2\right)} \tag{A48}$$

$$=:\mathcal{E}(\theta_0,\theta).$$

### c. The M-step

In the EM algorithm's "maximization" step, the above expectation $\mathcal{E}(\theta_0,\theta)$ (A48) needs to be maximized with respect to the parameter $\theta$. The parameter value maximizing the expectation then constitutes the next iteration's "new" $\theta_0$ value, for which then the expectation again is maximized, and so forth [8]. As one can see from expression (A48), maximisation of the expectation again amounts to mimimizing weighted least-squares, as in the Gaussian matched filter described above.

## 3. Pseudocode matched and Student-t filters

### a. Preliminaries

This section sketches actual implementations of Student-$t$ and (Gaussian) matched filters in comparison. In the following, we will use essentially the same conventions as before; we will be considering a time series $d$ of length $N$, sampled at a sampling interval of $\Delta_t$. The signal waveform here is assumed to be a linear combination of a sine- and a cosine-component ($s_{\mathrm{s},\theta}$, $s_{\mathrm{c},\theta}$), it has an associated arrival time parameter, and possibly additional parameters $\theta$ (as in (26)). Additional waveform parameters (other than amplitude, phase and time) are then commonly treated by running several matched filters corresponding to different values of $\theta$. The generalisation to the case of more than two linear signal components should be straightforward. The profile likelihood will be evaluated along a discrete grid of time points $\tau_i$ $(i=1,\ldots,m)$, where the special case of $\tau_i=i\Delta_t$ and $m=N$ is of particular interest. The filter's output each time is a single number, the maximized (logarithmic) likelihood ratio of signal vs. no-signal models. We will be making use of the inner product / quadratic form notation $\langle a,b;S\rangle$ as defined in (19). Implementations of the algorithms sketched in Sec. A 3 c and A 3 e are also provided in [35].

TABLE I. Matched filter, general implementation.

```
normS = $\langle s_{\mathrm{s},\theta},\ s_{\mathrm{s},\theta};\ S_1\rangle$
normC = $\langle s_{\mathrm{c},\theta},\ s_{\mathrm{c},\theta};\ S_1\rangle$
for ($i=1,\ldots,m$) do // loop over time points:
  for ($j=0,\ldots,N/2$) do // time-shift the data:
    $\tilde{d}'_j \;=\; \tilde{d}_j \times \exp(2\pi\mathrm{i}f_j\tau_i)$
  end for
  prodS = $\langle s_{\mathrm{s},\theta},\ d';\ S_1\rangle$
  prodC = $\langle s_{\mathrm{c},\theta},\ d';\ S_1\rangle$
  // compute log-likelihood ratio / profile likelihood:
  maxLLR[$i$] = $(\mathtt{prodS})^2/\mathtt{normS}+(\mathtt{prodC})^2/\mathtt{normC}$
end for
return max(maxLLR)
```

### b. The "Gaussian" matched filter: general implementation

The first algorithm (Tab. I) is a "naive" matched filter implementation that maximizes the likelihood (-ratio) over a given grid of $m$ time points ($\tau$). The algorithm mainly consists of a loop over time points, where for each time point the (conditional) likelihood is maximized over amplitude and phase. In order to match signal and data for a certain signal arrival time, the data $d$ are time-shifted against the signal waveforms $s_{\mathrm{s/c}}$. The eventual result is the profile likelihood evaluated at the specified time points, the maximum of which then constitutes the generalized likelihood ratio detection statistic that is returned.

### c. The "Gaussian" matched filter: efficient implementation

If the time points to be maximized over are taken to be the same as the data time series' points ($\tau_i=i\Delta_t$, $i=1,\ldots,m=N$), then the matched-filtering procedure may be implemented much more efficiently. The algorithm shown in Tab. II will give identical results to the previous, but it is more efficient as it takes advantage of a Fourier transform to essentially maximize over amplitude, phase

