Eur. Phys. J. C (2024) 84:1023
https://doi.org/10.1140/epjc/s10052-024-13400-6

THE EUROPEAN PHYSICAL JOURNAL C

Regular Article - Experimental Physics

Maximum entropy spectral analysis: an application to gravitational waves data analysis

Alessandro Martini[1,2,3,a], **Stefano Schmidt**[4,5,b], **Gregory Ashton**[6,c], **Walter Del Pozzo**[1,d]

[1] Dipartimento di Fisica Università di Pisa, and INFN Sezione di Pisa, 56127 Pisa, Italy
[2] Dipartimento di Fisica, Università di Trento, Povo, 38123 Trento, Italy
[3] INFN, Trento Institute for Fundamental Physics and Applications, Povo, 38123 Trento, Italy
[4] Institute for Gravitational and Subatomic Physics (GRASP), Utrecht University, Princetonplein 1, 3584 CC Utrecht, The Netherlands
[5] Nikhef, Science Park 105, 1098 XG Amsterdam, The Netherlands
[6] Royal Holloway University of London, London, UK

Received: 1 July 2024 / Accepted: 20 September 2024 / Published online: 8 October 2024

Abstract The maximum entropy spectral analysis (MESA) method, developed by Burg, offers a powerful tool for spectral estimation of a time-series. It relies on Jaynes' maximum entropy principle, allowing the spectrum of a stochastic process to be inferred using the coefficients of an autoregressive process AR(p) of order p. A closed-form recursive solution provides estimates for both the autoregressive coefficients and the order p of the process. We provide a ready-to-use implementation of this algorithm in a Python package called `memspectrum`, characterized through power spectral density (PSD) analysis on synthetic data with known PSD and comparisons of different criteria for stopping the recursion. Additionally, we compare the performance of our implementation with the ubiquitous Welch algorithm, using synthetic data generated from the GW150914 strain spectrum released by the LIGO-Virgo-Kagra collaboration. Our findings indicate that Burg's method provides PSD estimates with systematically lower variance and bias. This is particularly manifest in the case of a small (O(5000)) number of data points, making Burg's method most suitable to work in this regime. Since this is close to the typical length of analysed gravitational waves data, improving the estimate of the PSD in this regime leads to more reliable posterior profiles for the system under study. We conclude our investigation by utilising MESA, and its particularly easy parametrisation where the only free parameter is the order p of the AR process, to marginalise over the interferometers noise PSD in conjunction with inferring the parameters of GW150914

[a] e-mail: alessandro.martini-1@unitn.it (corresponding author)
[b] e-mail: s.schmidt@uu.nl
[c] e-mail: Gregory.Ashton@rhul.ac.uk
[d] e-mail: walter.delpozzo@unipi.it

1 Introduction

The problem of inferring the morphology and the defining parameters of deterministic signals superimposed to stochastic processes is one of the most wide spread and interesting problems in several areas of human activities. Whenever some form of model for the signal we are looking for is available, the problem is typically solved via the Wiener filter, defined as the whitening filter that maximises the signal-to-noise ratio, i.e. the relative power of the (known) signal over the power of the (known) underlying stochastic process. Hence, signal detection and characterisation requires accurate knowledge of (i) the shape of the signal we are looking for and (ii) the statistical properties of the stochastic process. The construction of signal models is typically driven either by physical or by mathematical arguments hence, although extremely difficult in general, it is doable. On the other hand, stochastic process models can be extremely difficult to construct, both for practical and theoretical reasons. A stochastic process is fully described by the knowledge of the probability distribution governing its realisations – the "paths" of the random variable under scrutiny – over the entire time axis, from $t = -\infty$ to $t = \infty$. Clearly this is not possible in practice. Therefore modelling a stochastic process either relies on modelling of the underlying physical processes, thus falling back onto the deterministic case, or on modelling the mathematical and statistical properties of the process, and potentially infering them from the process realisations. The study of the properties of stochastic processes is thus a crucial task in many fields of physics, astronomy, quantitative biology, as well as engineering and finance. Among the classes of stochastic processes, a key role is played by *wide-sense*

stationary processes. These are stochastic processes that display an invariance of their statistical properties, such as their two-point autocovariance function, with respect to translation of the independent variable, usually the time t. If $x(t)$ is a wide-sense stationary process, its statistical properties are completely determined by the knowledge of the (many-points) autocorrelation functions. In practice, one often has easy access to the two-point correlation function

$$C(\tau) = \mathbf{E}[x_t \cdot x_{t+\tau}] \tag{1}$$

or, equivalently, to the process *power spectral density* (PSD) $S(f)$. Thanks to the Wiener–Khinchin theorem, in wide-sense stationary processes, the two are in fact related by a Fourier transform:

$$S(f) = \int_{-\infty}^{\infty} \mathrm{d}\tau C(\tau) e^{-i2\pi f\tau}. \tag{2}$$

In the context of gravitational waves physics, e.g. [1], the PSD is introduced as

$$\mathbf{E}[\tilde{x}(f) \cdot \tilde{x}(f')] = S(f)\delta(f - f') \tag{3}$$

without highlighting its connection with the time structure of the process itself, thus masking some important properties that will be explored further in what follows. The latter definition in Eq. (3) gives, however, (i) a straightforward interpretation of the PSD: it measures how much signal "power" is located in each frequency; (ii) an operative way of estimating it for an unknown process.

An ubiquitous method for such a computation is due to Welch [2] and it is based on Eqs. (2–3). The PSD is obtained by slicing the observed realisation $x(t_1), \ldots, x(t_n)$ of the process $x(t)$ into many window-corrected batches and averaging the squared moduli of their Fourier transforms. This approach is equivalent [3,4] to taking the Fourier Transform of the windowed sample autocorrelation ρ_W, written as

$$\rho_W = \{W_0\rho_0, W_{\pm1}\rho_{\pm1}, \ldots, W_{\pm M}\rho_{\pm M}, 0, 0, \ldots\}, \tag{4}$$

where ρ is the empirical autocorrelation and M is the maximum time lag at which the autocorrelation is computed. The sequence W is a window function that can be chosen in several different ways, each choice presenting advantages and disadvantages for the final estimate of the PSD.

The choice of a window function is arbitrary and typically is made by trial and error, until a satisfactory compromise between variance and resolution of the estimate of PSD is reached. A high frequency resolution implies high variance and vice-versa. Besides the window function, Welch's method requires a number of arbitrary choices to be made, such as the number of time slices and the overlap between consecutive slices. All these knobs must be tuned by hand and their choice can dramatically affect the PSD estimation, hence begging the question of what the "best" PSD estimate is.

Another drawback of this approach is the requirement for the window to be 0 outside the interval in which the autocorrelation is computed. We are arbitrarily assuming $\rho_j = 0$ for $j > M$ and modifying the estimate (i.e. the data) if a non-rectangular window is chosen. Making assumptions on unobserved data and modifying the ones we have at our disposal introduces "spurious" information about the process that we, in general, do not really have.

A alternative approach providing a smooth PSD estimation, is to adopt a parametric model for the PSD and to fit its parameters to the data with a Reversible Jump Markov Chain Monte Carlo [5,6]. Despite being effective, this method is problem dependent, since it needs to make definite assumptions on the shape of the PSD. Moreover, it can be computationally expensive. For all the above reasons, we did not consider such methods in our work.

An appealing alternative, based on the maximum entropy principle [7–9], has been derived by Burg [10]. Being rooted on solid theoretical foundations, we will see that Burg's method, unlike Welch's, does not require any preprocessing of the data and requires very little tuning of the algorithm parameters, since it provides an iterative closed form expression for the spectrum of a stochastic stationary time series. Furthermore, it embeds the PSD estimation problem into an elegant theoretical framework and makes minimal assumptions on the nature of the data. Lastly and most importantly, it provides a robust link between spectral density estimation and the field of autoregressive processes. This provides a natural and simple machinery to forecast a time series, thus predicting future observations based on previous ones.

In this paper, we discuss the details of the maximum entropy principle, its application to the problem of PSD estimation with Burg's algorithm and the link between Burg's algorithm and autoregressive process. Our goal is to bring (again) to public attention maximum entropy spectral analysis, in the hope that it will be widely employed as a way out of the many undesired aspects of the Welch's algorithm (or other similar methods). To facilitate this goal, we based this study on `memspectrum`, a freely available, robust and easy-to-use python implementation of the algorithm described below.[1] We provide a thorough assessment of the performance of our code and we validate our results performing a number of tests on simulated and real data. We also compare our results with those of spectral analysis carried out with the standard Welch's method. In order to apply our model on a realistic setting, we analyse some time series of broad interest in the scientific community.

Our paper is organized as follows: we begin by briefly reviewing the theoretical foundations of the maximum entropy principle in Sect. 2. Section 3 presents the validation of Burg's method as well as of our implementation

[1] It is available at link: https://pypi.org/project/memspectrum/.

on simulated data. In Sect. 4 we compare the results from `memspectrum` with the Welch method; Sect. 5 presents a few applications to real time series, including the analysis of GW150914, and, finally, we conclude with a discussion in Sect. 6.

2 Theoretical foundations

The maximum entropy principle (MAXENT) is among the most important results in probability theory. It provides a way to uniquely assign probabilities to a phenomenon in a way that best represent our state of knowledge, while being non-committal to unavailable information. Its domain of application turned out to be wider than expected. In fact, thanks to [10], this method has also been applied to perform high quality computation of power spectral densities of time series.

After a short introduction to Jaynes' MAXENT (Sect. 2.1), we will review in detail Burg's technique of maximum entropy spectral analysis (MESA) and show that the estimate can always be expressed in an analytical closed form (Sect. 2.2). Next, we will discuss the interesting link between Burg's method and autoregressive processes (Sect. 2.3) and in Sect. 2.4 we will use such link to forecast a time series.

2.1 Maximum entropy principle

Before introducing the MAXENT principle, we will define, via some simple examples, the two core concepts of the problem and the roles they play in deductive inference: the 'evidence' and the 'information'. Let us start with the 'information' (or information entropy): it is a measure of the degree of uncertainty on the outcomes of some experiment and specifies the length of the message necessary to provide a full description of the system under study. As an example, no information is brought if we are studying a system whose outcome is certain (the outcome is known with probability $p = 1$), as in this case, a communication is not even needed. Shannon [11] proposed the quantity

$$I = \log_2 \frac{1}{p(x)} \tag{5}$$

to measure the quantity of information brought by an outcome x with probability $p(x)$. It is additive quantity as well as a monotonically decreasing function of $p \in [0, 1]$: the more uncertain the outcome, the higher the information it brings.

