An algorithm for the detection of extreme mass ratio inspirals in LISA data

Stanislav Babak¹, Jonathan R Gair² and Edward K Porter¹,³

¹ Albert Einstein Institute, Am Muehlenberg 1, D-14476 Golm, Germany
² Institute of Astronomy, Madingley Road, CB3 0HA Cambridge, UK
³ APC, UMR 7164, Université Paris 7 Denis Diderot, 10, rue Alice Domon et Léonie Duquet, 75205 Paris Cedex 13, France

E-mail: stba@aei.mpg.de and jgair@ast.cam.ac.uk

Received 24 February 2009, in final form 29 April 2009
Published 8 June 2009
Online at stacks.iop.org/CQG/26/135004

Abstract
The gravitational wave signal from a compact object inspiralling into a massive black hole (MBH) is considered to be one of the most difficult sources to detect in the LISA data stream. Due to the large parameter space of possible signals and many orbital cycles spent in the sensitivity band of LISA, it has been estimated that \(\sim 10^{35}\) templates would be required to carry out a fully coherent search using a template grid, which is computationally impossible. Here we describe an algorithm based on a constrained Metropolis–Hastings stochastic search which allows us to find and accurately estimate parameters of isolated EMRI signals buried in Gaussian instrumental noise. We illustrate the effectiveness of the algorithm with results from searches of the Mock LISA Data Challenge round 1B data sets.

PACS numbers: 04.20.—q, 04.30.Db, 04.30.Tv, 04.80.Nn

(Some figures in this article are in colour only in the electronic version)

1. Introduction
Extreme mass ratio inspirals (EMRIs)—stellar mass compact objects (COs) that are captured and spiral into massive black holes (MBHs) through the emission of gravitational radiation—are one of the most interesting sources for the future LISA mission [1]. The inspiral proceeds very slowly for these sources (the inspiral rate is proportional to the mass ratio which is typically \(1 : 10^5\)) so the CO spends a significant amount of time in the strong field region close to the MBH. The GW signal encodes information about the structure of the spacetime close to the central black hole which we hope to be able to extract in order to confirm or otherwise that the massive objects observed in galactic nuclei are indeed Kerr BHs, as we suppose [2–6].
EMRI observations may also be used to probe the stellar population in the central parsecs of galaxies and to measure the properties of astrophysical black holes to high precision [7]. This is possible because the signal is long lived and so a coherent phase integration should be able to determine the parameters of the binary to a high accuracy that is currently unprecedented in electromagnetic observations. We refer the reader to the review paper [8] for more details on the astrophysics that will be possible with EMRIs.

In order to scope out issues associated with LISA data analysis for EMRIs, we require waveform models that are cheap and easy to generate but still capture the main features of true EMRI waveforms. One such model of the signal is the so-called analytic kludge waveform [9]. It is a phenomenological template, constructed by putting together the most important physical elements: post-Newtonian expressions for the rate of change of the orbital parameters and frequencies (the Peter–Mathews approach), periastron precession and precession of the orbital plane around the spin axis of the MBH. While these waveforms are not faithful models of true EMRI signals, they capture the main qualitative features of the signals. This means that they should be sufficient to scope out issues associated with EMRIs, for instance, the accuracy we would expect to achieve in estimating the source parameters [9] and the level of the confusion background from cosmological EMRIs [10]. Because these waveforms are simple and fast to generate, they were chosen for use in the Mock LISA Data Challenge (MLDC). The MLDC was organized to stimulate the development of data analysis tools for LISA and to establish standard notations and conventions which allow comparison of different algorithms. There have been three challenges to date [11–15], which were aimed at different sources: SMBH binaries, Galactic white-dwarf binaries and EMRIs. For each of the two EMRI challenges, five data sets were released, each containing a single EMRI signal buried in instrumental noise.

The five EMRI data sets had some parameters drawn from priors common to all data sets, which were: the mass of CO $\mu \in U[9.5, 10.5] M_\odot$, the spin $S/M^2 \in U[0.5, 0.7]$, the plunge time $t_p \in U[1, 2]$ years and the eccentricity at plunge $e_{pl} \in U[0.15, 0.25]$; some parameters drawn from priors that were different for each data set: MBH mass $M \in U[0.95, 1.05] \times 10^6 M_\odot$ (high mass binary) with SNR $\in U[40, 110]$ (1.3.1 data set), $M \in U[4.75, 5.25] \times 10^6 M_\odot$ (medium mass binary) with SNR $\in U[70, 110]$ (1.3.2 data set) and with SNR $\in U[40, 60]$ (1.3.3 data set), $M \in U[0.95, 1.05] \times 10^6 M_\odot$ (low mass binary) with SNR $\in U[70, 110]$ (1.3.4 data set) and with SNR $\in U[40, 60]$ (1.3.5 data set). Note that the last two types of signals are in the category that are considered to be most likely to be seen by LISA: a $\sim 10 M_\odot$ BH falling into $\sim 10^6 M_\odot$ MBH [7]. More details on the parameter sets can be found in [13]. MLDC round 1 and round 1B had the same sets of priors for the five data sets. For both challenges, it has been shown that the signal can easily be detected with high confidence, but with parameters quite different from the true ones.

An important feature of EMRI signals is that they have many local maxima on the likelihood surface, which are quite well separated and can be as high as 75% of the true maximum. Search algorithms have a tendency to find secondary maxima quickly and then get stuck there. This represents a true detection but with incorrect parameters (see [15, 16] for results). However in this paper we will only regard a ‘detection’ as finding the global maxima in the likelihood, i.e., the true source parameters. Secondary maxima are likely to be the biggest problem for LISA data analysis. Some signals can be seen by eyes in a spectrogram or in the power spectral density of the data—the main problem is to estimate the source parameters with the best possible accuracy. A grid-based search (with a sufficiently fine grid) would be guaranteed to find the global maxima as it covers the whole parameter space, but the required number of templates is so high [17] that no one is presently considering doing it.

4 The ‘round 1’ EMRI data set was released as part of round 2 of the MLDC.
even with the tight priors provided in the MLDC. An alternative approach which has proven to be both efficient and accurate was first suggested in the context of LISA for non-spinning SMBH binary searches [18–20]. This approach is the semi-stochastic Metropolis–Hastings Monte Carlo (MHMC) method, where one constructs a search chain through the parameter space (these are not in general Markovian) and follows this up by a Markov chain Monte Carlo (MCMC) to sample the posterior distribution function. We say this approach is ‘semi-stochastic’ since, although successive points in the chains are chosen at random, they are chosen from directed proposal distributions. This method involves generating templates as the chain moves, but its power lies in the fact that the number of points at which templates must be generated in order to find the source is usually many fewer than in a full template grid. However, the chains can get stuck on local maxima. One needs to use the properties of the signal to encourage the chains to move off local maxima and explore the parameter space more widely.

By including tricks such as simulated annealing—‘heating’ the likelihood surface by scaling both the log-likelihood and the size of proposed jumps by a temperature factor—and frequency annealing—systematically increasing the range of frequencies included in the waveform template—MHMC has solved the search problem for non-spinning SMBH binaries [18–20]. The full utility of MHMC for EMRI searches has so far not been demonstrated. In the round 1B MLDC release, one of the five EMRIs (of the 1.3.1 type) was found by an MHMC technique [21] that used simulated annealing and began with searches on sub-segments of the data that were combined before finally running a long chain on the whole data set. The authors [22] also attempted to search for EMRIs in the round 1B data using MHMC and recovered close to the true parameters for one of the five signals (of the 1.3.2 type) before the Challenge deadline. However, in both round 1 and round 1B, the best performing algorithm was not MHMC based, but a time-frequency analysis [23, 24]. Such searches are easier to implement, and parameter estimation can be done if the EMRI is isolated and of sufficient brightness. However, the achievable accuracy of parameter estimation using such template-free techniques is not as good as template-based methods, and the algorithm will suffer when faced with source confusion.

