A principal component analysis of gravitational-wave signals from extreme-mass-ratio sources

Edward Joseph Bloomer

Submitted in fulfilment of the requirements for
the Degree of Ph.D.

University of Glasgow

Institute of Gravitational Research

Department of Physics and Astronomy

April 2010

Abstract

The Laser Interferometer Space Antenna (LISA) will detect the gravitational wave emissions from a vast number of astrophysical sources, but extracting useful information about individual sources or source types is an extremely challenging prospect; the large number of parameters governing the behaviour of some sources make exhaustively searching this parameter space computationally expensive.

We investigate the potential of an alternative approach, with a focus on detecting the presence of particular inspiraling binary source signals within a timeseries of gravitational wave data, and quickly providing estimates of their coalescence times. Specifically, we use Principal Component Analysis (PCA) to identify redundancy within the parameter space of Extreme Mass Ratio Inspiral (EMRI) sources and construct a new, smaller parameter space containing only relevant signal information. We then create a simple search method based on how gravitational wave signals project into this new parameter space.

Test cases indicate that a small number of principal components span a space occupied by the majority of EMRI spectrograms, but non-EMRI signals (including noise) do not inhabit this space. A PCA-based search method is capable of indicating the presence of gravitational waves from EMRI sources within a new test spectrogram.

The results of our PCA-based searches show that the method could be used to provide initial estimates of EMRI coalescence times quickly, to be used as initial data for a more thorough search.

Contents

List of Tables

List of Figures

Nomenclature

Waveform time Time with respect to the beginning of a searchable waveform.

$\mathbf{R_2}$ $1 - R$, used to reveal structure in residual measurements when $R \approx 1$.

R The residual; the difference in power between a spectrogram and its reconstruction using a set of eigenspectrograms, expressed as a fraction of the power contained by the original spectrogram.

$\mathbf{WP_N}$ The Weight Power of a test spectrogram in a searchable waveform containing only noise, but no EMRI signals.

$\mathbf{WP_S}$ The Weight Power of a test spectrogram in a searchable waveform.

eigenspace The space spanned by an eigenspectrogram basis set.

eigenspectrogram An eigenvector of the covariance matrix whose columns define spectrograms generated from parameters in the problem space. Eigenspectrograms are normalised and are therefore a basis vector for the eigenspace, but otherwise can be used interchangably with *principal component*.

Parameter space In the context of this thesis, the parameter space is the space spanned by the parameter values of EMRI analytic kludge waveforms. The EMRI signals generated from points within this parameter space are further manipulated, and the space they span is referred to as the problem space.

Principal component A vector accounting for as much variance in a data set as possible. Sets of principal components are orthogonal, and can reduce the dimensionality of the problem space.

Problem space The subspace of the parameter space spanned by the signals that will be analysed. In the context of this thesis the problem space is occupied by spectrograms containing EMRI signals, a subset of all possible spectrograms.

Searchable waveform The timeseries of gravitational wave strain measurements that will be searched.

Time from coalescence Time between a measurement and the coalescence time of an EMRI signal within a searchable waveform.

WP Weight Power. The total power of a spectrograms' projection into an eigenspace, calculated from the power contribution of each eigenspectrogram in the eigenspace.

Acknowledgements

I don't think anyone ever finds writing a thesis *easy* (and those that do should share that secret), but I will admit to finding it more difficult than I had ever imagined. For their assistance in producing this piece of work and throughout my time as a PhD student at the University of Glasgow, I am indebted to a number of people and wish to acknowledge them below.

First, I am extremely grateful to my supervisors Graham Woan and Siong Heng for their support, encouragement, insight and expertise, without which I could not possibly have completed my research, nor extracted anything useful from it. Their help was truly invaluable. Thanks too to STFC for providing the funding which allowed me to support myself during the research.

I would further like to express my gratitude to the members of the Institute of Gravitational Research and indeed the department of Physics and Astronomy as a whole; I have been in the enviable position of being able to work in a building full of people I admire on a personal and professional level. The same has been true of those members of the gravitational wave community around the world that I have met, with such consistency that I have begun to suspect some sort of benevolent conspiracy. Even so, I must single out my officemates over the years; Matt Pitkin, Matt Edgar and Chris Messenger, who were always willing to offer guidance and encouragement and continually put up with queries that began "um, this might by a stupid question, but..."[1] with good grace.