We can generalize the definition of information in the case where two different outcomes E_1, E_2, with given probabilities P_1 and P_2, are possible. To gain some intuition on the problem, we ask ourselves which are the probability assignments that make the outcome more uncertain (i.e. maximize the information). If P_1 and P_2 are largely different, for instance $P_1 = 0.999$ and $P_2 = 0.001$, we are allowed to believe that event E_1 will occur almost certainly, considering E_2 to be a very implausible outcome. The information content will be very low. On the other hand, most unpredictable situation happens when

$$P_1 = P_2 = \frac{1}{2} :$$

this describes a situation of 'maximum ignorance' and the information content of such system must be high. Any generalization of Eq. (5), must then have its maximum when $P_1 = P_2$. For N events, the system with the highest possible information content is when:

$$P_1 = \ldots = P_N = \frac{1}{N} :$$

Shannon [11] showed that the only functional form satisfying continuity with respect to its parameters, additivity and that has a maximum for equal probability events is:

$$H[p_1, \ldots, p_N] = -\sum_{i=1}^{N} p_i \log p_i, \tag{6}$$

which can be interpreted as the 'expected information' brought by an experiment with N possible outcomes each with its own probability p_i. In the continuous case:

$$H[p(x)] = -\int p(x) \ln p(x) dx, \tag{7}$$

We call the functional H information entropy.[2]

We now turn to the core of our problem: how can we assign probabilities to a set of events keeping into account our knowledge of the system and, at the same time, ensure it is non-committal towards unavailable knowledge? The "knowledge" at our disposal about the system under investigation is what we define 'evidence' and any probability assignment is given such evidence, in agreement with Cox [12] construction of probability. In the case above, our knowledge on the system is only the total number N of different outcomes – this is a minimal requirement. Of course, more complex evidence constraints can be applied.

It is very common that the constraints provided by the evidence are not enough for setting the probabilities for each event: in this case, it is reasonable to assume that the probability assignment should make the experiment as unpredictable as possible.[3] In other words, the information entropy content introduced by the probability assignment should be as large

[2] In defining the information entropy as in Eq. (7), we are implicitly assuming a uniform measure over the parameter space. In case of a non-uniform measure $m(x)$, the definition generalises to $H[p(x)] = -\int p(x) \ln \frac{p(x)}{m(x)} dx$.

[3] In [9] this statement is made more precise and justified more thoroughly, with arguments based on combinatorial analysis.

as possible, in accordance with the available evidence. MAXENT formalises this reasoning by stating that probabilities should be assigned by maximizing uncertainty (information entropy) using evidence as a constraint. This defines a variational problem, where the information entropy functional $H[p_1, \ldots, p_N]$, defined in Eq. (6), has to be maximized.

The maximisation of the entropy, supplemented by evidence in the form of constraints to which the sought-for probability distribution must obey, gives rise to several of the most common probability distributions commonly employed in statistics. In the cases of interest, evidence is used to constraint, via Lagrange Multipliers, the momenta of the probability distribuiton we are seeking to evaluate. For instance, whenever the only constraint available is the normalization of the probability distribution (i.e. no evidence is available), the entropy is maximised by the uniform distribution. If we have evidence to constraint the expected value, the information entropy is maximised by the exponential distribution.

Of particular relevance for our purposes is the case in which, in addition to the mean, also the variance is known: MAXENT leads to the Gaussian distribution. This derivation is particularly interesting from the foundational point of view, since it provides a deeper insight into the ubiquitous Gaussian distribution. Indeed, it is not only the limit distribution provided by the central limit theorem for finite variance processes but it is also the distribution that maximizes the entropy for a fixed mean and variance: from the MAXENT principle, it is the correct probability distribution to assign if the mean and covariance are the only quantities that fully define our process. In some sense, we can interpret the central limit theorem as the natural 'statistical' evolution toward a configuration that maximizes entropy in repeated experiments.

For this work, we are especially interested in the multidimensional case. Suppose we have a vector of measurements $(x(t_1), \ldots, x(t_n)) = (x_1, \ldots, x_n)$ that we conveniently express as a single realization of an unknown stochastic process $x(t)$ and we have information about the expectation value of the process $\mu(t)$ and on the matrix of autocovariances $C_{ij} \equiv C(t_i, t_j)$, then the MAXENT distribution is the n-dimensional multivariate Gaussian distribution [13]:

$$p\left((x_1, \ldots, x_n)|I\right) = \frac{1}{(2\pi \det C)^{k/2}} \exp\left(-\frac{1}{2}\sum_{i,j}(x_i-\mu_i)(x_j-\mu_j)C^{-1}_{ij}\right). \tag{8}$$

For a wide-sense stationary process the mean function is independent of time, hence it can be redefined to be equal to zero without loss of generality, and the auto-covariance function is dependent only on the time lag $\tau \equiv t_i - t_j$. One can thus choose a sampling rate Δt so that $C_{ij} = C((i-j)\Delta t)$. The autocovariance matrix thus becomes a Toeplitz matrix[4]. Toeplitz matrices are asymptotically equivalent to circulant matrices and thus diagonalized by the discrete Fourier transform base [14]. Some simple algebra shows that the time-domain multivariate Gaussian can be transformed into the equivalent frequency domain probability distribution:

$$p\left((\tilde{x}_1, \ldots, \tilde{x}_{n/2})|I\right) = \frac{1}{(2\pi \det S)^{n/2}} \exp\left(-\frac{1}{2}\sum_{ij}\tilde{x}_i S^{-1}_{ij}\tilde{x}_j\right), \tag{9}$$

where the matrix $S_{ij} = S_i\delta_{ij}$ is an $n \times n$ diagonal matrix whose elements are the PSD $S(f)$ calculated at frequency f_i. Many readers will recognise the familiar form of the Whittle likelihood that stands at the basis of the *matched filter* method [15] and of gravitational waves data analysis, [1,16, e.g.]. Thanks to MAXENT, the problem of defining the probability distribution describing a wide-sense stationary process is thus entirely reduced to the estimation of the PSD or, equivalently, the autocovariance function.

2.2 Maximum entropy spectral analysis

In principle, if the autocorrelation was known exactly (i.e. at every time $\tau \in (-\infty, +\infty)$), the computation of the PSD would reduce to a single Fourier transform (i.e. Eq. (2)). However, in any realistic setting, we are dealing with a finite number of samples N from the process. In such cases, the single periodogram is not a consistent estimator for the power spectral density, since its variance doesn't decrease when the sample size increases. Moreover, the error σ_k in the estimate of the autocorrelation after k steps increases as $\sigma \sim 1/\sqrt{N-k}$,[5] so that only few values for the autocorrelation function can actually be computed reliably. This bring us to the core of the problem: how to give an estimate from partial (and noisy) knowledge of the autocorrelation function? MAXENT can guide us in this task without any a priori assumptions on the unavailable data.[6]

[4] We remind the reader that a Toeplitz matrix is a matrix in the form:

$$\begin{pmatrix} a_0 & a_1 & a_2 & \ldots & \ldots & \ldots & a_n \\ a_{-1} & a_0 & a_1 & \ldots & \ldots & \ldots & a_{n-1} \\ a_{-2} & a_{-1} & a_0 & \ldots & \ldots & \ldots & a_{n-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{-n+1} & \ldots & \ldots & \ldots & a_{-1} & a_0 & a_1 \\ a_{-n} & \ldots & \ldots & \ldots & a_{-2} & a_{-1} & a_0 \end{pmatrix}$$

[5] This is easily understood: when computing the autocorrelation at order k, only $N-k$ examples of the product $x_i x_{t+k}$ are available and the variance of the average value goes as the inverse of the square root of the points considered.

[6] Indeed this is the largest difference with the most common Welch method. The latter assumes that the unknown values of the autocorrelation are 0. Clearly, this assumption is unjustified and MAXENT is a good way to relax this assumption.

As in the previous examples, one needs to set up a variational problem where the entropy, Eq. (7), is maximized subject to some problem-specific constraints. In our case, they are (i) the PSD estimate has to be non-negative; (ii) its Fourier transform has to match the sample autocorrelation (wherever an estimate of this is available).

Before doing so, there is a technicality to solve: the definition of entropy depends on a probability distribution, not on the PSD. It can be shown [17,18, e.g.] that the variational problem can be formulated in terms of the power spectral density $S(f)$ alone by considering our signal as the result of the filtering a white noise process using a filter with transfer function $T(f)$ equal to $S(f)$.[7] The difference in entropy between the input and the output time series (i.e. the entropy gain) obtained by such filter applied on white noise is:

$$\Delta H = \int_{-Ny}^{Ny} \log S(f) df. \tag{10}$$

where Δt is sampling rate and $Ny \equiv \frac{1}{2\Delta t}$ is the Nyquist frequency. Thus maximising Eq. (10) is equivalent to maximizing Eq. (7).

Before maximizing the entropy gain, we need to include the evidence available as a form of mathematical constraints for the assignment of $S(f)$. This is equivalent in imposing that the variational solution $S(f)$ for the PSD matches the empirical autocorrelation. Let us define a realization of a stochastic process $(x_1, \ldots, x_N)$ with sample autocorrelations $\bar{r}_k$, $k = 0, \ldots, N/2$, then the PSD must satisfy the following equation:

$$\int_{-Ny}^{Ny} S(f) e^{i 2\pi f k \Delta t} df = \bar{r}_k. \tag{11}$$

Thus, by maximizing Eq. (10) with constraints in Eq. (11), we can give an estimate of the spectrum given a time series sample. This approach on PSD computation provides a result consistent with the empirical autocorrelation function whenever this is available and, at the same time, it does not make any assumption for the unavailable estimates for the autocorrelation at large time lags.

Remarkably, the variational problem admits a closed-form analytical expression for $S(f)$. The expression was first found by [10]:

$$S(f) = \frac{P_N \Delta t}{\left(\sum_{s=0}^{N} a_s z^s\right)\left(\sum_{s=0}^{N} a_s^* z^{-s}\right)}, \tag{12}$$

where Δt is the sampling interval of the time series, $z = \exp(2\pi i f \Delta t)$, $a_0 = 1$. The vector obtained as $(1, a_1, \ldots, a_N)$ is also known as the *prediction error filter*. The coefficients $a_s (s > 0)$, together with an overall multiplicative scale factor P_N, are to be determined by an iterative process (called Burg's algorithm) At least, two implementations of Burg's algorithm are available in the literature, labeled as 'Standard' and 'Fast' in the `memspectrum` package. The 'Standard' method is slower but more stable, while 'Fast' trades stability for speed. On simulated stationary data, both versions typically yield similar results, while our tests with real gravitational waves data seems to indicate that the 'Fast' implementation introduces noise into the PSD estimate.[8] A comparison of the computational times for Standard MESA implementation and Fast implementation (together with Welch's) is provided in Appendix C.