In this paper, we describe for the first time a complete template-based EMRI search that is able to detect and recover accurate parameters for bright, isolated EMRI sources buried in instrumental noise, with parameters drawn from any of the five canonical MLDC EMRI source types. This search technique is based on our previous MHMC search, but with several improvements which we describe below.

The origin of all the secondary maxima on the EMRI likelihood surface is in the characteristics of the signal. An EMRI signal is composed of many harmonics of the three fundamental orbital frequencies (of the radial \(r\)-motion, the azimuthal \(\phi\)-motion and the polar \(\theta\)-motion), which are evolving in time. These harmonics vary in strength (amplitude), and the local maxima arise from matching the phase of the strongest (or of several strong) harmonics for some period of time. It is possible for a signal with very different parameters to match the dominant harmonic very well for the whole duration of the signal but miss completely all the other harmonics. We have tried to exploit this property by using several chains to identify the dominant harmonic and then impose a constraint between the fundamental frequencies that fixes the frequency of the dominant harmonic at some reference time. The key idea of our search is to determine the frequency of the dominant harmonic/harmonics using several local maxima, and this was used for the 1B submission [22]. However, since the round 1B MLDC deadline, we have improved our search technique in three important ways. We have changed the parametrization of the signal so that it is now specified by the three orbital frequencies at some reference time, \(t_{\text{ref}}\), and we have changed the proposal distributions accordingly.
We use two main proposal distributions: a normal multivariate Gaussian aligned with the
eigendirections of the Fisher matrix and a variation of the Metropolis random walk which we
will describe later. The second important improvement was to release the constraint after a
certain point and let the chains correct the frequency of the dominant harmonic at
time. The signal can be described by harmonics of three fundamental orbital frequencies:
\( f_\nu \equiv \dot{\nu}/(2\pi) \), \( f_\alpha \equiv \dot{\alpha}/(2\pi) \), where the dot denotes a derivative with respect to time.
Frequency \( \nu \) is the orbital frequency, i.e., the frequency of successive periapse passages, \( f_\gamma \) is the frequency of precession of the perihelion and \( f_\alpha \) is the frequency of precession of the orbital plane that arises due to the spin–orbit coupling. The frequencies evolve according to
\[
\frac{d\nu}{dt} = \frac{96}{10\pi} (\mu/M^3)(2\pi M \nu)^{1/3}(1 - e^2)^{-9/2} \left[ [1 + (73/24)e^2 + (37/96)e^4](1 - e^2)
+ (2\pi M \nu)^2/3[(1273/336) - (2561/224)e^2 - (3885/128)e^4
- (13147/5376)e^6] - (2\pi M \nu)(S/M^2) \cos \lambda(1 - e^2)^{-1/2}
\times [(73/12) + (1211/24)e^2 + (3143/96)e^4 + (65/64)e^6]\right],
\]
\[
\frac{df_\nu}{dt} = \left[ (2\pi v M)^{2/3}(1 - e^2)^{-1}\right] \left[ 5 + \frac{7}{4}(2\pi v M)^{2/3}(1 - e^2)^{-1}(26 - 15e^2) \right]
- 12 \cos \lambda(S/M^2)(2\pi M \nu)(1 - e^2)^{-3/2} \left[ \frac{d\nu}{dt} \right]
+ \left[ 6\nu(2\pi v M)^{2/3}(1 - e^2)^{-1}\right] \left[ 1 + \frac{11}{2}(2\pi v M)^{2/3}(1 - e^2)^{-1}\right]
- 18\nu \cos \lambda(S/M^2)(2\pi M \nu)(1 - e^2)^{-3/2} \frac{e}{(1 - e^2)^{3/2}} \frac{de}{dt},
\]
\[
\frac{df_a}{dt} = 2\nu(S/M^2)(2\pi M\nu)(1-e^2)^{-3/2}
\left(\frac{1}{\nu}\frac{dv}{dt} + \frac{3e}{1-e^2}\frac{de}{dt}\right),
\]
(3)

\[
\frac{de}{dt} = -\frac{e}{15}(\mu/M^2)(1-e^2)^{-7/2}(2\pi M\nu)^{8/3}(304 + 121e^2)(1-e^2)(1 + 12(2\pi M\nu)^{2/3})
\]
\[
-\frac{1}{56}(2\pi M\nu)^{2/3}((8)(16705) + (12)(9082)e^2 - 25211e^4)
\]
\[
+ \mu/2(\mu/M^2)\cos\lambda(2\pi M\nu)^{11/3}(1-e^2)^{-4}[13(1364/5)
\]
\[
+ (5032/15)e^2 + (263/10)e^4]
\]
(4)

The harmonic structure of the signal is best seen when using a static source frame defined by the direction of the spin of the MBH, which is assumed to be constant in this model. The radiative frame is then constructed using the direction of propagation (or the direction to the source from the solar system barycenter (SSB)) and the spin direction of the MBH. The advantage of those two frames is that they are static and all the time dependence is encoded in the amplitude and phases of the harmonics explicitly. In the original analytic kludge paper [9], the waveform was expressed relative to a precessing frame, tied to the orbital angular momentum, which makes it more complicated to compute the harmonic decomposition. In the static SSB frame, the signal takes the following form:

\[
h \sim \mu(2\pi M\nu(t))^{2/3} \sum_{l,n,m} A_{l,n,m}(e(t)) e^{i(l\phi(t) + n\psi(t) + m\alpha(t))}.
\]
(5)

The amplitude of each harmonic depends on the source location (ecliptic coordinates), orientation of the spin and the orbital eccentricity. These expressions are known analytically, but are messy so we do not include them explicitly here. We have examined the amplitudes of the harmonics for a wide range of parameters and find that harmonics of the perihelion precession with \(l \neq 2\) are significantly suppressed. We can also neglect the contribution from harmonics of the orbital frequency with \(n > 5\) for orbital eccentricities less than \(e \sim 0.65\). Moreover, by construction, the analytic kludge waveforms are quadrupolar and therefore only harmonics of the orbital plane precession frequency with \(m \in [-2, 2]\) are allowed. This will not be the case for real EMRI signals, and more sophisticated models include higher multipoles [26, 27]. Knowledge of the analytic form of the harmonic amplitudes and the restriction of the number of harmonics, to as few as \(\sim 4\)–\(8\) dominant harmonics in most cases, allows us to simplify the template and make its generation more efficient. The amplitudes of the harmonics depend on Bessel functions, with argument, \(ne(t)\), that is usually small, so a further simplification follows by expanding these as Taylor series and truncating at the desired level of accuracy.

The technique of time-delay interferometry (TDI) [28] will be used in order to cancel the laser noise in the LISA data. The basic idea of this technique is to combine the data sent and received by different spacecraft with time delays chosen to cancel the common laser noise component. The LISA response function is consequently somewhat complicated, although it can be significantly simplified in the long wavelength limit \(\omega_{GW}L \ll 1\) [29] (L is LISA’s arm length ≈ 16.7 s and \(\omega_{GW}\) is the GW frequency). For an EMRI into a lower mass MBH (the 1.3.4/1.3.5 type source), the frequency of the GWs can be quite high and neither the long wavelength nor rigid adiabatic approximations [29] are valid. Our code uses the full response but with time delays applied only to \(h^{SSB}\) and not to the LISA motion—treating LISA as a solid rotating triangle. This is overkill for the high mass MBH EMRIs (1.3.1 type) (and

\[5\] We are working in geometrical units \(G = c = 1\).
probably for the medium mass, 1.3.2 type, sources as well), but we decided to use the same
codes for all searches. We have saved on computing time for the higher mass systems by using
a lower sampling rate during the integration of the orbital motion and then up-sampling while
generating the TDI streams. We also use linear interpolation to compute the time delayed
data from a regularly sampled SSB time series, rather than a more complicated and expensive
interpolation scheme.