TABLE II. Matched filter, efficient implementation.

```
normS = $\langle s_{\mathrm{s},\theta},\ s_{\mathrm{s},\theta};\ S_1\rangle$
normC = $\langle s_{\mathrm{c},\theta},\ s_{\mathrm{c},\theta};\ S_1\rangle$
for ($j=0,\ldots,(N-1)$) do // convolve data and signals:
  convS[$j+1$] = $\tilde{d}_j \times \tilde{s}^*_{\mathrm{s},\theta,j} \,/\, S_1(f_j)$
  convC[$j+1$] = $\tilde{d}_j \times \tilde{s}^*_{\mathrm{c},\theta,j} \,/\, S_1(f_j)$
end for
// apply Fourier transforms:
FTS = DFT(convS)
FTC = DFT(convC)
for ($i=1,\ldots,N$) do // profile likelihood (-ratio):
  maxLLR[$i$] = $\left(\frac{\Delta_t}{N}\right)^2\left(\frac{(\mathtt{FTS}[N+1-i])^2}{\mathtt{normS}}+\frac{(\mathtt{FTC}[N+1-i])^2}{\mathtt{normC}}\right)$
end for
return max(maxLLR)
```

TABLE III. Student-$t$ filter, general implementation.

```
LL0 = $\log(p(d, S_1, \nu))$ // log-likelihood noise-only model
for ($i = 1, \ldots, m$) do // loop over time points:
  for ($j = 0, \ldots, N/2$) do // time-shift the data:
    $\tilde{d}'_j = \tilde{d}_j \times \exp(2\pi \mathrm{i} f_j \tau_i)$
  end for
  // EM-iterations:
  $k = 1$;   $\Delta_{\mathtt{LLR}} = 1$;   LLRprev = 0;   $S_1^\star = S_1$
  while ($\Delta_{\mathtt{LLR}} > \Delta_{\max}$) and ($k \leq k_{\max}$) do
    normS = $\langle s_{\mathrm{s},\theta},\ s_{\mathrm{s},\theta};\ S_1^\star \rangle$
    normC = $\langle s_{\mathrm{c},\theta},\ s_{\mathrm{c},\theta};\ S_1^\star \rangle$
    prodS = $\langle s_{\mathrm{s},\theta},\ d';\ S_1^\star \rangle$
    prodC = $\langle s_{\mathrm{c},\theta},\ d';\ S_1^\star \rangle$
    $\hat{\beta}_{\mathrm{s}}$ = prodS/normS
    $\hat{\beta}_{\mathrm{c}}$ = prodC/normC
    $\hat{n} = d' - \big(\hat{\beta}_{\mathrm{s}} s_{\mathrm{s},\theta} + \hat{\beta}_{\mathrm{c}} s_{\mathrm{c},\theta}\big)$ // vector of noise residuals
    LL1 = $\log(p(\hat{n}, S_1, \nu))$ // log-likelihood signal model
    LLR = LL1 − LL0 // log-likelihood ratio
    $\Delta_{\mathtt{LLR}}$ = LLR − LLRprev
    LLRprev = LLR
    for ($j = 0, \ldots, N/2$) do // adapt the spectrum:
      $S_1^\star(f_j) = \frac{\nu_j}{\nu_j+2}\, S_1(f_j) + \frac{2}{\nu_j+2} \frac{2\Delta_t}{N} \left|\hat{n}_j\right|^2$
    end for
    $k = k + 1$
  end while
  maxLLR[$i$] = LLR // profile likelihood (-ratio)
end for
return max(maxLLR)
```

and time simultaneously (see also Sec. II D). In practice, one may want to restrict the profile likelihood maximization (line 13) to the subset of sensible time-shifts that do not "wrap" the signal circularly around the data's end points. Instead of a Fourier transform, one could also implement an inverse DFT and would then also not need to time-reverse the result's indices (line 11).

### *d. The Student-t filter: general implementation*

This algorithm (see Tab. III) again is a "general" version of the Student-$t$ filter, analogous to the general matched filter (Sec. A 3 b), where the set of time points $\tau$ is not restricted. The EM-algorithm here is applied at the level of each single amplitude/phase maximization conditional on some time shift $\tau_i$. The EM component requires the specification of a threshold $\Delta_{\max}$ on the improvement in logarithmic maximized likelihood ratio (e.g. $10^{-6}$), and a threshold $k_{\max}$ on the number of EM iterations (e.g. 100). The Student-$t$ likelihood function

$$p(x, S, \nu) \;\propto\; \exp\!\left(-\sum_j \tfrac{\nu_j+2}{2} \log\!\left[1 + \frac{1}{\nu_j} \frac{\left|\tilde{x}_j\right|^2}{\frac{N}{4\Delta_t}\, S_1(f_j)}\right]\right)$$

(see also (29)) only needs to be computed up to a proportionality constant here, as only the likelihood *ratio* is of eventual interest.