[1] They usually were. I still maintain there must have been a few sensible questions during the three years, but I cannot think of a single one.

Finally, I shall risk offending some by placing a great number of family members and friends into a group I shall call “people-who-have-no-more-than-a-passing-interest-in-gravitational-waves”. It is my incredible good fortune that they are genuinely too numerous to name individually, but I would have been lost without those whose support had nothing to do with my work. Although perhaps possible in some technical sense, the last three years would have been unbearable without your company. Thank you.

Declaration

This thesis presents my original research work. Every effort has been made to clearly acknowledge the contributions of others where appropriate, and include references to the relevant literature.

The work was done under the guidance of Graham Woan and Ik Siong Heng of the Institute for Gravitational Research, part of the Department of Physics and Astronomy for the University of Glasgow.

May 1, 2010

Chapter 1

Introduction

We are on the verge of transforming our understanding of the universe by directly observing *gravitational waves*. While the observations of astrophysical sources in electromagnetic spectrograms have produced a vast wealth of scientific discoveries, the gravitational wave spectrum stands separate from (but complimentary to) these. Thus, detection of the gravitational waves allows us to search a realm of source properties that are otherwise imperceptible; an exhilarating prospect, but one replete with its own quirks and difficulties. This Chapter provides a brief introduction to gravitational waves; the principles of detection, likely sources and some current data analysis techniques.

1.1 Introduction to gravitational waves

Einstein's theory of general relativity revolutionised the way in which physicists thought about the nature of spacetime. The theory revised the Newtonian concept of gravity, describing it in terms of the geometry of space-time curved by the presence of massive compact objects. Further, changes to an object's gravitational field can generate 'ripples' in the background curvature that propagate from their source-gravitational waves.

The 'wave' nature of gravitational radiation is a consequence of small perturbations to the local spacetime metric (the underlying geometry of the spacetime) in the linearised Einstein Equations, a weak-field approximation of a more general relation between the matter and energy distribution and the curvature of spacetime known as the Einstein Equations. The solution to the linearised equations is most

simply expressed as plane waves with a propagation speed equal to that of the speed of light.

To determine the luminosity and amplitude of gravitational waves from a system of masses, we calculate the traceless quadrupole moment tensor of that system. The Einstein equations are too difficult to solve analytically for a general case of a system of gravitating masses, but in a multipole expansion representing the distribution of mass-energy in the system the mass quadrupole is the leading order that can have a nonzero time derivative. The mass monopole, representing the total mass-energy in the system does not change over time, nor does the dipole representing the centre of mass of the system. The contributions of higher-order multipoles are present as well, but are much smaller. If we ignore them, we are left with the quadrupole approximation, the lowest-order method for expressing the amplitude of radiated gravitational waves.

The classic illustration of the effects of gravitational waves is the deformation of particles arranged in a ring. Gravitational waves propagating perpendicular to the plane of the ring will alter the proper distance between particles and the centre of the ring (as shown in Fig. 1.1). In contrast to electromagnetic waves with polarisation states rotated 90° with respect to each other, gravitational wave polarisations are rotated only 45°. These two polarisation are labelled *plus* (+) and *cross* (×), and the gravitational strains (a dimensionless measure of the magnitude of the spacetime perturbation in terms of the proper distance between particles) are denoted h_+ and $h_\times$.

It is difficult to provide an estimate for the 'typical' strain produced from astrophysical sources because there is no typical source of emission, although Schutz (1) provides a quick estimate of 10^{-20} produced by a binary system of two solar-mass stars 5 kpc from Earth. However, in the quadrupole approximation for the gravitational radiation the strain of the waves is inversely proportional to the distance from the source, so more distant sources will produce even smaller strains.

Thus far there has been no direct observation of these gravitational waves made by measuring the change in proper distance between particles, but general relativity has enjoyed considerable success; amongst other triumphs, it was able to predict the precession of Mercury's orbit, and the orbital decay of the Hulse-Taylor binary pulsar (2). Therefore, there is extremely strong motivation to determine just how viable the the theory is as a description of our universe.