We have verified our waveform templates against full analytic kludge templates generated
using Synthetic LISA [30], by computing the overlap, which is the inner product,

\[
(s|h) = 2 \int \frac{\hat{s}^* \hat{h}^* + \hat{h}^* \hat{s}^*}{S_b(f)} \, df, 
\]

between two normalized signals, \((s|s) = (h|h) = 1\). In the expression above, a tilde denotes
the Fourier transform and \(S_b(f)\) is the one-sided noise power spectral density. We have found
that the overlap between our approximate model and the accurately computed templates is in
the range (0.93–0.99) depending on the source parameters, in particular the mass of the MBH
(the overlap is usually higher for high mass MBH EMRIs). The loss in the overlap comes
primarily from mismatches in the amplitude, while the phase is tracked very well. This is to
be expected, as we do not make any approximations in our computation of the evolution of the
orbital parameters and frequencies. The small mismatch between the template and the signal
will lead to a bias in the parameters estimated for the signal. A mismatch in amplitude will
primarily affect the estimated signal-to-noise ratio/luminosity distance for the source, while
a phase error will lead to errors in all the intrinsic parameters. We have found that our model
is very faithful, with typical model-induced parameter errors for MLDC source types being
\(~1–2\sigma\), where \(\sigma\) is the parameter error as estimated from the Fisher matrix. In other words,
we expect the model error to be of the similar size, but no larger than the error in parameter
recovery that arises from instrumental noise in the detector. This is confirmed by the results
of our search, as summarized in tables 1 and 2. We see that our parameter recovery was very
good except for the SNR which was as much as \(~5\%\) different in two of the low mass MBH
cases.

3. Search algorithm

In this section, we will describe the overall search algorithm. In practice, there were some
differences between the searches for each source and we will discuss these source-specific
details in the following section. Our search consists of three stages. In the first stage, we
look for the ‘footprints’ of the signal and for points from which we can seed our subsequent
chains. In the second stage, we construct chains using the identified properties of the signal,
via a constrained Metropolis Monte Carlo search on half-year long segments of data. The
final stage is to refine the parameters of the signals by extending the duration of the templates
to the total length of observation (this is similar to what was done in [21]). In the following
subsections, we will give details of each of the three stages.

3.1. Stochastic search

As mentioned previously, the basis of our search method is to identify as many strong local
maxima in the likelihood as possible and then to use the information encoded in these maxima
to direct the search toward the true solution. The first stage is very simple and remarkably
efficient. In spirit, it is similar to using a random template bank, as described in [31]. We
generate template waveforms for the last half a year of inspiral with parameters randomly
chosen from within uniform priors and record their likelihoods. For greater efficiency, we include maximization of the log-likelihood over the distance, plunge time and three orbital phases at plunge. We maximize over the plunge time in the usual way, by computing the correlation of the template with the data (instead of the inner product). The maximized value of the plunge time is then used for constructing a new filter, and we compute the likelihood maximized over phases and distance. The maximization over the distance is done in the usual way: the log-likelihood (up to a constant factor) is given by

\[ \Lambda = -\sum_I (x_I - h_I|x_I - h_I) \sim \sum_I 2(x_I|h_I) - (h_I|h_I), \]

where \( I = \{A, E\} \) runs over orthogonal TDI streams [28], which here play the role of independent detectors, \( x_I = n_I + s_I \) are the corresponding TDI data which are combined out of the noise \( n_I \) and a signal \( s_I, h_I \) is a template and the inner product is defined in equation (6).

An amplitude factor, \( A \), (inversely proportional to the luminosity distance to the source) can be factored out \( h_I = A \hat{h} \) and maximized over analytically. The maximum likelihood estimator for the amplitude is

\[ A = \frac{\sum_I (x_I|h_I)}{\sum_I (h_I|h_I)}, \]

and then the maximized log of likelihood is

\[ \Lambda_{\text{max}}(A) = \frac{\left(\sum_I (x_I|h_I)\right)^2}{\sum_I (h_I|h_I)}. \]

This value is sometimes referred to as SNR\(^2\), since if \( s_I = h_I \), it reduces to \( \sum_I (h_I|h_I) \), which is the square of the matched-filtering signal-to-noise ratio. Note that equation (9) is not sensitive to the sign of the inner product.

Maximization over the three initial orbital phases is more involved. We consider a waveform template which is constructed out of three bright harmonics only. The brightness of each \( m \)-harmonic for a given set of source parameters depends only on the inclination angle \( \lambda \) and on the inclination of the MBH’s spin to the direction to the source from the SSB [24]. The prior range on plunge eccentricity is such that the dominant \( n \) harmonics would be \( n = 2 \) and/or \( n = 3 \). This suggests that we want to use the following harmonics \( n_0 = n_1 = 2, m_0 \neq m_1, n_2 = 3, m_2 = m_0 \), with \( m_0 \) the brightest of \( m_0, m_1 \). The three initial phases for harmonics \( h^{(i)} \) are

\[ \Phi_i = n_i \phi_0 + 2 \gamma_0 + m_i \alpha_0. \]

Each harmonic is of form \( \cos (\Phi_i + \tilde{\phi}^{(i)}(t)) \) and hence may be decomposed as

\[ h^{(i)} = A_i^0 \cos \Phi_i \tilde{h}^{(i)}(0) - A_i^0 \sin \Phi_i \tilde{h}^{(i)}(\pi/2), \]

where \( \tilde{h}^{(i)}(0) \) means that it is taken at the zero initial phase. The three-harmonic template can therefore be written as

\[ h^c = h^{(0)} + h^{(1)} + h^{(2)} = \sum_{j=0}^5 a_j h^j. \]

The cross-harmonic overlaps, i.e., \( (h^i|h^j) \) for \( i \neq j \), are in general small because individual harmonics cannot cross in the time-frequency plane, and because the orbital parameters vary on a much longer timescale than the orbital period (and therefore \( \sin \Phi(t) \) and \( \cos \Phi(t) \) are approximately orthogonal). If we included these terms, the maximum could still be found
easily, but it would involve the calculation of 15 cross-harmonic overlaps and this additional computational overhead would not be justified by the modest improvement in amplitudes that would result. Therefore, we omit all the cross harmonic terms and analytically maximize the likelihood of our template \( h^{r} \) over all values of the constants \( a_{ij} \), in a similar way to the \( F \)-statistic,

\[
a_{2i} = \frac{\sum_{j} (x_{ij} \hat{h}_{ij}^{(1)}(0))}{\sum_{j} (\hat{h}_{ij}^{(1)} \hat{h}_{ij}^{(1)})}, \quad a_{2i+1} = \frac{\sum_{j} (x_{ij} \hat{h}_{ij}^{(1)}(\pi/2))}{\sum_{j} (\hat{h}_{ij}^{(1)} \hat{h}_{ij}^{(1)})},
\]

where \( i = 0, 1, 2 \). This leads to the following maximum likelihood estimators for the amplitude and phase of the harmonics:

\[
\Phi_{0}' = \arctan \left( \frac{a_{2j+1}}{a_{2j}} \right), \quad A_{0}' = \sqrt{a_{2j}^{2} + a_{2j+1}^{2}}. \tag{14}
\]