The number N of such coefficients is a choice that shall be made by the user and indeed it is the only hyperparameter that needs to be tuned. The details of the derivation and the actual form for the coefficients a_s can be found in Appendix A.

2.3 Autoregressive process analogy

The application of MESA is not limited to spectral estimates, but it also provides a link between spectral analysis and the study of autoregressive processes (AR) [19]. An autoregressive stationary process of order p, AR(p), is a time series whose values satisfy the following expression:

$$x_t - b_1 x_{t-1} - b_2 x_{t-2} \ldots b_p x_{t-p} = \nu_t \tag{13}$$

where $b_1, \ldots, b_p$ are real coefficients and ν_t is white noise with a given variance σ^2. Thus, an AR(p) process models the dependence of the value of the process at time t from the last p observations, thus being potentially able to model complex autocorrelation structures within observations.

Thanks to Wold's theorem [20], every stationary time series can be represented as an autoregressive process: this ensures that maximum entropy estimation is faithful and general; it turns out that the maximum entropy principle provides a representation of the time series as an $AR(p)$ process and Burg's algorithm computes the corresponding autoregressive coefficients that are suitable to model the available data.

To show the analogy, we compute the PSD $S_{AR(p)}$ of an $AR(p)$ process and we show that it is formally equivalent to the PSD obtained in Eq. (12). This will also provide a direct expression for the autoregressive coefficients b_i and

[7] A filter with transfer function $T(f)$ takes in input a time series x_t and outputs a times series y_t such that:

$$T(f) = \frac{\tilde{y}(f)}{\tilde{x}(f)}$$

where $\tilde{x}(f)$ denotes the Fourier transform of x_t (and similarly for y_t)

[8] For this reason, it is advisable to use the 'Standard' implementation whenever possible. In most case of numerical instability in the 'Fast' method, `memspectrum` will send a warning to user.

for the noise variance σ^2. We start taking the z transform[9] of Eq. (13):

$$\sum_t x_t z^t - \sum_i b_i z^i \sum_t x_{t-i} z^{t-i} = \sum_t \nu_t z^t \,. \tag{14}$$

Calling $\tilde{x}(z)$ and $\tilde{\nu}(z)$, the transformed quantities, in the z domain, the process takes the form:

$$\tilde{x}(z) = \frac{\tilde{\nu}(z)}{\left(1 - \sum_{n=1}^{p} b_n z^n\right)}. \tag{15}$$

Since we assumed a wide-sense stationary process, $\tilde{x}(z)$ is analytic both on and inside the unit circle. Taking its square value and evaluating it on the unit circle $z = e^{-\imath 2\pi f \Delta t}$, from the definition of spectral density one obtains:

$$S_{AR(p)}(f) = |\tilde{x}(z)|^2 = \frac{|\tilde{\nu}(f)|^2}{\left|1 - \sum_{n=1}^{p} b_n e^{\imath 2\pi f n \Delta t}\right|^2}. \tag{16}$$

The numerator is the spectral density of white noise ν_t, i.e. its (constant) variance σ^2.

Equations (16) and (12) are equivalent, if we identify $b_i = -a_i$ and $P_N \Delta t = \sigma^2$. This shows that the MAXENT estimation of the PSD models the observed times series as an AR process and provides a *fit* for the autoregressive coefficients. Furthermore, as a consequence of Wold's theorem, there is the theoretical guarantee that every stationary time series can be modelled faithfully by the MAXENT.

2.4 Forecasting

The link between MESA and AR processes is of particular interest. Given the solution to Burg's recursion to determine the a_k, we automatically obtain the coefficients of the equivalent AR process, hence we are able to exploit Eq. 13 to perform *forecasting*, thus providing plausible future observations, conditioned on the observed data. Indeed, for an AR(p) process the conditional probability $p(x_t|x_{t-1}, \ldots, x_{t-p})$ of the observation at time t with respect to the past p observation has the form:

$$p(x_t|x_{t-1}, \ldots, x_{t-p}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_t - \sum_{i=1}^{p} b_i x_{t-i}}{\sigma}\right)^2\right]. \tag{17}$$

The interpretation of Eq. (17) is straightforward: x_t follows a Gaussian distribution with a fixed variance and a mean value $m_t = \sum_{i=1}^{p} b_i x_{t-i}$ computed from past observations. Equation (17) provides then a well defined probability framework for predicting future observations: this is a very useful feature of MESA, that does not have an equivalent in any other spectral analysis computation methods.

2.5 Whitening

The theory of the AR processes can be also applied to the problem of whitening a time series. Given a time series, x_t, the whitening operation produces another time series x_t^W such that:

$$x_t^W = \mathcal{F}^{-1}\left[\frac{\tilde{x}(f)}{\sqrt{S(f)}}\right] \tag{18}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform of a frequency series. If x_t is a realization of gaussian noise (see Eq. (9)) with PSD $S(F)$, the whitened time series x_t^W is just white noise (i.e. uncorrelated samples from a normal gaussian).

From Eq. (13), remembering that $b_i = -a_i$, it's straightforward to derive an expression for the whitened time series x_t^W:

$$x_t^W = \frac{1}{\sqrt{P_N}} \sum_{i=0}^{p} a_i x_{t-i} \tag{19}$$

This amounts to a convolution of the time series x_t with the kernel $(1, a_1, \ldots, a_p)$, plus a variance rescaling. Performing a convolution is an appealing alternative to evaluating Eq. (18) directly.

3 Validation of the model

MESA provides a recursive formula for computing the coefficients a_k in Eq. (12). The number M of such coefficients is equivalent to the maximum order of the autocorrelation $\bar{r}_m$ considered. In an ideal scenario, this would be equal to the number of points the autocorrelation is computed at (equivalent to the length of data considered). However, the computation of high order coefficients of the autocorrelation is unstable and for high enough m, as the estimation for $\bar{r}_m$ shows a very high variance, broadly scaling as $\sim \left(\sqrt{M-m}\right)^{-1}$.

It is then clear that the choice of the number of samples of the discrete autocorrelation to consider is important: on the one hand it is advisable to include as much knowledge of the autocorrelation as possible, leading to include all the known $\bar{r}_m$; on the other hand, including values of the autocorrelation that are not reliably estimated, can be counterproductive. The order M of the autocorrelation to be considered (or, equivalently, the order M of the underlying autoregressive process) is the only tuning parameter of MESA and a careful balance between these two necessities must be made when applying the algorithm.

[9] The z transform is the discrete-time equivalent of the Laplace transform, thus taking a discrete time-series and returning a complex frequency series.

The remainder of this section is devoted to an extensive study on how to make such choice. In Sect. 3.1, we are going to define two different *loss functions* to measure how well the algorithm is able to reproduce a known PSD. The basic idea is to validate, as the autoregressive order considered increases, the performance of the algorithm results by measuring the loss function and pick, among the orders the one that yields better results. The performance of the different losses will be assessed by answering to two questions: (i) how well the AR order is recovered and (ii) how well the measured PSD is able to whiten the input time series. This will be discussed Sects. 3.3 and 3.4.

3.1 Choice of the autoregressive order

Guided from numerical experiments, an indication on the upper bound to the autoregressive order M_{max} is [21]:

$$M_{max} = 2N/\ln(2N), \tag{20}$$

where N is the number of observed points in the time-series. However, this is just a plausible upper limit on the order of the AR process m and the optimal algorithm could employ fewer points. We then need a more sophisticated method for computing the right value for m. We summarise them below:

- **Final prediction error** The first criterion is due to [22]. It was proposed that m should be chosen as the length that minimizes the error when the filter is used as a predictor, the *final prediction* error (FPE):

$$FPE(m) = \mathbb{E}\left[\left((x_t - \hat{x}_t)^2\right)\right] \tag{21}$$

  with $\hat{x}_t = \sum_{i=1}^{M} a_i x_{t-i}$. Asymptotically minimizing FPE is equivalent to minimizing the quantity:

$$\mathcal{L}_{\mathrm{FPE}}(m) = P_m \frac{N+m+1}{N-m-1} \tag{22}$$

  with P_m being the estimated noise variance at order m, see Eq. (33). In the $N \to \infty$ limit, remembering $m_{max} \sim 2N/\log(2N)$, Akaike's loss function is equivalent to the minimization of the variance P_m of the white noise of the underlying $AR(p)$ model.
- **Variance maximum (VM)** This second criterion [23] is based on a similar assumptions to FPE. It minimises the actual value of the least squares (instead of relying to asymptotical behaviour), using a normalising factor that takes into account the m degrees of freedom necessary to estimate the forward prediction error filter a_k.

  The quantity to be minimised is

$$VM(m) = \frac{1}{N-2m} \sum_{t=m}^{N} \left(x_t - \sum_{i=1}^{m} a_i x_{t-i} \right)^2 \tag{23}$$

  The package implementation of VM loss function takes advantage of a recursive re-writing of the above formula, as in Eqs. (27) and (28) of [24].
  Several other criteria are available in the literature [25,26] and some are implemented in the `memspectrum` package. We don't report them in this paper since they didn't show any additional merit with respect to the aforementioned loss functions

Once a loss function is selected, the choice of the best recursion order is straightforward: we solve the Levinson recursion [27] until M_{max}, as given in Eq. (20), iterations are reached. Then, the order m is selected to be the one that minimizes the specified loss function.

In a real implementation of the algorithm, computing all the recursion up to M_{max} can result in a significant waste of computational power: the optimal value is often $m_{opt} << M_{max}$ and, in such cases, computing all the values of m until M_{max} is not useful. In practice, we can apply an *early stop* procedure: every few iterations we look for the best order of m_{opt}; if this value does not change for a while, we assume that a good (local) minimum of the loss function is found and the computation is stopped.

The following sections will be devoted to the study of the statistical properties of the loss functions introduced above: we need to understand which choice provides the best quality in the reproduction of some known power spectral densities. In the following paragraph, we will discuss three different comparison (one qualitative and two quantitative) of the two proposed loss functions.

3.2 How accurate are the reconstructed PSDs?

In our initial qualitative comparison, depicted in Fig. 1, we juxtapose the reconstruction of a known a-priori power spectral density with those obtained using the two distinct loss functions. The black, dotted line in the plot represents our chosen reference PSD, released together with the GWTC-1 catalog [28,29], and computed for the LIGO Handford with the BayesLine package [5,6,30,31]. The PSD is the median over the reconstructed posterior distribution for the GW150914 event.