TABLE IV. Student-$t$ filter, efficient implementation.

```
    LL0 = $\log(p(d, S_1, \nu))$ // log-likelihood noise-only model
    // EM-iterations:
    $k = 1$;   $\Delta_{\mathtt{LLR}} = 1$;   LLRprev = 0;   $S_1^\star = S_1$
    while ($\Delta_{\mathtt{LLR}} > \Delta_{\max}$) and ($k \leq k_{\max}$) do
5:    // the "plain" matched filter:
      normS = $\langle s_{\mathrm{s},\theta},\ s_{\mathrm{s},\theta};\ S_1^\star \rangle$
      normC = $\langle s_{\mathrm{c},\theta},\ s_{\mathrm{c},\theta};\ S_1^\star \rangle$
      for ($j = 0, \ldots, (N-1)$) do
        convS[$j+1$] = $\tilde{d}_j \times \tilde{s}^*_{\mathrm{s},\theta,j} \,/\, S_1^\star(f_j)$
10:     convC[$j+1$] = $\tilde{d}_j \times \tilde{s}^*_{\mathrm{c},\theta,j} \,/\, S_1^\star(f_j)$
      end for
      FTS = DFT(convS)
      FTC = DFT(convC)
      for ($i = 1, \ldots, N$) do
15:     maxLLR[$i$] = $\left(\frac{\Delta_t}{N}\right)^2 \left(\frac{(\mathtt{FTS}[N+1-i])^2}{\mathtt{normS}} + \frac{(\mathtt{FTC}[N+1-i])^2}{\mathtt{normC}}\right)$
      end for
      // end of "plain" matched filter.
      // Determine best-fitting template, residuals, etc.:
      $i_{\max} = \arg\max_i$ maxLLR[$i$]
20:   for ($j = 0, \ldots, N/2$) do // time-shift the data:
        $\tilde{d}'_j = \tilde{d}_j \times \exp(2\pi \mathrm{i} f_j \tau_{i_{\max}})$
      end for
      prodS = $\langle s_{\mathrm{s},\theta},\ d';\ S_1^\star \rangle$
      prodC = $\langle s_{\mathrm{c},\theta},\ d';\ S_1^\star \rangle$
25:   $\hat{\beta}_{\mathrm{s}}$ = prodS/normS
      $\hat{\beta}_{\mathrm{c}}$ = prodC/normC
      $\hat{n} = d' - \big(\hat{\beta}_{\mathrm{s}} s_{\mathrm{s},\theta} + \hat{\beta}_{\mathrm{c}} s_{\mathrm{c},\theta}\big)$ // vector of noise residuals
      LL1 = $\log(p(\hat{n}, S_1, \nu))$ // log-likelihood signal model
      LLR = LL1 − LL0 // log-likelihood ratio
30:   $\Delta_{\mathtt{LLR}}$ = LLR − LLRprev
      LLRprev = LLR
      for ($j = 0, \ldots, N/2$) do // adapt the spectrum:
        $S_1^\star(f_j) = \frac{\nu_j}{\nu_j+2}\, S_1(f_j) + \frac{2}{\nu_j+2} \frac{2\Delta_t}{N} \left|\hat{n}_j\right|^2$
      end for
35:   $k = k + 1$
    end while
    return LLR
```

### *e. The Student-t filter: efficient implementation*

The Student-$t$ filter also may be implemented more efficiently in case the signal arrival times to maximize over is taken to be the time points of the original time series ($\tau_i = i\Delta_t$, $i = 1, \ldots, m = N$, as in Sec. A 3 c). This implementation (Tab. IV) then requires to move the level at which the EM-agorithm is applied from the conditional maximization over amplitude and phase to the joint amplitude/phase/time maximization; effectively this implementation iteratively runs several matched filters (see lines 6–16) while adapting the noise spectrum in between. It is unclear whether or how the level at which the EM-algorithm is applied affects the results; as noted in Sec. III D, the likelihood may be multimodal and different implementations might end up with differing maximization results, but whether this actually poses a problem in practice is not obvious. Computationally, this latter