With the choice of harmonics given above, we can obtain the initial orbital phases from the maximum likelihood estimates of the harmonic amplitudes and phases:

\[
\phi_{0} = \frac{\Phi_{0}' - \Phi_{2}'}{m_{0} - m_{2}}, \tag{15}
\]

\[
\gamma_{0} = \frac{1}{2} \left[ -\Phi_{0}' \frac{m_{0} m_{0} - m_{1} m_{2}}{(m_{0} - m_{2})(m_{0} - m_{1})} + \Phi_{0}' \frac{m_{0}}{m_{0} - m_{1}} + \Phi_{2}' \frac{m_{2}}{m_{0} - m_{2}} \right], \tag{16}
\]

\[
\alpha_{0} = \frac{\Phi_{0}' - \Phi_{2}'}{m_{0} - m_{1}}. \tag{17}
\]

After a new plunge time is determined from the correlation analysis, we estimate the initial phases using the above method. This guarantees that we have chosen the optimal phases if at least one of the harmonics in the template matches the signal. We then use the maximized phases and distance to compute the log-likelihood, equation (9). We will refer to this as the ‘3-harmonic physical’ likelihood as the final likelihood is computed using the physical waveform template with the maximized phases and distance.

Using three harmonics allows us to get rid of the three phase parameters and distance in the search. The orientation of the black hole spin is also an extrinsic quantity that only affects the relative brightness of the different \( m \) harmonics. This can be maximized over automatically in a similar way, although the statistic is slightly more complicated as the spin orientation also affects the relative contributions of the plus and cross polarizations to a given detector channel. This technique was described in [32] for galactic binary systems, but we apply it here to each harmonic of the EMRI signal. The LISA detector response in a given channel, \( I \), for a single harmonic, \( \hat{h}_{i}^{r} \), say, may be written as [32]

\[
\hat{h}_{i}^{r} = h_{+}^{(i)} f_{i}^{(+)} + h_{\times}^{(i)} f_{i}^{(\times)} = h_{+}^{(i)} (u_{i}(t) \cos(2\psi) + v_{i}(t) \sin(2\psi)) + h_{\times}^{(i)} (v_{i}(t) \cos(2\psi) - u_{i}(t) \sin(2\psi)), \tag{18}
\]

The polarization \( \psi \) depends on the direction of the MBH spin, but the detector modulation functions, \( u_{i}(t) \) and \( v_{i}(t) \), depend only on the sky location of the source. As before, we may approximate the two polarizations of the harmonic as \( h_{+}^{(i)} = A_{+}^{(i)} \cos(\Phi_{0}' + \Phi_{i}'(t)) \) and \( h_{\times}^{(i)} = A_{\times}^{(i)} \sin(\Phi_{0}' + \Phi_{i}'(t)) \). The two amplitudes \( A_{+}^{(i)} \) and \( A_{\times}^{(i)} \) contain a common time-dependent part, which we denote as \( A(t) \), and a time-independent piece that is different in the two cases (for a circular equatorial binary the time-independent parts are \( 1 + (\hat{L} \cdot \hat{\mathbf{N}})^{2}/D \) and \(-2(\hat{L} \cdot \hat{\mathbf{N}})/D \), here \( \hat{L} \) and \( \hat{\mathbf{N}} \) are unit vectors along the orbital angular momentum and the
works as follows: given a data set \( s(t) \), we may therefore write

\[
h^{(i)}_I = a^{(i)}_1 A(t) u_I(t) \cos(\tilde{\phi}(t)) + a^{(i)}_2 A(t) v_I(t) \sin(\tilde{\phi}(t)) + a^{(i)}_3 A(t) u_I(t) \sin(\tilde{\phi}(t)) + a^{(i)}_4 A(t) v_I(t) \cos(\tilde{\phi}(t))
\]

All the dependence of the waveform on phase angles, distance and the spin orientation is now encoded in the four \( a^{(i)} \) coefficients. If we assume these are independent, we can maximize over them automatically by inverting the matrix equation

\[
\begin{pmatrix}
U & Q & 0 & P \\
Q & V & -P & 0 \\
0 & -P & U & Q \\
P & 0 & V & 0
\end{pmatrix}
\begin{pmatrix}
a^{(i)}_1 \\
a^{(i)}_2 \\
a^{(i)}_3 \\
a^{(i)}_4
\end{pmatrix}
= \begin{pmatrix}
\sum_I (x_I | h^{(i)}_{11}) \\
\sum_I (x_I | h^{(i)}_{12}) \\
\sum_I (x_I | h^{(i)}_{13}) \\
\sum_I (x_I | h^{(i)}_{14})
\end{pmatrix}
\equiv x
\]

where

\[
U = \sum_I (h^{(i)}_{11} | h^{(i)}_{11}) \\
V = \sum_I (h^{(i)}_{12} | h^{(i)}_{12}) \\
P = \sum_I (h^{(i)}_{13} | h^{(i)}_{13}) \\
Q = \sum_I (h^{(i)}_{14} | h^{(i)}_{14})
\]

and the summation is over all the independent LISA data channels. In this case, the four basis waveforms are not orthogonal and so we need to include the cross terms in the matrix inversion. The maximized SNR\(^2\) for this harmonic is then given by \( \rho_{i0}^2 = x^T \cdot M^{-1} \cdot x \). This can be repeated for any number of harmonics, \( N \), and an \( 'N\)-harmonic maximized' likelihood computed by summing \( \sum_{i=1}^N \rho_{i0}^2 \) (NB as before, we ignore cross-contamination between harmonics since this is minimal). This maximized likelihood does not force consistency between the maximizing parameters. However, we found it to be a useful statistic since it reduces the parameter space to be searched and ensures that any point where one of the harmonics has good phase overlap with a harmonic of the true signal returns a high value of the statistic and will therefore be explored further.

3.2. Search on subsets of data

During the first stage of the search, the longer we run, the more high likelihood points we are able to find. We typically use about 100–200 CPUs for several days in this phase, and normally identify a few dozen distinct secondaries with SNR of about 20%–40% of the maximum. Having identified these secondaries, we move on to the next stage of the search, which we call ‘improve frequencies–fix–release’. This stage involves repeated iteration of two steps. For this whole stage, we divide the data stream into half-year long segments. There are two reasons why we carry out the search on short duration segments at first: (i) it is much faster to generate short templates, so the search proceeds more quickly; (ii) the accuracy of estimating the eigenvalues and eigenvectors of the Fisher matrix for our main proposal drops faster to generate short templates, so the search proceeds more quickly; (ii) the accuracy of estimating the eigenvalues and eigenvectors of the Fisher matrix for our main proposal drops with an increase in the duration of the template. For the first step, ‘improve frequencies’, we run an MCMC using the Metropolis rejection/acceptance rule [33]. The MCMC technique works as follows: given a data set \( \hat{s}(t) \) and a set of templates \( \hat{h}(t; \vec{x}) \), we choose a starting point, \( \vec{x} \), in the parameter space. We then propose a jump to another point, \( \vec{y} \), in the space by drawing from a certain proposal distribution, \( q(\vec{y}|\vec{x}) \), and evaluate the Metropolis–Hastings ratio

\[
H = \frac{\pi(\vec{y}) p(s|\vec{y}) q(\vec{x}|\vec{y})}{\pi(\vec{x}) p(s|\vec{x}) q(\vec{y}|\vec{x})}.
\]
Here $\pi(\vec{x})$ are the priors of the parameters, which, in our analysis, were taken to be uniform distributions within the ranges allowed by the MLDC. The function $p(s|\vec{x})$ is the likelihood