To conduct this analysis, we generated 1000 noise time-series whose power spectral densities matches the reference PSD by construction [32]. The sampling rate and observation time were fixed at $df = 2048$ Hz and $T = 5$ s, respectively. For each noise realization, we employed both the FPE and

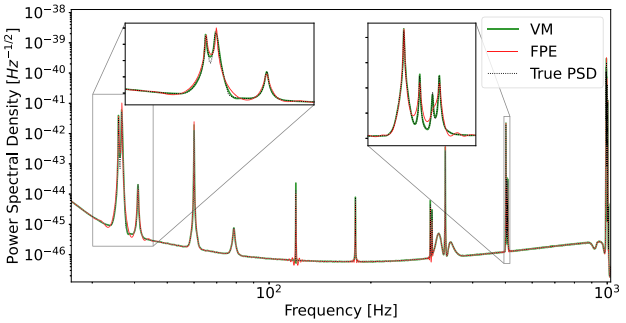

Fig. 1 Comparison of the ensemble average PSD estimate

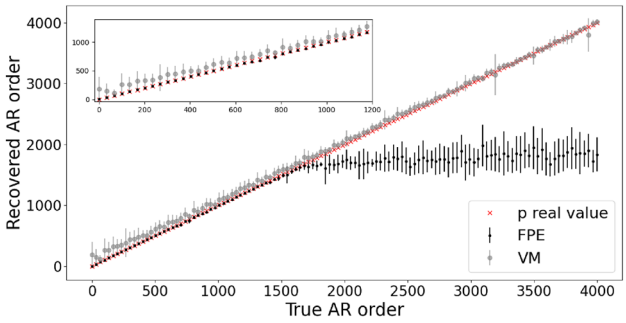

Fig. 2 Reconstructed value for the autoregressive order plotted against the true value of the autoregressive order. The reconstructed autoregressive orders are computed from a time series randomly drawn with an $AR(p)$ model, with the two different loss functions under investigation

the VM loss functions to estimate the PSD. Ultimately, we compared the reference PSD against the ensemble average of these two estimation methods.

The FPE-derived estimate, represented by the red line, effectively identifies and reconstructs peaks across both high and low frequency ranges with commendable accuracy. However, as illustrated in the inset plots, FPE struggles when confronted with structured peaks-those containing subordinate modes. In such cases, FPE accurately captures the primary mode but overlooks the subsidiary peaks.

On the other hand, the VM estimate, depicted as the continuous green line, excels in reconstructing both dominant and subordinate modes with remarkable precision. VM appears to prioritize comprehensive mode reconstruction, while FPE emphasize an accurate reconstruction of major modes while potentially neglecting more intricate sub-peaks. Additional figures of merit are inserted in Appendix B

3.3 How well is the AR order recovered?

Moving to our second comparison, we now focus on another crucial aspect: how accurately each loss function estimates the autoregressive (AR) order, which represents the number of employed a_k coefficients.

Here, the `memspectrum` package proves quite useful. It allows us to assign a specific order to the reconstructed autoregressive filter and use the resulting coefficients to forecast time series. With these tools in hand, we generated various time series, each with a different autoregressive order ranging from m = 0 to m = 4000.

To ensure reliability, we created 30 distinct time series for each autoregressive order. This approach lets us compute both the mean and variance, giving us insights into the accuracy of each loss function's order estimation. This analysis provides valuable information about how well each method performs in estimating the Autoregressive order across a broad spectrum of scenarios.

The results are reported in Fig. 2. The injected autoregressive order's true value is depicted by the red line. The estimations yielded by the two loss functions are illustrated alongside, accompanied by error bars indicating one standard deviation.

The plot reveals two distinct regions: one with "short" autoregressive orders (m = 0 to around m = 1600) and another with "long" autoregressive orders (starting from m = 1600).

In the first region, both loss functions provide comparable results that generally match the actual autoregressive order. FPE performs particularly well, offering estimates close to the injected order and with minimal error bars. VM performs slightly worse than FPE in this range, overestimating complexity and showing larger error bars.

Moving into the second region ($m > 1600$), a shift in performance becomes apparent. FPE's estimates tend to stabilize at a certain autoregressive value. However, as the injected model becomes more complex beyond this point, FPE's accuracy in recovering the true order diminishes, and its variance increases. In contrast, VM performs better in this range, closely following the actual behavior and consistently recovering the true order within one standard deviation. To conclude, VM appears to prioritize complexity in its approach. In contrast, FPE seems to lean toward synthesis, emphasizing accurate reconstruction of not too complex models.

3.4 How well can MESA whiten the data?

In Sect. 2.5, we showed how autoregressive coefficients and noise variance estimate P can jointly be used to create a whitening filter, as in Eq. (18) and. To complete our investigation, we compare how well these whitening filters work when obtained from the two different loss functions we've been studying.

For this test, we employed the same set of time-series data as described in Sect. 3.2. Each time series underwent the whitening process using the autoregressive filter derived from its corresponding loss function. We then evaluated the resulting whitened time series against a zero-mean, univariate normal distribution using the Anderson–Darling test.

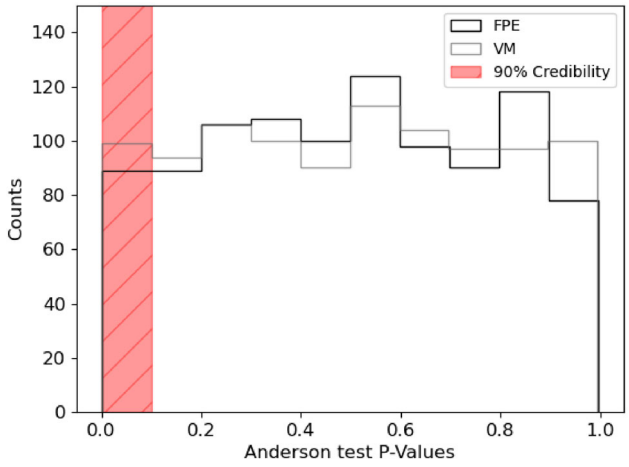

Fig. 3 Histogram of the P-values obtained with an Anderson Darling test on the whitened time series, against a univariate, 0 mean normal distirbuition

The results are reported as an histogram for the obtained p-values in Fig. 3 together with the chosen critical region of $p < 0.1$, representing a 90% confidence level. In this region, there is no statistical difference between the two. Infact, the total number of counts c in this bin are respectively $c_{VM} = 100 \pm 10$ and $c_{FPE} = 89 \pm 9$, affirming the absence of a pronounced discrepancy between the two. In essence, this final examination underscores a shared proficiency in whitening between the two loss functions, showing that very long filters are not needed to obtain a comparlable result in whitening. Both methods showcase comparable results for whitening scopes.

From our previous discussions, it's evident that both FPE and VM have their own strengths, and the choice between them greatly depends on the specific analysis requirements. In our analysis, VM tends to provide more accurate PSD estimates and often results in longer autoregressive filters. However, in cases where the underlying model is simple, there is a risk of VM overestimating complexity and generating patterns that don't truly reflect the data.

On the other hand, FPE is a good option for reconstructing processes without introducing unnecessary complexity. However, it might underestimate the complexity of the data, particularly in scenarios involving secondary peaks or in the low-frequency region.

Lastly, it's worth noting that FPE holds the advantage of lower numerical complexity due to its straightforward calculations involving simple arithmetic. In contrast, VM requires more complex computations, dealing with arrays that might be very long depending on the analysed data.

4 Comparison with Welch method

We perform a *qualitative* comparison between the performance of the MESA and of the standard Welch algorithm. In this, we cannot avoid to be only qualitative. Indeed, as the results of the comparison are problem dependent, it is very hard to quantify this in a single metric. Although similar studies can be drawn from any other PSD, in this section we focus on a single PSD and we try to generalize some observations that we make. We used the same reference PSD used for the comparison of the two losses in the previous section [5,6,30,31].

We simulate data[10] from the PSD used for the analysis of the event GW150914 and we employ both Welch's method and MESA to estimate the spectrum. We vary the length of the data used for the estimation: this is also useful to assess how the computation depends on the data available. We set the total observation time $T = 1, 5, 8, 10, 100, 1000$ s. The observation time of $8\,s$ is inserted since it is the observation time over which the reference PSD is computed. For the MESA algorithm, we choose the VM loss function. For the Welch algorithm, we employ a Tukey window with the shape parameter α equal to 0.4 (see scipy documentation), an overlap fraction of 1/2 for the segments and a length of segments $L = 512, 1024, 2048, 8192, 32768$ points, depending on the observation time. In all cases, the sampling rate is set to 4096 Hz. For the Welch algorithm, we use the standard implementation provided by the python library `scipy` [33,34]. The results from both methods are summarized in Figs. 4 and 5 respectively.

First of all, we note that using a longer time series results in a better estimation of the PSD, especially at low frequencies. This is somehow obvious: longer data streams probe lower frequencies thanks to Nyquist's theorem as well as providing better estimates for the FFT, in the Welch case, and the sample autocorrelation, for MESA.

We also note that MESA converges (Figs. 4 and 5) to the underlying spectrum much faster than Welch's method, providing a better estimate even in the case of short time series. Although observed at every frequency, this behaviour is more evident in the low frequency region. An accurate profile reconstruction can be obtained with MESA using a 5 s-strain only, while Welch method requires at least 10 s of data to obtain a comparable profile". Such a difference persists when the 8 s data strain is considered. Furthermore, MESA is able to model all the details of the peak at around ~ 40 Hz (even with $T = 100$ s), while the Welch's algorithm fails to do so even with an observation time of $T = 1000$ s.

Another important element is the noise of the spectral estimation: we find that the PSD estimation provided by the

[10] This is to ensure that we have a baseline PSD to compare the data with.