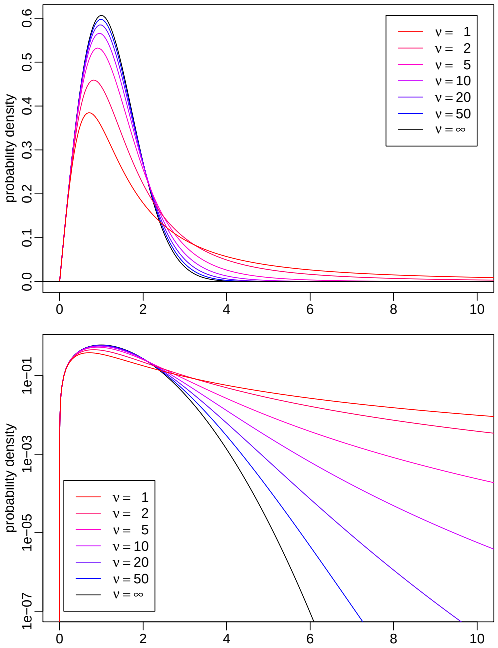

FIG. 6. Probability density functions of “Student-Rayleigh” distributions for varying degrees of freedom $\nu$ and fixed scale $\sigma^2 = 1$. For $\nu = \infty$, the distribution corresponds to the usual (“Gaussian”) Rayleigh distribution.

implementation should be much easier, though. Another difference to note is that while the matched filter allows to return the profile likelihood as a function of time (the “*SNR time series*”), only the Student-$t$ filter implementation from Sec. A 3 d is able to provide this, while the more efficient implementation will only return the overall maximum.

### 4. The “Student-Rayleigh” distribution

#### a. Relation to the F-distribution

The noise power’s probability distribution under the Student-$t$ model (see (33), Sec. IV C) may be related to Snedecor’s $F$-distribution. Firstly, real and imaginary parts of the $j$th element of the discretely Fourier-transformed vector $n$ follow a multivariate (*bi*variate) Student-$t$ distribution (see Sec. III A). Let $A$ and $B$ be independent Gaussian random variables with zero mean and standard deviation $\sigma$. Let furthermore $C$ be a $\chi^2_\nu$-distributed random variable with $\nu$ degrees of freedom. Then the random vector

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \frac{1}{\sqrt{C/\nu}} \begin{pmatrix} A \\ B \end{pmatrix}$$

follows a bivariate Student-$t$ distribution with a diagonal covariance matrix, exactly like the real and imaginary components of $\tilde{n}(f_j)$ [8]. The root-mean-square figure corresponding to the power then may be written as

$$\sqrt{X^2+Y^2} = \sqrt{2\sigma^2 \, \frac{\left(\left(\frac{A}{\sigma}\right)^2 + \left(\frac{B}{\sigma}\right)^2\right) / \, 2}{C \, / \, \nu}} = \sqrt{2\sigma^2 \, D}, \tag{A49}$$

where the random variable $D$, being a ratio of $\chi^2$-distributed random variables that are normalized by their respective degrees-of-freedom, follows an $F(2,\nu)$-distribution with 2 and $\nu$ degrees of freedom [7].

#### b. Probability density function, etc.

In the Gaussian noise model (see Sec. II B), the noise power at the $j$th frequency bin, $\bigl|\tilde{n}(f_j)\bigr|$, follows a Rayleigh distribution with probability density function

$$f_{\mathrm{R}}(x|\sigma) = \tfrac{x}{\sigma^2} \, \exp\bigl(-\tfrac{x^2}{2\sigma^2}\bigr), \tag{A50}$$

where the scale parameter $\sigma$ is given as $\sigma = \sqrt{\frac{N}{4\Delta_t} S_1(f_j)}$.