$$p(s|\vec{x}) = C e^{-<s-h(\vec{x})|s-h(\vec{x})>/\Theta},$$

where $C$ is a normalization constant and $\Theta = 2$ without annealing. This jump is then accepted with probability $\alpha = \min(1, H)$, otherwise the chain stays at $\vec{x}$. In our search, we use the Metropolis rejection/acceptance rule which simplifies the above by assuming the proposal $q(\vec{y}|\vec{x})$ is symmetric, so the ratio (21) is just the product of the likelihood ratio with the prior ratio. In this stage we also include simulated annealing, which means that $\Theta$ is allowed to vary from 2. This has the effect of smoothing and flattening the likelihood surface, which makes it easier for the chain to move around and climb up the surface to the maximum. The idea is to have a high heat initially, to encourage the chain to explore widely and find the global maximum, and then cool the surface so the chain locks into the vicinity of the maximum. We vary the temperature as the chain advances according to a schedule of the form

$$\Theta = 2 \times \begin{cases} \left(\frac{\text{SNR}_0}{\text{SNR}}\right)^3 & \text{SNR} \leq \text{SNR}_0 \\ 1 & \text{SNR} > \text{SNR}_0 \end{cases},$$

where $\text{SNR}_0$ is typically 6–7. This annealing scheme is used at the beginning of the search and helps to find the trace of the signal quickly by exploring widely in the large parameter space. At later stages of the search, we also made use of thermostated annealing, as described in [19, 20], which encourages the search chains to explore the vicinity of identified maxima.

During this stage of the search, we employ a different parametrization of the template, by prescribing the three orbital frequencies at some reference time $t_{\text{ref}}$ (usually chosen in the middle of the segment) instead of the mass and spin of the MBH. We continue to use the ‘3-harmonic physical’ likelihood statistic to maximize over the waveform phases. We start $n_1$ chains, where $n_1$ is the number of interesting, i.e., high SNR, parameter space points found in stage 1 of the search. The idea behind this second stage is that the points found in the first stage are not far in the parameter space from maxima (local/secondary or global/primary) on the likelihood surface. The MCMC is efficient at moving to these maxima. To date, we have been unsuccessful in attempts to make the chains efficiently jump between local maxima until they find the global one. Instead we use the information stored in each maximum to guide the search in the right direction.

The MCMC search employs two main proposals, $q(\vec{y}|\vec{x})$. The first uses jumps within the scaled ambiguity ellipsoid (defined by the eigenvectors and eigenvalues of the variance–covariance matrix). Following [34, 35], we introduce the metric on the parameter space:

$$d\mathbf{x}^2 = g_{\mu\nu} \, dx^\mu \, dx^\nu,$$

where $h_{\mu\nu} = \partial h/\partial x^\mu$ and $x^\mu$ are parameters of the template. We can determine from the metric its eigenvectors, $\mathbf{V}_i$, and eigenvalues, $\lambda_i$, and hence write $g = V^T \Lambda V$, where $V$ is the matrix of eigenvectors and $\Lambda$ is the diagonal matrix of eigenvalues. We introduce a new parametrization

$$d\mathbf{x}^2 = W^T \mathbf{W}, \quad W_i = Y_i \sqrt{\lambda_i}, \quad Y = V^T \, d\mathbf{X},$$

where $d\mathbf{X}$ is a vector of parameters. We choose a value of $d\mathbf{x}^2$ according to a Gaussian distribution, $[N(0, 1)]$, and choose the direction vector $dW$ randomly oriented on the hypersphere with radius $d\mathbf{x}$. In other words, this proposes jumps on the surface of the ambiguity ellipsoid scaled according to the chosen $d\mathbf{x}$. Another similar proposal which was used for the search of some data sets makes normal jumps in the eigendirections. This latter proposal was first suggested in [19]. Note that for both proposals, the jumps are further scaled by
Figure 1. Harmonics \( n = 2, m = [-2, 2] \) for the best points identified in the first stage of the search. For each chain run in the first stage, we identify the point in parameter space with the highest SNR found by that chain. For those parameters, we compute the frequencies of all the waveform harmonics at a time, \( t_{\text{ref}} \), in this case in the middle of the third half year. The horizontal axis ‘index’ labels the chain from which the point was taken and does not refer to harmonic indices. We see that all the points agree very well on the dominant harmonic, which in this case is \( m = 2 \) (triangles down), and this harmonic has the smallest dispersion in the figure.

the temperature when using the simulated annealing scheme to ensure larger jumps when the surface is ‘hot’.

The second proposal which we have found to be efficient early in the search is based on an estimation of the frequency of the dominant harmonic. For each of the high SNR points identified in the first stage of the search, we can compute the frequency of all harmonics at reference time, \( t_{\text{ref}} \). In general, all the points agree on the frequencies of the dominant harmonics with a small dispersion, \( \sigma \). An example from our blind search is shown in figure 1. One can clearly see that all the points agreed about the frequency of the harmonic \( n = 2, m = 2 \), although some points matched this harmonic with the \( m = 1, m = 0 \) or even \( m = -1 \) harmonic of the template (but with completely wrong parameters). The scatter reflects the relative amplitude of the harmonics: the weakest will have the largest dispersion.

From these first-stage results, we can obtain an estimate of the frequency of the dominant harmonic at \( t_{\text{ref}} \), \( F_{n=2,m=2} \), and the error in that frequency, \( \sigma_{n=2} \). These are computed as the mean and standard deviation of the high SNR points identified in the first search stage. In figure 1, the points lying on the line with frequency \( \sim 0.00536 \) Hz would be used to obtain this estimate. In a similar way, we can compute frequencies \( f_\alpha \) and \( f_\gamma \) at \( t_{\text{ref}} \) for all the high SNR points identified in the first stage of the search and estimate the mean and standard deviation in these frequency estimates—\( \bar{f}_\gamma, \sigma_\gamma \) for \( f_\gamma \) and \( \bar{f}_\alpha, \sigma_\alpha \) for \( f_\alpha \). Having precomputed these quantities, we implement a proposal in the search which proposes a new value of \( F_{n=2,m=2} \) from \( N(\bar{F}_{n=2,m=2}, \sigma_{n=2}) \), a new value of \( f_\gamma \) from \( N(\bar{f}_\gamma, \sigma_\gamma) \) and a new value of \( f_\alpha \) from \( N(\bar{f}_\alpha, \sigma_\alpha) \), and then computes the proposed value of \( \nu = F_{n=2,m=2}/2 - f_\gamma - f_\alpha \). This proposal works well in the beginning of the chains to refine the frequencies given other parameters.

We run the MCMC chains until the search reaches a static state, at which point we stop the chains and start the ‘fix frequency’ step. First, we need to understand what we have detected.
To do this, we evaluate the ‘$N$-harmonic maximized’ $F$-statistic, as described earlier, for the parameter values at the best points found by each chain. This tells us which harmonics of the templates have locked onto harmonics of the true signal. We claim a detection with a harmonic if the SNR in that harmonic is $\geq 5$. Different chains detect different harmonics, although in the majority of cases they detect the dominant one. We can then plot the frequencies at $t_{\text{ref}}$ of all the template harmonics that have SNR $\geq 5$, on a figure similar to figure 1. Although the indices of the harmonic in the template do not always correspond to the indices of the harmonic of the signal that it has matched, it is always clear from such a plot where the harmonics of the signal actually lie. If the chains do not agree on the identification of the harmonics, it is not always possible to tell, for instance, whether the dominant harmonic of the signal is $m = 2$ or $m = 1$. In order to estimate $f_\gamma$ uniquely, it is necessary to correctly identify the indices of at least one harmonic. If the identification is ambiguous, we run further chains for both possibilities. Once the harmonic indices of the detected harmonics have been inferred, we apply a least-squares fit to determine the three fundamental orbital frequencies. Harmonics with high SNR do not always lie closer to the true frequencies than lower SNR points. We see sometimes that lower SNR points match the frequency of the harmonics very well, but fail to fit the other waveform parameters, so the derivatives of frequency do not match.