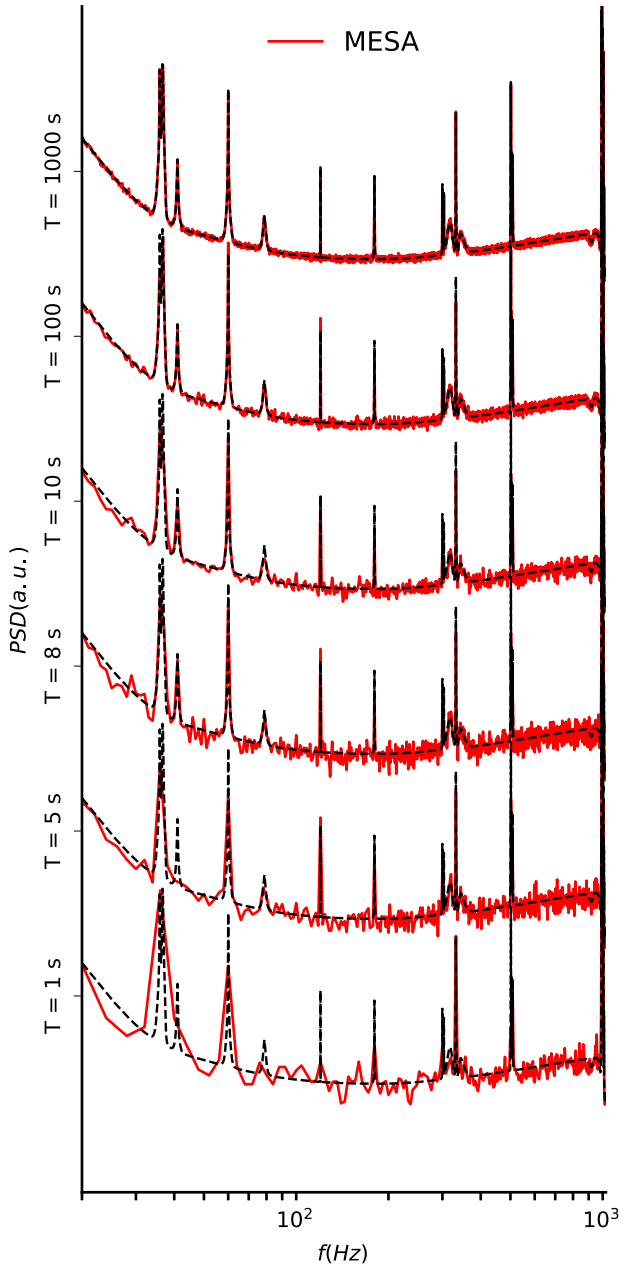

Fig. 4 Comparison between analytic (dashed line) and estimated (red line) spectrum. The estimation is performed with Maximum Entropy method on *synthetic* data, with an increasing observation time $T = 1, 5, 8, 10, 100, 1000$ s

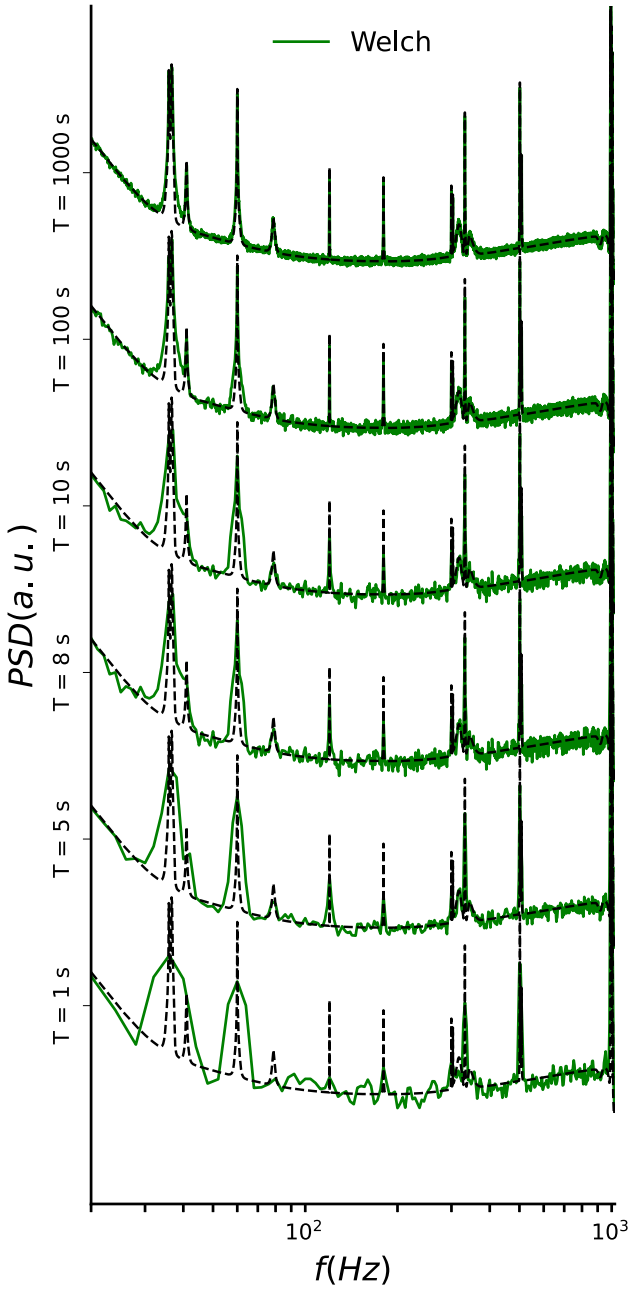

Fig. 5 Comparison between analytic (dashed line) and estimated (green line) spectrum. The estimation is performed with Welch's method on *synthetic* data with an increasing observation time $T = 1, 5, 10, 100, 1000$ s

Welch's method is noisier (i.e. has a large number of spurious peaks) compared to the PSD measured with MESA and FPE loss function. This is especially true at high frequencies and for long observation times T.

Finally, as already discussed Welch's method is very dependent on the choice of window function. A Tukey window with aforementioned parameters is what we found to be the best compromise between noise and accuracy for the reconstruction, but different choices can be made, possibly providing more accurate results than the ones reported here. However, we want to stress that this fact does not invalidate our discussion but reinforces it: one of the most appealing advantages of MESA is the minimal amount of fine tuning required.

It would also be interesting to perform a detailed comparison between our results and the ones from the FastSpec algorithm [35], both in terms of operational speed and ability

to whiten GW data. Such study will be carried on in future work.

5 Marginalisation over the noise distribution: application to GW parameter estimation

Define the data hypothesis D as the statement that the data $D = S + N$ with S and N some deterministic signal and some noise hypotheses. Typically, in this formulation one is choosing both a functional form for the signal of interest $S \equiv$ "$h(t;\theta)$" and some parametric form $f(t)$ for the noise distribution $N \equiv$ "$n(t) \sim f(t)$". Some well established math then leads to the usual Bayesian framework for parameter estimation, see [36] for an application to gravitational wave physics. This procedure is very robust as long as the choice of noise distribution is indeed representative of the underlying process. Let us relax the N hypothesis by defining a *residuals* hypothesis R as $R = D - S$. This might seem a very trivial statement, but it has a non-trivial application: given $d(t) = h(t;\theta) + n(t)$ where $h(t;\theta)$ is our signal model, defined by a set of parameters θ, the residuals $r(t) \equiv d(t) - h(t;\theta)$. Formally, no reference to the noise process is present anymore. Under MAXENT, we can model $r(t) \sim \mathrm{AR}(k)$ with k the *unknown* order of the process to be inferred from the residuals, either via one of the aforementioned loss functions or even by marginalising over it while exploring the signal space. Moreover, we can *always* write $p(r(t)|N\,I)$ as in Eq. (9) once we know k, with the PSD given in Eq. (16), whatever the noise process actually is. In other words, we care only about maximising the information entropy in the distribution of the residuals.

Hence, as an application of MESA, and its implementation in `memspectrum`, we analyse GW150914 [37] using a Bayesian framework that allows for the marginalisation of the order k of the AR(k) process representing the residuals data stream. Although the inference is essentially unchanged compared to the standard case, see [36], there are some substantial modification to the likelihood construction. Since MESA is applicable to time-domain data, all calculations prior to the Fourier transform must be performed in time domain, thus increasing the computational cost by a non-negligible amount. We shall refer the time-of-arrival parameter t_c of the GW to the geocenter. At each iteration of the inference algorithm, we sample a vector $\theta \equiv \theta_{GW} \cup k$.[11] For each inteferometer j, therefore, we need to compute a time-delay Δt_j to compute the antenna response functions $F_{j,+}(t + \Delta t_j)$, $F_{j,\times}(t + \Delta t_j)$ as well as the correct time-shift for the GW template

$$h_j(t) = \sum_{p=+,\times} F_{j,p}(t + \Delta t_j) h_p(t + \Delta t_j; \theta_{GW}) \tag{24}$$

that we use to compute the time-domain residuals $r_j(t) = d_j(t) - h_j(t)$. We apply `memspectrum` to $r_j(t)$ with the *fixed* value of k and calculate the detector likelihood for $\tilde{r}_j$ using Eq. (9) and PSD as in Eq. (12). The coherent likelihood is then given by the product of the individual likelihoods. As our analysis template, we adopt the fast machine learning based MLGW model [38], an aligned spin model trained on TEOBResumS [39], that has been shown to perform well on LVK events detected during O1 and O2 [38]. Our sampler is a nested sampling algorithm [40] and the specific inference model is implemented as part of `granite`, a dedicated inference model for ground-based interferometric detectors. We compare our results with the combined posterior samples[12] available from GWOSC [41] and available at `https://zenodo.org/records/6513631`. As a comparison, we also performed an analysis using a fixed PSD estimated off-source with MESA, performing an averaging procedure analogous to what if performed in the Welch method.

In Figs. 6, 7 and 8 we show the posteriors for the set of intrinsic parameters, extrinsic parameters and reconstructed waveform from our analyses. The posterior samples in red show the results coming from marginalising over the PSD, the ones in green show the results from our fixed PSD anal-

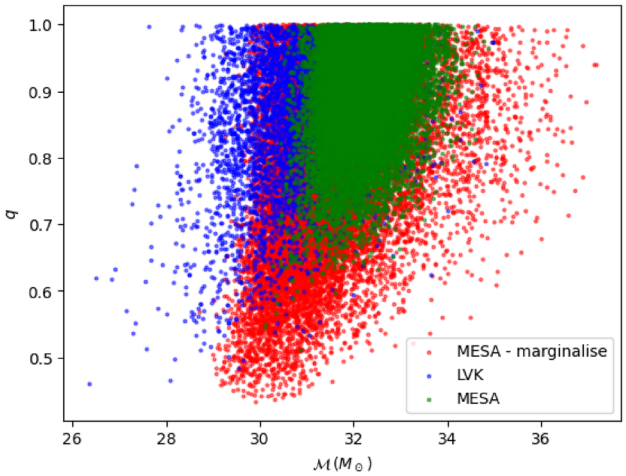

Fig. 6 Posterior samples for $\mathcal{M}$ and mass ratio q from the LVK (blue), using `memspectrum` to estimate with a fixed PSD (green) and using `memspectrum` to marginalise over the PSD (red). The samples are largely consistent among the models, with the MESA model providing a more conservative estimate

[11] We indicate the set of all GW parameters (component masses, spins, luminosity distance, etc.) with θ_{GW}.

[12] In particular, we compare against the posterior samples given in the file `IGWN-GWTC2p1-v2-GW150914_095045_PEDataRelease_mixed_nocosmo.h5`.