The analogue “Student-Rayleigh” probability distribution in the Student-$t$ noise model (see Sec. III A) is defined through its density function

$$f_{\mathrm{SR}}(x|\sigma,\nu) = \tfrac{x}{\sigma^2} \, f_{F(2,\nu)}\bigl(\tfrac{x^2}{2\sigma^2}\bigr), \tag{A51}$$

where $f_{F(2,\nu)}(\cdot)$ is the probability density function of an $F(2,\nu)$-distribution with 2 and $\nu$ degrees of freedom. Similarly, the cumulative distribution function and quantile function are given by

$$F_{\mathrm{SR}}(x|\sigma,\nu) = F_{F(2,\nu)}\bigl(\tfrac{x^2}{2\sigma^2}\bigr) \quad \text{and} \tag{A52}$$

$$Q_{\mathrm{SR}}(p|\sigma,\nu) = \sqrt{2\sigma^2 \, Q_{F(2,\nu)}(p)}, \tag{A53}$$

where $F_{F(2,\nu)}(\cdot)$ and $Q_{F(2,\nu)}(\cdot)$ are the $F$-distribution’s cumulative distribution function and quantile function.

Fig. 6 illustrates probability density functions of “Student-Rayleigh” probability distributions for varying degrees of freedom $\nu$. For $\nu = \infty$, the distribution corresponds to the usual (“Gaussian”) Rayleigh distribution. Note in particular the differing tail behaviour (analogous to Fig. 1) that is apparent especially in the logarithmic plot.

[1] K. S. Thorne. Gravitational radiation. In S. W. Hawking and W. Israel, editors, *300 years of gravitation*, chapter 9, pages 330–358. Cambridge University Press, Cambridge,

1987.

[2] B. F. Schutz. Gravitational wave astronomy. *Classical and Quantum Gravity*, 16(12A):A131–A156, December 1999.

[3] C. Röver, R. Meyer, and N. Christensen. Modelling coloured residual noise in gravitational-wave signal processing. *Classical and Quantum Gravity*, 28(1):015010, January 2011.

[4] G. L. Turin. An introduction to matched filters. *IRE Transactions on Information Theory*, 6(3):311–329, June 1960.

[5] N. Choudhuri, S. Ghosal, and A. Roy. Contiguity of the Whittle measure for a Gaussian time series. *Biometrika*, 91(4):211–218, 2004.

[6] L. S. Finn. Detection, measurement, and gravitational radiation. *Physical Review D*, 46(12):5236–5249, December 1992.

[7] A. M. Mood, F. A. Graybill, and D. C. Boes. *Introduction to the theory of statistics*. McGraw-Hill, New York, 3rd edition, 1974.

[8] A. Gelman, J. B. Carlin, H. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman & Hall / CRC, Boca Raton, 1997.

[9] J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer-Verlag, 2nd edition, 1985.

[10] R. Prix and B. Krishnan. Targeted search for continuous gravitational waves: Bayesian versus maximum-likelihood statistics. *Classical and Quantum Gravity*, 26(20):204013, October 2009.

[11] A. C. Searle. Monte-Carlo and Bayesian techniques in gravitational wave burst data analysis. *Arxiv preprint 0804.1161 [gr-qc]*, April 2008.

[12] C. Röver, M.-A. Bizouard, N. Christensen, H. Dimmelmeier, I. S. Heng, and R. Meyer. Bayesian reconstruction of gravitational wave burst signals from simulations of rotating stellar core collapse and bounce. *Physical Review D*, 80(10):102004, November 2009.

[13] K. Cannon, A. Chapman, C. Hanna, D. Keppel, A. C. Searle, and A. Weinstein. Singular value decomposition applied to compact binary coalescence gravitational-wave signals. *Physical Review D*, 82(4):044025, August 2010.

[14] P. Jaranowski, A. Królak, and B. Schutz. Data analysis of gravitational-wave signals from spinning neutron stars: The signal and its detection. *Physical Review D*, 58(6):063001, September 1998.

[15] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied linear statistical models*. McGraw-Hill, New York, 4th edition, 1996.

[16] C. Röver, C. Messenger, and R. Prix. Bayesian versus frequentist upper limits. *Arxiv preprint 1103.2987*, February 2011.

[17] B. Allen et al. Observational limit on gravitational waves from binary neutron stars in the galaxy. *Physical Review Letters*, 83(8):1498–1501, August 1999.