Having determined frequency estimates, we force the orbital frequencies to be fixed at these values and then run a search on the remaining parameters. To do this, we use an MCMC search again, but we now use the ‘$N$-harmonic maximized’ likelihood to evaluate the acceptance ratio. The harmonics included in the maximization are those which have been identified in the preceding analysis. Initially, we turn off jumps in the orbital frequencies, but only until the SNR has reached a value which is better than the best attained by any of the chains prior to fixing the orbital frequencies. We then release the constraint and repeat the ‘improve frequencies’ step, but we now continue to use the ‘$N$-harmonic maximized’ likelihood in this and subsequent iterations. The aim of this procedure is to first find a good guess for the frequencies, then refine the other waveform parameters, then refine the estimates of the frequencies again and so on. If the initial guess is not too bad, a high SNR is achieved quite quickly when re-adjusting the other parameters. We repeat the procedure ‘improve frequencies–fix–release’ several times if required, but this iteration need not usually be repeated more than three times. In figure 2, we show the final result of a run on the data from the low-mass binary (1.3.4 type). In this example, the signal was found in the first and third half years of the data after only one iteration.

Before we conclude this subsection, we should reiterate that we need to detect at least three harmonics in order to be able to estimate the orbital frequencies. Moreover, two of these must have different $n$-numbers and two must have different $m$ and the same $n$. If the harmonics detected are not sufficient to determine the orbital frequencies, one can just use the frequency-refining proposal described above or use an $N - k$-harmonic template with the $k$ already identified harmonics excluded. This forces the search to look for other harmonics.

### 3.3. Parameter finalization

A good indication that we have found the signal is that the highest SNR chains cluster in parameter space. In other words, if all the chains with SNR close (say $\geq 90\%$) to the maximum SNR found across all chains, have similar parameters. A further indication is that these ‘best’ parameters are approximately the same for the different data subsegments. An illustrative example is shown in figure 2, where we see that the first half-year and third half-year searches are producing similar results.
Figure 2. Non-blind search for the low-mass binary. Results are shown for the search of the first (top) and third (bottom) half-year long segments. We show SNR versus parameter value for all chains for two cases: the inclination angle, \( \lambda \), (left) and the MBH spin parameter (right). The vertical line in each plot indicates the true value of the parameter.

At this stage, the accuracy of the recovered parameters is relatively poor because we have been searching shorter data segments—the total SNR is therefore lower and there is greater degeneracy between waveform parameters over a short duration of signal (one can accurately fit a small number of cycles in many distinct ways). The final stage of the search involves first reparametrizing the templates from all the chains in the different segments of data by their frequencies etc at a common reference time (usually \( t_{\text{ref}} = 0 \)) and then increasing the duration of the template first to one year and then to two. We then run an MCMC search with these longer waveforms, with chains starting at each of the high SNR points identified in the previous stage of the search. In figure 3, we show how the parameter estimation improves as we increase the template duration from half a year to two years.

We could continue to use the ‘\( N \)-harmonic maximized’ likelihood to begin with as we lengthen the waveforms, as it is more efficient to use than the physical model. However, there are two caveats to using this maximized likelihood—(i) If the number of harmonics included is large, it is very slow, as we need to maximize the likelihood for each harmonic (although, in practice, 5–8 harmonics are usually enough to build up an SNR that is comparable to the full 25 harmonic SNR). (ii) The maximization leads to a smoother but larger ambiguity (error) ellipsoid in parameter space. The smoothness helps the chains to reach the global maxima, but the fact that the maximum is quite flat gives a larger error in parameter estimation.

For these reasons, at this stage, we switch to physical templates, although continue to use the ‘3-harmonic physical’ likelihood to maximize the three phase angles automatically. To seed this final analysis, we need an estimate of the spin orientation, which is characterized by the two parameters \( \theta_K \) and \( \phi_K \). These represent the ecliptic colatitude and azimuth to which
the spin of the black hole would point if the black hole was translated to the solar system barycenter. An estimate for these parameters can be obtained from the estimated amplitudes of the harmonics, but the analytic expressions are so complicated that the inversion would have to be done numerically. Instead, we compute the likelihood for various spin orientations and take the one giving the highest value. However, there is a four-fold degeneracy in the angle $\phi_K$ and a strong degeneracy in $\theta_K$. This is shown in figure 4 for the blind search for the high mass MBH EMRI. This plot is color-coded by SNR, and the points were chosen randomly from a uniform distribution over the $\theta_K - \phi_K$ plane. One can see that it is hard to distinguish between four different values of $\phi_K$ and it is almost flat in $\theta_K$. To deal with this, we ran chains for several different choices of $\theta_K, \phi_K$ and in the end took the chain with highest SNR, which did find the correct values.

4. Results

In this section, we will describe some source-specific peculiarities encountered while we were analyzing the data sets. In each case, the signal was detected at a different stage of the search algorithm described in the previous section, and we will attempt to explain why this was
Figure 4. Degeneracy in the determination of the orientation of the MBH spin for the blind search of the high-mass binary (1.3.1 type).

Table 1. Results of the analysis of the five ‘blind’ data sets used in MLDC Challenge 1B.3. These are 1B.3.1–1B.3.5 going from top to bottom. The analysis of these data sets was not blind, as it was mostly finished after the parameters were released.

| Type   | ν (mHz) | μ/M⊙ | M/M⊙ | e₀     | θₜ | φₜ | λ | a/M² | SNR  |
|--------|---------|------|------|--------|-----|----|----|------|------|
| True   | 0.1920421 | 10.296 | 9517952 | 0.21438 | 1.018 | 4.910 | 0.4394 | 0.69816 | 120.5 |
| Found  | 0.1920437 | 10.288 | 9520796 | 0.21411 | 1.027 | 4.932 | 0.4384 | 0.69823 | 118.1 |
| True   | 0.34227777 | 9.771 | 5215577 | 0.20791 | 1.211 | 4.6826 | 1.4358 | 0.63796 | 132.9 |
| Found  | 0.34227742 | 9.769 | 5214091 | 0.20818 | 1.172 | 4.6822 | 1.4364 | 0.63804 | 132.8 |
| True   | 0.3425731 | 9.697 | 5219668 | 0.19927 | 0.589 | 0.710 | 0.9282 | 0.53326 | 79.5  |
| Found  | 0.3425712 | 9.694 | 5216925 | 0.19979 | 0.573 | 0.713 | 0.9298 | 0.53337 | 79.7  |
| True   | 0.8514396 | 10.105 | 955795 | 0.45058 | 2.551 | 0.979 | 1.6707 | 0.62514 | 101.6 |
| Found  | 0.8514390 | 10.106 | 955544 | 0.45053 | 2.565 | 1.012 | 1.6719 | 0.62534 | 96.0  |
| True   | 0.8321840 | 9.790 | 1033413 | 0.42691 | 2.680 | 1.088 | 2.3196 | 0.65829 | 55.3  |
| Found  | 0.8321846 | 9.787 | 1034208 | 0.42701 | 2.687 | 1.053 | 2.3153 | 0.65770 | 55.6  |

a The columns are: radial orbital frequency, mass of the CO, mass of the MBH, eccentricity at t = 0, ecliptic co-latitude, ecliptic longitude, inclination angle λ, spin of MBH, SNR recovered by template

the case. This section will be divided into two subsections. The first one is dedicated to the non-blind searches which were the basis for the algorithm development and tuning. The second subsection gives details of the ‘blind’ searches we did to test the search pipeline.