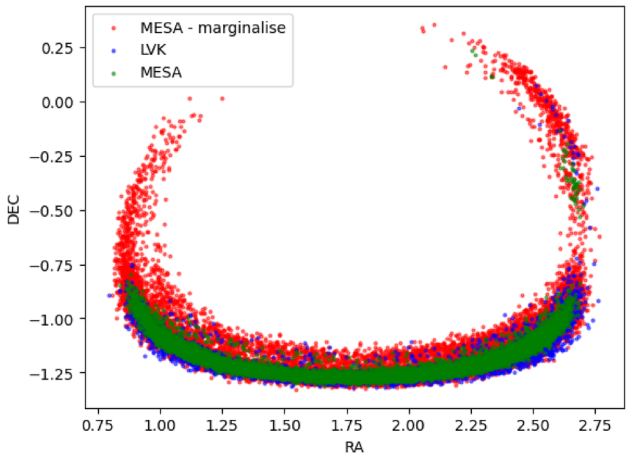

Fig. 7 Posterior samples for $\mathcal{M}$ and mass ratio q from the LVK (blue), using `memspectrum` to estimate with a fixed PSD (green) and using `memspectrum` to marginalise over the PSD (red). The samples are largely consistent among the models, with the MESA model providing a more conservative estimate

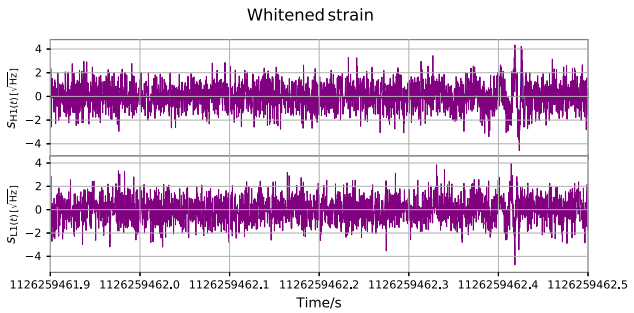

Fig. 8 Whitened reconstructed waveforms and data from our analysis for the Hanford detector (top panel) and the Livingston detector (bottom panel). The shaded turquoise area indicates the 90% credible region over the waveforms space while the purple contours indicate the 90% credible regions over the whitened data

ysis, while the ones in blue are the samples release by the LVK. While the fixed PSD run show posteriors that are largely consistent with the LVK results, the posteriors from the marginalisation run are still in general consistent with what has been released by the LVK, but showing wider credible regions. This is expected since our likelihood includes additional uncertainty due to the explicit sampling over the process order, hence the PSD. For the particular 4 s of data, sampled at 4096 Hz, the recovered orders are $k_{H1} = 1107^{+9}_{-5}$ and $k_{L1} = 1146^{+8}_{-8}$, Fig. 9. The corresponding PSDs and uncertainties are shown in Fig. 10. The full joint posterior distribution recovered when marginalising over the AR order is shown in Appendix D.

We conclude this section with a comparison with Ref. [42], where a similar attempt at characterising the effect of marginalising over the PSD was made. The aforementioned work adopted a very different methodology compared to ours, the author use independent draws from the posterior over the PSD as inferred by BayesLine and repeat a

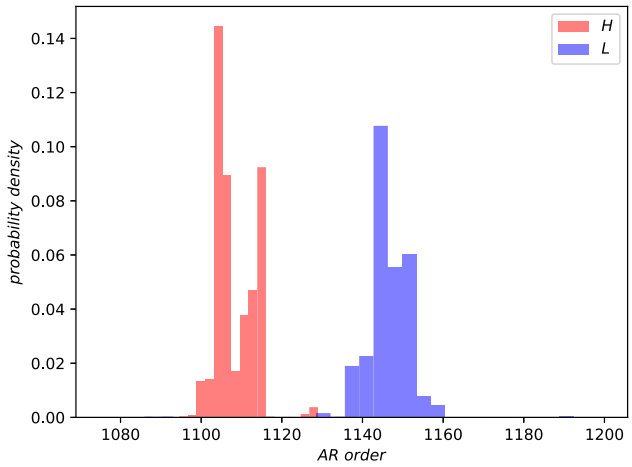

Fig. 9 Posterior distributions for AR process orders in the Hanford (red) and Livingston (blue)

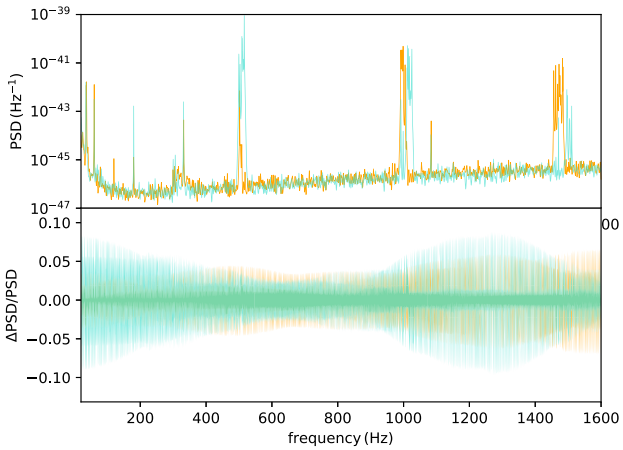

Fig. 10 Top panel: posterior PSDs for the Hanford (orange line) and Livingstone (blue line) as inferred by our analysis. Bottom panel: relative uncertainty around the median for both the Hanford (blue) and the Livingston (orange) PSDs

standard frequency domain analysis for several signals. The results reported in Ref. [42] show a similar trend to ours: they observe a general widening of the sky position posteriors, and little effect (few %) on the intrinsic parameters width, while we observe a substantial increase in the width of the chirp mass and mass ratio posteriors. We believe this difference to be due both to the very different model we employed as well as on the small (200) number of samples from the PSD they used. The estimated uncertainty over the PSD is however consistent, of the order of (5–10%).

6 Summary and discussion

We presented a case study of the application of Maximum Entropy principle to the realm of spectral estimation. Albeit the methodology hereby presented is grounded on solid the-

oretical foundations and its merits are widely recognised, Maximum Entropy methods have yet to be adopted routinely in the study of problems related to time series. The superior nature of maximum entropy methods, and in particular of Burg's method, is exemplified by the closed form estimate of the power spectral density and by the theoretical bridge between spectral analysis and AR processes. Moreover, the method presents, in our view, two main advantages when compared with more traditional ones; first there is no need to choose an arbitrary window function to correct the data and, second it provides as straightforward way to compute predictions given past observations. Accompanying this work, we provide a publicly available `Python` implementation, called `memspectrum`, that we used to perform the numerical studies presented in this work.

Since the order of the AR process is not yet determined by the theory, we opted for an in-depth investigation of several proposals in the literature and found that different loss functions are required for different situations, with the FPE loss function being the most indicated to deal with gravitational wave data. Along these lines, we directly compared the PSDs computed with MESA with the canonical Welch's algorithm. As outlined in Sect. 4, MESA provides PSD estimates with smaller variance and better accuracy than Welch algorithm. The use of MESA is particularly useful for short time series samples, where Welch's method is outperformed in both precision and confidence. As an examples, Figs. 4 and 5 illustrate that MESA's performance over a 8-second interval is more closely aligned with Welch's performance over a 100-second interval than Welch's performance over the 8-second interval alone. This is due to the better variance-bias tradeoff provided by MESA. Longer segments are in fact obtained by a linear interpolation of the original PSD. For MESA, this procedure allows the computation of the estimate on a larger number of data points, thus reducing the variance of the estimate in both cases. For the Welch method, having longer intervals allows the use of longer sub-segments of the original data, further contributing to lowering the bias, especially in the low-frequency domain.

This observation suggests a promising avenue to pursue in future developments of gravitational waves data analysis: for short time series, comparable with the length of binary black hole systems as observed by LIGO, Virgo and KAGRA, the computational cost of MESA is moderate and the inferred PSD is an accurate representation of the true underlying PSD. By applying MESA to 4 s of data in correspondence to GW150914, we demonstrated that it is possible to simultaneously estimate the signal and noise parameters, hence effectively marginalise over the noise PSD, without the need to

- assume a specific functional form for the PSD;
- estimate the PSD in an off-source segment of data.

Both items are of particular interest for several reasons that we shall discuss in what follows. Several proposals exist in the literature attempt to marginalise over the PSD, mostly using a parametric model for the PSD [36,43,44]. MAXENT fixes the functional form for us exploiting the correspondence with AR processes, providing a one-parameter family of models that are particularly easy to sample, thus grounding the noise properties marginalisation in solid theoretical foundations and in an easy-to-use numerical implementation. The latter point is also particularly relevant, especially in the context of future GW detectors. Future detectors are in fact expected to be operating in the signal dominated regime, with several sources – potentially from different classes – constantly present within the detectors' data streams. In these cases the common procedure of estimating the PSD from off-sources segments is bound to fail and or provide biases inferences. MAXENT and MESA model and are relevant *only for the segment of data under consideration*, and make no assumptions over what is not part of the analysis. We believe, and we will show in a future study, that using MESA can be a natural solution for computing single-source posteriors whenever multiple sources are overlapping. This is possible since, in our formulation, everything that we did not label as *signal* will be part of the residuals, over which we apply MESA.

Furthermore, MESA provides a simple, but robust and quite accurate, albeit for short times, predictor for the time series. This fact is remarkable and can be used in time series analysis for several purposes. As an example, an anomaly detection pipeline could be built using the forecasts of MESA: the predictions can form a baseline to compare the actual observations with. Whenever the observed data are outside the expectations, an anomaly detection can be claimed. Of course such predictions can be done with a more accurate (perhaps nonlinear) model; however MESA has the advantage of being simple and fast to construct, while providing decent predictions. At the same time, several instruments present gaps in their data stream, for instance LISA is expected to show such gaps (e.g. [45] and references therein), MESA forecasting capabilities could be used to fill those gaps with predicted data from past observations. In conclusion, we reiterate that MESA is a theoretically sound, computationally feasible and reliable way of studying the properties of stochastic processes and we hope that the investigations presented in this work will further stimulate developments and applications of this method.

Acknowledgements We are grateful to S. Biscoveanu, D. Laghi, M. Maugeri, C. Rossi and S. Shore for useful comments and discussions. This research has made use of data, software and/or web tools obtained from the Gravitational Wave Open Science Center (https://www.gw-openscience.org/), a service of LIGO Laboratory, the LIGO Scientific Collaboration and the Virgo Collaboration. LIGO Laboratory and Advanced LIGO are funded by the United States National Sci-

ence Foundation (NSF) as well as the Science and Technology Facilities Council (STFC) of the United Kingdom, the Max-Planck-Society (MPS), and the State of Niedersachsen/Germany for support of the construction of Advanced LIGO and construction and operation of the GEO600 detector. Additional support for Advanced LIGO was provided by the Australian Research Council. Virgo is funded, through the European Gravitational Observatory (EGO), by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale di Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by institutions from Belgium, Germany, Greece, Hungary, Ireland, Japan, Monaco, Poland, Portugal, Spain.

Data Availability Statement My manuscript has associated data in a data repository. [Authors' comment: The GW data are publicly available at [gravitational waves open science center]: https://gwosc.org/.]

Code Availability Statement My manuscript has associated code/software in a data repository. [Authors' comment: The code/software [memspectrum] is available at https://pypi.org/project/memspectrum.]