[18] B. Allen, W. G. Anderson, P. G. Brady, D. A. Brown, and J. D. E. Creighton. Findchirp: an algorithm for detection of gravitational waves from inspiraling compact binaries. *Arxiv preprint gr-qc/0509116*, September 2005.

[19] D. A. Brown. Using the Inspiral program to search for gravitational waves from low-mass binary inspiral. *Classical and Quantum Gravity*, 22(18):S1097–S1107, September 2005.

[20] M. Was et al. On the background estimation by time slides in a network of gravitational wave detectors. *Classical and Quantum Gravity*, 27(1):015005, January 2010.

[21] W. S. Gosset. The probable error of a mean. *Biometrika*, 6(1):1–25, March 1908.

[22] H. H. Kelejian and I. R. Prucha. Independent or uncorrelated disturbances in linear regression: An illustration of the difference. *Economics Letters*, 19(1):35–38, 1985.

[23] T. S. Breusch, J. C. Robertson, and A. H. Welsh. The emperor's new clothes: a critique of the multivariate $t$ regression model. *Statistica Neerlandica*, 51(3):269–286, December 2001.

[24] K. L. Lange, R. J. A. Little, and J. M. G Taylor. Robust statistical modeling using the $t$ distribution. *Journal of the American Statistical Association*, 84(408):881–896, December 1989.

[25] J. Geweke. Bayesian treatment of the independent Student-$t$ linear model. *Journal of Applied Econometrics*, 8:S19–S40, December 1993.

[26] D. R. Divgi. Robust estimation using Student's $t$ distribution. CNA Research Memorandum CRM 90-217, Center for Naval Analyses, Alexandria, VA, USA, December 1990.

[27] J. B. McDonald and W. K. Newey. Partially adaptive estimation of regression models via the generalized $t$ distribution. *Econometric Theory*, 4(3):428–457, December 1988.

[28] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust statistics: The approach based on influence functions*. Wiley, New York, 1986.

[29] P. J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley, 2nd edition, 2009.

[30] J. D. Creighton. Data analysis strategies for the detection of gravitational waves in non-Gaussian noise. *Physical Review D*, 60(2):021101, July 1999.

[31] B. Allen, J. D. E. Creighton, É. É. Flanagan, and J. D. Romano. Robust statistics for deterministic and stochastic gravitational waves in non-Gaussian noise: Frequentist analyses. *Physical Review D*, 65(12):122002, June 2002.

[32] B. Allen. $\chi^2$ time-frequency discriminator for gravitational wave detection. *Physical Review D*, 71(6):062001, March 2005.

[33] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–38, 1977.

[34] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, March 1938.

[35] C. Röver. bspec: Bayesian spectral inference, version 1.3, 2011. R package. URL: `http://cran.r-project.org/package=bspec`.

[36] T. Mäkeläinen, K. Schmidt, and G. P. H. Styan. On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed-size samples. *The Annals of Statistics*, 9(4):758–767, July 1981.

[37] T. Damour, B. R. Iyer, and B. S. Sathyaprakash. Comparison of search templates for gravitational waves from binary inspiral. *Physical Review D*, 63(4):044023, January 2001.

[38] B. P. Abbott et al. LIGO: the Laser Interferometer Gravitational-wave Observatory. *Reports on Progress in Physics*, 72(7):076901, July 2009.

[39] P. D. Welch. The use of Fast Fourier Transform for the

estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, AU-15(2):70–73, June 1967.

[40] T. Tanaka and H. Tagoshi. Use of new coordinates for the template space in a hierarchical search for gravitational waves from inspiraling binaries. *Physical Review D*, 62(8):082001, October 2000.

[41] M. B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1–17, March 1968.

[42] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

[43] S. Waldman. Rayleigh distributions for H1, L1 for S6a. LIGO-Virgo collaboration internal report, November 2009.

[44] J. Slutsky et al. Methods for reducing false alarms in searches for compact binary coalescences in LIGO data. *Classical and Quantum Gravity*, 27(16):165023, August 2010.

[45] J. Veitch and A. Vecchio. Bayesian approach to the follow-up of candidate gravitational wave signals. *Physical Review D*, 78(2):022001, July 2008.