4.1. Round 1B analysis

While developing the algorithm, we analyzed the five ‘challenge’ data sets that were released within the MLDC round 1B. The bulk of the search tuning and development was done after the submission deadline, so we knew parameters of the signals. We used our knowledge of the parameters only to identify when the search was going in the wrong direction so that we could try other techniques and to identify when we had detected the signal. The results of these searches, for the intrinsic source parameters, are summarized in table 1.
Table 2. Results of the blind analysis of two data sets. For this analysis, we used the MLDC 1B.3.1 and 1B.3.4 ‘training’ data sets. Our analysis was blind in the sense that the search was run end-to-end without reference to the parameters used to generate the data sets. Our results were then compared to the known parameters at the end.

| Type   | $v$ (mHz) | $\mu/M_\odot$ | $M/M_\odot$ | $\epsilon_0$ | $\theta_S$ | $\phi_S$ | $\lambda$ | $a/M^2$ | SNR |
|--------|-----------|---------------|-------------|---------------|-------------|---------|--------|--------|-----|
| True   | 0.1674472 | 10.131        | 10397935    | 0.25240       | 2.985       | 4.894   | 1.2056  | 0.65101| 52.0|
| Found  | 0.1674462 | 10.111        | 10375301    | 0.25419       | 3.023       | 4.857   | 1.2097  | 0.65148| 51.7|
| True   | 0.9997672 | 9.7478        | 975650      | 0.360966      | 1.422       | 4.95339 | 0.5113  | 0.65005| 122.9|
| Found  | 0.9997626 | 9.7479        | 975610      | 0.360966      | 1.422       | 4.95339 | 0.5113  | 0.65007| 116.0|

The signal was detected at different stages of the search in the various cases. The easiest signals for this algorithm to find appear to be those from medium mass MBHs (data sets 1.3.2 and 1.3.3). We believe the reason for this is that the frequency evolution of the harmonics is not so great that the harmonics are hard to detect (which is true for the low-mass MBH case), but it is sufficient that the parameter space is not too degenerate (unlike the high-mass MBH case which is very degenerate). The second signal in table 1, 1.3.2, was detected without any iterations in the ‘improve frequencies–fix–release’ phase of the search. The second medium mass binary, 1.3.3, required more work (one iteration), primarily because of the lower SNR of this source.

The high- and low-mass MBH systems are more difficult to detect. The signal from the high-mass MBH EMRI has a lot of secondary maxima which are quite strong compared to the primary. These arise because the evolution in these systems is very slow, so it is easy to match harmonics for long periods of time with very different parameters. These secondary maxima are well separated and lie all over the parameter space. The analyzed signal was even worse than usual, because the inclination of the spin of the black hole to our line of sight was such that almost all the signal power was concentrated in a single $m$-harmonic for each $n$ (we encountered a similar case in our blind search and will discuss this later).

The low-mass MBH EMRI is our canonical EMRI, as we expect these systems to dominate the event rate [7, 17]. For these systems, the harmonic frequencies evolve rather significantly over the inspiral and so a template needs to match both the frequency and frequency derivative of a harmonic rather accurately in order to get high SNR. This means there are fewer secondaries, but, at the same time, it also implies that the global maximum is rather ‘sharp’ in parameter space, which is reflected in the better accuracy of recovered parameters. Usually these signals require more time on the first stage of the search (or a larger number of CPUs). The loudest signal (1.3.4—fourth in the table) was more difficult to detect because of its orientation ($\lambda$ is close to $\pi/2$). This caused us to make an incorrect guess of the dominant harmonic when doing the phase maximization. We had to use an $N$-harmonic template in the search with $N = 9$. The second signal (1.3.5) was not peculiar, but we had to do two iterations of the ‘improve frequencies–fix–release’ part of the search since our first guess of the orbital frequencies was not very good due to the lower signal SNR.

4.2. Blind tests

Since the high- and low-mass MBH EMRIs were the hardest to find, we decided to test the algorithm pipeline by performing blind tests on one data set containing a high-mass MBH EMRI, and one data set containing a low-mass MBH EMRI. We use the MLDC round 1B ‘training’ data sets 1.3.1 and 1.3.4 for this analysis, although we did not consult the parameter key until we had finished the search. The results are presented in table 2. We used two criteria
to determine the end point of the search and claim a detection: (i) several chains converged to the same result; (ii) the SNR of all harmonics in these best chains was comparable to and no less than the best SNR of the corresponding harmonic found in all the other chains.

For the high-mass MBH signal we did not encounter any difficulties. The search with the $N$-harmonic template resulted in quite a large error bar on the parameters because of the degeneracies in the parameter space and because of the relatively low SNR of this source. The degeneracy in $\theta_K, \phi_K$ for this source is illustrated in figure 4. The signal was found after one iteration of the ‘improve frequencies–fix–release’ stage.

The signal from the low-mass MBH EMRI proved much more interesting and difficult. Almost all the power was concentrated in the $m = 2$ harmonic (with $n = 2, 3, 4$). The ‘$N$-harmonic maximized’ SNR$^2$ for each of the 25 harmonics ($n = 1, \ldots, 5, m = -2, \ldots, 2$) are given in the matrix

$$F = \begin{pmatrix}
  m = -2 & -1 & 0 & 1 & 2 \\
  n = 1 & 1.98 & 1.38 & 1.52 & 5.94 & 205.53 \\
  n = 2 & 2.14 & 0.66 & 2.75 & 178.61 & 4677.62 \\
  n = 3 & 1.06 & 3.13 & 2.39 & 103.74 & 2109.76 \\
  n = 4 & 5.22 & 1.70 & 0.78 & 13.35 & 576.45 \\
  n = 5 & 2.85 & 1.61 & 3.27 & 4.73 & 145.50
\end{pmatrix}.$$ (25)

We were only able to detect three $m = 2$ harmonics ($n = 2, 3, 4$) with the chains, but a second $m$-harmonic is required in order to estimate the three frequencies. To achieve this, we used the method mentioned above: we constructed an $N$-harmonic template that did not include the harmonics which had already been identified in the search. This allowed us to find the $n = 2, m = 1$ harmonic and hence make a preliminary estimation of all three orbital frequencies. We then required three iterations of the ‘improve frequencies–fix–release’ search to reach the final answer. At this stage, we were confident about the quality of the detection and, when we compared to the true parameters, we had indeed reached a very high accuracy for all parameters.

5. Discussion

We have described an algorithm for the detection of EMRI signals in LISA data. This algorithm is based on running multiple MCMC search chains simultaneously. However, it relies on the key refinement that all the secondary solutions that different chains have identified are used together to constrain further movement of the chains. It is clear from the results in tables 1 and 2 that the algorithm is able to robustly and accurately find the true solution, in the simplified situation that we are searching for a single, bright EMRI source buried in Gaussian instrumental noise. We have successfully found the source in seven out of seven mock data streams, even when the parameters were such that the waveform had unusual features, e.g., orbital inclination close to $\lambda = \pi/2$. This gives us reason to expect that the algorithm will work equally well in all comparable situations, independent of the parameters of the source.