Funded by SCOAP[3].

A Details of PSD computation

A.1 MESA solution

We derive the expression for the MAXENT spectral estimator following the approach proposed by [10]. Unlike the standard approach, we do not enforce the constraints in Eq. (11) with the standard Lagrange multipliers approach. We write instead the PSD $S(f)$ as the Fourier transform of the sample autocorrelation function:

$$S(f) = \frac{1}{2Ny} \sum_{n=-\infty}^{\infty} \bar{r}_n e^{-\imath 2\pi n \Delta t}, \tag{25}$$

and, plugging it in the entropy gain expression Eq. (10), we obtain:

$$\Delta H = \int_{-Ny}^{Ny} \log \left(\frac{1}{2Ny} \sum_{n=-\infty}^{\infty} \bar{r}_n e^{-\imath 2\pi f n \Delta t} \right) df. \tag{26}$$

Note that this expression already takes into account the constraints in Eq. (11).

We now introduce a set of coefficients λ_s, defined as the derivative of ΔH with respect to the autocorrelation function r_s. Explicitly they are:

$$\lambda_s := \frac{\delta H}{\delta \bar{r}_s} = \frac{1}{2Ny} \int_{-Ny}^{Ny} S(f)^{-1} e^{-\imath 2\pi f s \Delta t} df \tag{27}$$

and we will show that $S(f)^{-1}$ can be written as a Fourier expansion in terms of such coefficients. Then, the determination of the values for the λ_s uniquely solves the problem of power-spectral density estimation.

Some properties for the coefficients can be worked out easily. First, since $S(f)$ is real, the λ_s show the property

$$\lambda_s = \lambda_{-s}^*.$$

The second property is obtained considering that the autocorrelation function r_n can only be computed for a finite time interval $n \in [-N, N]$ and that the PSD estimation must not depend on the unavailable values r_n: this is part of the constraint in Eq. (11) This requirement can be implemented as:

$$\frac{\delta H}{\delta \bar{r}_s} = 0 \text{ for } |s| > N,$$

that means

$$\lambda_s = 0 \text{ for } |s| > N.$$

From Eq. (27) and from the properties above, is easily seen from the properties of the Fourier transform that $S(f)$ can be expressed via a Fourier series

$$S(f)^{-1} = \sum_{s=-N}^{N} \lambda_s e^{-\imath 2\pi f s \Delta t}. \tag{28}$$

Defining $z = e^{-\imath 2\pi f \Delta t}$ the previous Fourier expansion becomes a Laurent polynomial in z:

$$S(f)^{-1} = \lambda_0 + \sum_{s=1}^{N} \lambda_s z^s + \sum_{s=1}^{N} \lambda_s^* z^{-s}. \tag{29}$$

It is easy to show that if z_0 is a root for the polynomial $(z_0^*)^{-1}$ is also a root: for every root laying outside the unit circle there will be another root inside of it and vice-versa. These properties allow us to rewrite the Fourier expansion (29) as [46]:

$$S(f) = \frac{P_N \Delta t}{\left(\sum_{s=0}^{N} a_s z^z\right)\left(\sum_{s=0}^{N} a_s^* z^{-s}\right)} \tag{30}$$

with $a_0 = 1$ and Δt the uniform sampling interval for the time series. The vector obtained as $(1, a_1, \ldots, a_N)$ is the prediction error filter. The power spectral density $S(f)$ is uniquely determined if both the prediction error filter and P_N coefficients are computed.

To compute the a_s is convenient to plug into Eq. (11) the Laurent Polynomial exansion for $S(f)$ Eq. (30) and then integrating over z (taking values on $\mathbb{S}^1$). In this way the equation

becomes:

$$\frac{P_N}{2\pi\imath}\oint_{\mathbb{S}^1}\frac{z^{-s-1}}{\sum_{n=0}^{N}a_n z^n\sum_{n=0}^{N}a_n^* z^{-n}}dz=\bar{r}_s. \tag{31}$$

Substituting $s\to s-r$, multiplying by a_s^* and summing over s, the previous equation becomes

$$\sum_{s=0}^{N}a_s\bar{r}_{s-r}=\frac{P_N}{2\pi\imath}\oint\frac{z^{r-1}}{\sum_{s=0}^{N}a_s z^s}dz \tag{32}$$

For a wide-sense stationary processes, all the poles lay outside the unit circle so that the previous integral can be easily computed obtaining the following, well known, equations:

$$\sum_{s=0}^{N}a_s\bar{r}_{r-s}=P_N \quad \text{if } r=0 \tag{33}$$

$$\sum_{s=0}^{N}a_s\bar{r}_{r-s}=0 \quad \text{if } r\neq 0. \tag{34}$$

A.2 Levinson recursion

The solution of the Eqs. (33–34) fully determines the functional form of the power spectral density estimator (30). The method for solving the equations is called the Levinson–Durbin recursion [27] and it is described in the following. For each order N of the iteration we define the quantities:

$$\Delta_N=\sum_{n=0}^{N}a_n\bar{r}_{N-n+1} \tag{35}$$

$$c_N=-\frac{\Delta_N}{P_N}, \tag{36}$$

The Levinson recursion computes the Nth order quantities given the $N-1$th order quantities:

$$P_N=P_{N-1}\left(1-|c_{N-1}|^2\right) \tag{37}$$

and

$$\begin{pmatrix}1\\a_1\\\vdots\\a_{N-1}\\a_N\end{pmatrix}=\begin{pmatrix}1\\b_1\\\vdots\\b_{N-1}\\0\end{pmatrix}+c_{N-1}\begin{pmatrix}0\\b_{N-1}^*\\\vdots\\b_1^*\\1\end{pmatrix}. \tag{38}$$

where b holds the value of the a_s coefficients at order $N-1$. The 0-th order element can be easily initialized reminding that $a_0=1$ (always) and that P_0 can be determined from (33). Its values turns out to be:

$$P_0=R(0), \tag{39}$$

Δ_0 and c_0 are uniquely determined from their definitions and they are:

$$\Delta_0=R(1);\quad c_0=-\frac{R(1)}{R(0)}. \tag{40}$$

These expressions allow us to compute **a** and P_N to any order by simply iterating (37) and (38). Substituting them in Eq. (30) the problem of the estimation for the power spectral density via maximum entropy principle is solved. Burg's method for spectral analysis is solved via Levinson is implemented in the released `memspectrum` package. Another faster recursion method is available in [47] and it is also available in `memspectrum`.

B Additional plots for validation

In this section we report additional plots for the Validation. Figure 11 represents the difference between the estimated PSD and the reference PSD, normalised by the latter. Figures 12 and 13 show the reference PSD (in red) against all the 1000 reconstruction obtained with FPE and VM respectively.

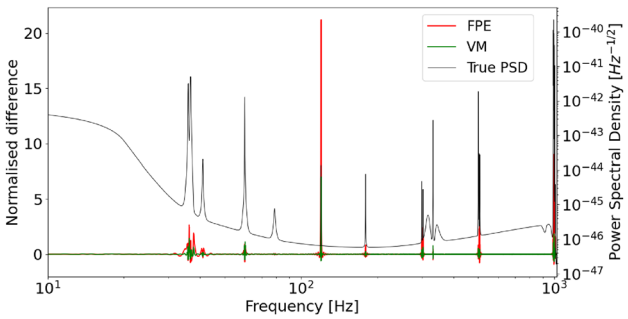

Fig. 11 Normalised difference of reconstructed PSD with respect to the reference PSD

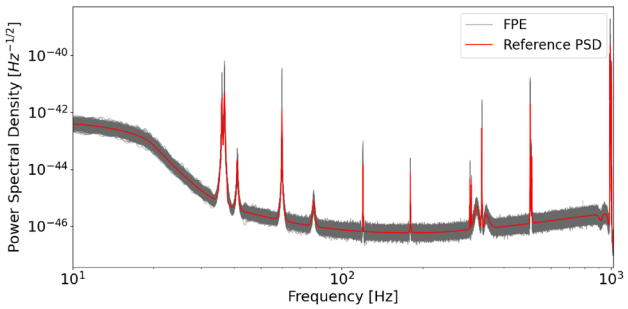

Fig. 12 The 1000 reconstruction of the reference PSD obtained with the FPE loss

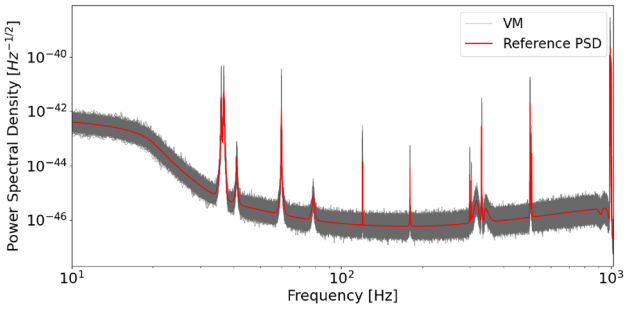

Fig. 13 The 1000 reconstruction of the reference PSD obtained with the VM loss

C Computational time for both MESA methods and Welch

In this appendix we shortly introduce the computational times required by the MESA method (considering both the standard and fast implementation) and the Welch method. They are just inserted to give an idea of what the time differences between the methods are. These are obtained via the python %timeit special function, run on a personal machine (Table 1).

Table 1 Comparison of the computational times for the estimate of the power spectral densities with our implementation of MESA (both standard and fast implementations) and Welch's method

Computational times			
Batch length	MESA std	MESA fast	Welch
1 s	22 ms ± 1.22 ms	19.6 ms ± 620 µs	335 µs ± 9.24 µs
5 s	158 ms ± 21.7 ms	42.4 ms ± 353 µs	839 µs ± 4.61 µs
10 s	187 ms ± 11.6 ms	51.5 ms ± 3.67 ms	1.74 ms ± 64.3 µs
100 s	1.96 s ± 338 ms	205 ms ± 5.09 ms	18.8 ms ± 140 µs
1000 s	17.1 s ± 605 ms	1.33 s ± 17.4 ms	235 ms ± 3.69 ms

D Full posterior distribution for GW150914

Here we report the full posterior distribution obtained from our PSD marginalisation model. The prior posterior distributions assumed are identical to what presented in Ref. [48] for all overlapping parameters (Fig. 14).