This is the first algorithm to be published in the literature that has been demonstrated to be able to detect and determine parameters for a ‘typical’ EMRI signal—a $10M_\odot$ black hole falling into a $10^6M_\odot$ black hole—which we expect to dominate the LISA event rate [7, 17]. It is particularly gratifying that our parameter recovery is now reaching the theoretical level that was estimated from Fisher matrix analyses [9]—MBH mass and spin determinations at the level of $10^{-4}$, and sky position accuracies of $10^{-3}$. While we would expect the Fisher matrix to accurately represent the shape of the global maximum for these high SNR sources, it is a purely local analysis and therefore does not account for the presence of secondary maxima.
The fact that our algorithm can now find the global maxima from among the bright secondaries bodes well for using EMRI sources for high precision astrophysical observations [7, 8].

The algorithm can still be improved further. The main problem at present is that the identification of secondaries in order to determine constraints on the orbital frequencies is done by hand. We stop the search chains after stage 1, and then by hand examine the results in order to estimate suitable proposal distributions for the frequencies in the second stage. In principle, this could all be done automatically, although it would require communication between different chains. If each chain had information about the best points found by all the other chains, then an adaptive proposal could be constructed from this information. The resulting search would be more like a population MCMC search. The other area of the search that could potentially be improved is the way in which we use annealing. The annealing scheme was borrowed almost verbatim from SMBH searches [18] and we have not attempted to optimize it for the EMRI problem. While this is clearly not affecting the ultimate convergence of our search, the convergence speed might be improved by modifying the annealing scheme. Other MCMC variants, for instance, parallel tempering, might also improve the efficiency of the search. Parallel tempering has been demonstrated in LISA searches to great effect [36], but we have no immediate plans to include it in the search pipeline.

The algorithm as described here has only been demonstrated in a significantly simplified scenario—detection of a single, bright EMRI buried in purely instrumental Gaussian noise. The real LISA data stream will be very different and is expected to contain many thousands of resolvable signals which will be overlapping in time and frequency, in addition to a noise foreground from galactic compact binaries and non-Gaussian instrumental artifacts etc. It is not clear how well this search will perform under those circumstances, since it relies on being able to identify all the secondary peaks in the likelihood surface that are associated with the same signal. The relative SNRs and track shapes will provide powerful discriminators for this purpose, but there will inevitably be problems distinguishing a dim sideband harmonic of a bright source from the dominant harmonic of a similar but much more distant source. The best way to explore these complications is to attempt to analyze more realistic data sets. The next round of the MLDC, Challenge 3, includes an EMRI data set that contains five overlapping signals of low SNR. We will begin to explore source confusion by using that data set as a test case. As future MLDC releases become increasingly realistic, we will analyze them in order to demarcate where this algorithm fails and how it can be improved to cope with this greater realism.

Another aspect where LISA EMRI data analysis requires further work is in the choice of waveform template family employed in the search. The analytic kludge waveforms employed here capture the main features of true EMRI signals, but will not be faithful, i.e., they will not be a good match to a true waveform of nature with the same parameters. However, they may still be ‘effective’ in the sense that analytic kludge waveforms might mimic a true waveform with different parameters, or at the very least might match individual harmonics of the true waveform well. This needs to be assessed as a matter of some urgency and can be done by injecting a more accurate numerical kludge [26] or perturbative [27] waveform into a data stream and then searching for it with an algorithm, such as this one, built on the analytic kludge model. If the analytic kludge waveforms are effective and the mapping between the recovered parameters and the true parameters can be understood, then more accurate waveforms would only be required for the final, parameter refinement, part of the search. However, it is more likely that the analytic kludge will not be sufficiently effective and so more faithful waveforms, such as the numerical kludge [26], will be required throughout the search pipeline. Nonetheless, the basic form of the search outlined here, based on harmonic
identification and refinement, should still work and only the practical implementation will need to be updated.

Acknowledgments

The work of SB and EKP was supported in part by DFG grant SFB/TR 7 ‘Gravitational Wave Astronomy’ and by DLR (Deutsches Zentrum für Luft- und Raumfahrt). JG acknowledges support from the Royal Society and thanks the Albert Einstein Institute for hospitality and support while this work was being completed. We would also like to thank the Aspen center for physics where part of this work was conducted.

References

[1] Danzmann K et al 1998 LISA Laser Interferometer Space Antenna, Pre-Phase: A Report Report MPQ 233 Max-Planck-Institute fur Quantenoptic
[2] Ryan F D 1995 Phys. Rev. D 52 5707
[3] Collins N A and Hughes S A 2004 Phys. Rev. D 69 124022 (arXiv:gr-qc/0402063)
[4] Glampedakis K and Babak S 2006 Class. Quantum Grav. 23 4167 (arXiv:gr-qc/0510057)
[5] Gair J R, Li C and Mandel I 2008 Phys. Rev. D 77 024035 (arXiv:0708.0628)
[6] Babak L and Cutler C 2007 Phys. Rev. D 75 042003 (arXiv:gr-qc/0612029)
[7] Gair J R 2008 arXiv:0811.0188
[8] Amaro-Seoane P, Gair J R, Freitag M, Miller M C, Mandel I, Cutler C J and Babak S 2007 Class. Quantum Grav. 24 113
[9] Babak L and Cutler C 2004 Phys. Rev. D 69 082005 (arXiv:gr-qc/0310125)
[10] Babak L and Cutler C 2007 Phys. Rev. D 75 042003 (arXiv:gr-qc/0612029)
[11] Arnaud K A et al 2006 AIP Conf. Proc. 873 619 (arXiv:gr-qc/0609105)
[12] Arnaud K A et al (Mock LISA Data Challenge Task Force) 2006 AIP Conf. Proc. 873 625 (arXiv:gr-qc/0609106)
[13] Arnaud K A et al 2007 Class. Quantum Grav. 24 S551 (arXiv:gr-qc/0701170)
[14] Arnaud K A et al 2007 Class. Quantum Grav. 24 S529 (arXiv:gr-qc/0701139)
[15] Babak S et al 2007 Class. Quantum Grav. 25 184026 (arXiv:0806.2110 [gr-qc])
[16] Gair J R, Barack L, Creighton T, Cutler C, Larson S L, Phinney E S and Vallisneri M 2004 Class. Quantum Grav. 21 1595
[17] Gair J R, Barack L, Creighton T, Cutler C, Larson S L, Phinney E S and Vallisneri M 2004 Class. Quantum Grav. 21 1595
[18] Gair J R, Barack L, Creighton T, Cutler C, Larson S L, Phinney E S and Vallisneri M 2004 Class. Quantum Grav. 21 1595
[19] Cornish N J and Cutter M C, Mandel I, Cutler C J and Babak S 2007 Class. Quantum Grav. 24 1595
[20] Cornish N J and Cutter M C, Mandel I, Cutler C J and Babak S 2007 Class. Quantum Grav. 24 1595
[21] Jaranowski P, Krolik A and Schutz B F 1998 Phys. Rev. D 58 063001 (arXiv:gr-qc/9804014)
[22] Babak S, Fang H, Gair J R, Glampedakis K and Hughes S A 2007 Phys. Rev. D 75 024005 (arXiv:gr-qc/0607007)
[23] Drasco S and Hughes S A 2006 Phys. Rev. D 73 024027 (arXiv:gr-qc/0509101)
[24] Tinto M and Dhurandhar S V 2005 Living Rev. Rel. 8 (http://www.livingreviews.org/lrr-2005-4)
[25] Owen B J and Sathyaprakash B S 1999 Phys. Rev. D 60 022002 (arXiv:gr-qc/9804018)
[26] Key J S and Cornish N J 2009 Phys. Rev. D 79 043014 (arXiv:0812.1590)