References

1. L.S. Finn, Phys. Rev. D **46**(12), 5236–5249 (1992). https://doi.org/10.1103/physrevd.46.5236
2. P. Welch, IEEE Trans. Audio Electroacoust. **15**(2), 70 (1967)
3. N.R. Lomb, Astrophys. Space Sci. **39**(2), 447 (1976). https://doi.org/10.1007/BF00648343
4. J.D. Scargle, Astrophys. J. **263**, 835 (1982). https://doi.org/10.1086/160554
5. N.J. Cornish, T.B. Littenberg, Class. Quantum Gravity **32**(13), 135012 (2015). https://doi.org/10.1088/0264-9381/32/13/135012
6. T.B. Littenberg, N.J. Cornish, Phys. Rev. D (2015). https://doi.org/10.1103/physrevd.91.084034

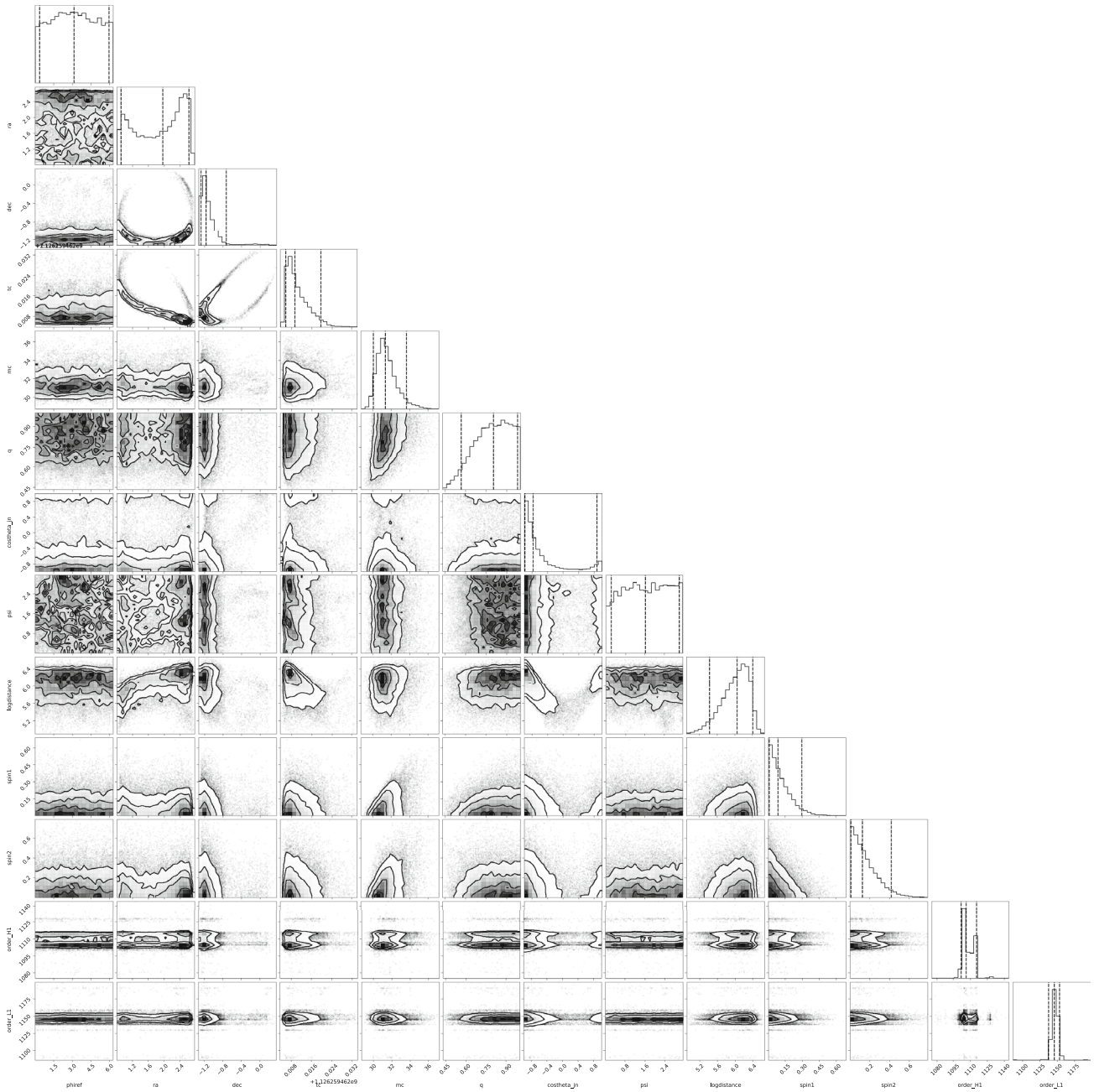

Fig. 14 Full posterior distribution for GW150914 when marginalising over the AR process orders

7. E.T. Jaynes, Phys. Rev. **106**, 620 (1957)
8. E. Jaynes, G. Bretthorst, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, 2003)
9. E.T. Jaynes, Proc. IEEE **70**(9), 939 (1982). https://doi.org/10.1109/PROC.1982.12425
10. J. Burg, Maximum entropy spectral analysis. in Stanford Exploration Project (Stanford University, 1975). https://books.google.it/books?id=Xug_AAAAIAAJ
11. C.E. Shannon, Bell Syst. Tech. J. **27**(3), 379 (1948)
12. R.T. Cox, Am. J. Phys. **14**(1), 1 (1946). https://doi.org/10.1119/1.1990764
13. P. Gregory, *Multivariate Gaussian from Maximum Entropy* (Cambridge University Press, Cambridge, 2005), pp.450–454. https://doi.org/10.1017/CBO9780511791277.020
14. R.M. Gray, *Toeplitz and Circulant Matrices: A Review* (Now Foundations and Trends, 2006). https://doi.org/10.1561/0100000006
15. D.W.F.P.M. Woodward, W. Higinbotham, *Probability and Information Theory, with Applications to Radar*, 2nd edn. (Pergamon Press, 1964). http://cds.cern.ch/record/2031792
16. B. Allen, W.G. Anderson, P.R. Brady et al., Phys. Rev. D **85**, 122006 (2012). https://doi.org/10.1103/PhysRevD.85.122006
17. J.G. Ables, Astron. Astrophys. Suppl. **15**, 383 (1974)
18. M. Bartlett, Louvain Econ. Rev. **34**(2), 227 (1968). https://doi.org/10.1017/S077045180004077X
19. T.J. Ulrych, T.N. Bishop, Rev. Geophys. **13**(1), 183 (1975). https://doi.org/10.1029/RG013i001p00183
20. H. Wold, J. Inst. Actuar. **70**(1), 113–115 (1939). https://doi.org/10.1017/S0020268100011574
21. J.G. Berryman, Geophysics **43**(7), 1384 (1978)
22. H. Akaike, Ann. Inst. Stat. Math. (1998). https://doi.org/10.1007/978-1-4612-1694-0_11
23. S. Kay, Modern spectral estimation, in *Prentice-Hall Signal Processing Series*. (Prentice-Hall, Hoboken, 1988)
24. E. Cuoco, G. Calamai, L. Fabbroni et al., Class. Quantum Gravity **18**(9), 1727–1751 (2001). https://doi.org/10.1088/0264-9381/18/9/309
25. A.R. Rao, R.L. Kashyap, L. Mao, Water Resour. Res. **18**(4), 1097 (1982). https://doi.org/10.1029/WR018i004p01097
26. R.J. Bhansali, Ann. Stat. **14**(1), 315 (1986). https://doi.org/10.1214/aos/1176349858
27. N. Levinson, J. Math. Phys. **25**(1–4), 261 (1946)
28. B. Abbott, R. Abbott, T. Abbott et al., Phys. Rev. X (2019). https://doi.org/10.1103/physrevx.9.031040
29. B. Abbott, R. Abbott, T. Abbott, et al. LIGO Document P1900011-Power Spectral Densities (PSD) release for GWTC-1. LIGO Document Service: https://dcc.ligo.org/LIGO-P1900011/public (2019)
30. N.J. Cornish, T.B. Littenberg, B. Bécsy et al., Phys. Rev. D **103**(4), 044006 (2021). https://doi.org/10.1103/PhysRevD.103.044006
31. K. Chatziioannou, C.J. Haster, T.B. Littenberg et al., Phys. Rev. D (2019). https://doi.org/10.1103/physrevd.100.104004
32. A.J. Owens, J. Geophys. Res. Space Phys. **83**(A4), 1673 (1978)
33. C.R. Harris, K.J. Millman, S.J. van der Walt et al., Nature **585**, 357–362 (2020). https://doi.org/10.1038/s41586-020-2649-2
34. P. Virtanen, R. Gommers, T.E. Oliphant et al., Nat. Methods **17**, 261 (2020). https://doi.org/10.1038/s41592-019-0686-2
35. T. Gupta, N.J. Cornish, Phys. Rev. D **109**, 064040 (2024). https://doi.org/10.1103/PhysRevD.109.064040
36. J. Veitch, V. Raymond, B. Farr et al., Phys. Rev. D **91**, 042003 (2015). https://doi.org/10.1103/PhysRevD.91.042003
37. B. Abbott, R. Abbott, T. Abbott et al., Phys. Rev. Lett. (2016). https://doi.org/10.1103/PhysRevLett.116.061102
38. S. Schmidt, M. Breschi, R. Gamba et al., Phys. Rev. D **103**(4), 043020 (2021). https://doi.org/10.1103/PhysRevD.103.043020
39. A. Nagar, S. Bernuzzi, W. Del Pozzo et al., Phys. Rev. D **98**(10), 104052 (2018). https://doi.org/10.1103/PhysRevD.98.104052
40. J. Veitch, W.D. Pozzo, M. Williams et al., johnveitch/cpnest: fix for python < 3.8 versioning (2020). https://doi.org/10.5281/zenodo.4109277
41. R. Abbott, O. Bulashenko et al., Astrophys. J. Suppl. **267**(2), 29 (2023). https://doi.org/10.3847/1538-4365/acdc9f
42. S. Biscoveanu, C.J. Haster, S. Vitale, J. Davies, Phys. Rev. D **102**, 023008 (2020). https://doi.org/10.1103/PhysRevD.102.023008
43. T.B. Littenberg, M. Coughlin, B. Farr, W.M. Farr, Phys. Rev. D (2013). https://doi.org/10.1103/physrevd.88.084044
44. M.C. Edwards, R. Meyer, N. Christensen, Phys. Rev. D (2015). https://doi.org/10.1103/physrevd.92.064011
45. Q. Baghi, J.I. Thorpe, J. Slutsky et al., Phys. Rev. D **100**, 022003 (2019). https://doi.org/10.1103/PhysRevD.100.022003
46. T.E. Barnard, The maximum entropy spectrum and the Burg technique. Technical Report No. 1: Advanced Signal Processing. NASA STI/Recon Technical Report N (1975)
47. K. Vos, A fast implementation of Burg's algorithm. (2013). https://opus-codec.org/docs/vos_fastburg.pdf
48. R. Abbott, T.D. Abbott, F. Acernese et al., Phys. Rev. D **109**, 022001 (2024). https://doi.org/10.1103/PhysRevD.109.022001