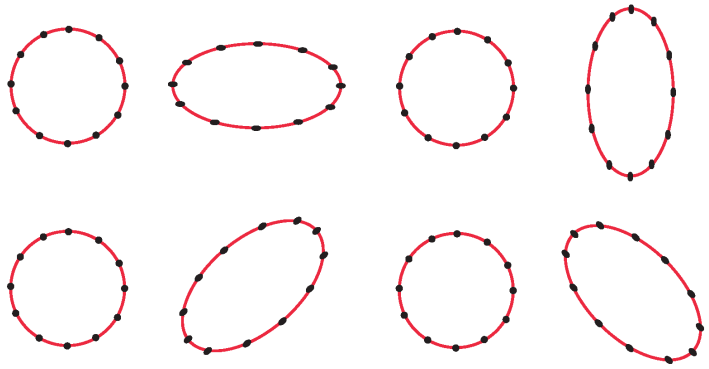


Figure 1.1: The effect of gravitational waves on a ring of particles. The upper row shows the distortion cause by a + polarised gravitational wave propagating perpendicular to the plane of the ring, and the lower row a × polarised wave propagating in the same direction.

1.1.1 Gravitational wave sources

There are a number of possible sources of gravitational wave emission, but until direct observations are made we cannot be certain how insightful our predictions are. In particular, some of our uncertainty about the event rate of certain sources of emission and the spatial density of sources are a result of relying on electromagnetic observations; such observations may not be useful in estimating these values, since we may have sources emitting gravitational radiation that we cannot detect electromagnetically. Nevertheless, expected sources generally fall within a few categories.

1.1.1.1 Burst sources

Core-collapse supernovae, leading to the formation of black holes and neutron stars, will result in a burst of gravitational waves that accompanies this violent explosion (3). Uncertainty regarding the collapse process (and how non-spherical this will be) and the precise form of the emission prevent accurate predictions of the waveform produced in this type of event, but rough estimates do suggest that current ground-based detectors could be fairly confident of detecting these sources (4),.

Additionally, binary systems of compact astrophysical objects (neutron stars or black holes) will radiate energy as gravitational waves, reducing the distance between

the binary components. The evolution of these systems will cause the emitted gravitational waves to increase in frequency as time goes on, referred to as a *chirp*; the observation of the Hulse-Taylor binary orbit evolution and the degree to which it matches the changes predicted by general relativity as a result of gravitational radiation leaves little doubt that this is the responsible mechanism(5). Eventually, the two objects will merge and the remaining object will ringdown, producing yet more gravitational waves.

Each of these periods of evolution is worthy of study in its own right, but of particular significance is the prediction that the disruption of neutron stars by black hole companions is one of the possible sources of gamma-ray bursts, suggesting that coincident searches could be made for these objects in two different regimes, cementing the collaborative relationship between the gravitational wave research community and that of other astrophysical research.

1.1.1.2 Periodic sources

There will be no gravitational waves emitted from objects that are perfectly symmetric around their rotation axis due to the quadrupolar nature of the radiation; some kind of asymmetry is necessary. The typical detectable source of this kind is a rotating neutron star deformed by 'bumps', producing an asymmetrical distribution of its mass and hence gravitational wave emission (3). These bumps may be produced by distortions in the solid crust of the star or the result of magnetic pressure (the magnetic field and the rotation axis may be misaligned), but whatever the mechanism supportable deformations are likely to be small.

Current efforts in this area include observations of the Crab pulsar (PSR B0531+21) in the Crab Nebula. The LIGO gravitational wave observatories have failed to detect gravitational waves from this neutron star pulsar, placing an upper limit on contribution, through the the mechanism of gravitational radiation, to the loss of kinetic energy that would cause the observed spindown rate of the pulsar (6).

1.1.1.3 Stochastic sources

Large numbers of unresolved discrete gravitational wave sources will produce a stochastic background of gravitational radiation (3). Additionally, backgrounds will be produced by processes in the early universe such as the Big Bang and inflation, and detecting these sources would allow us to probe the fundamental nature of the

universe in the early stages of its development (where energies were beyond anything we could hope to recreate in a laboratory), and would subsequently allow us a greater understanding of the cosmological make up of the universe.

1.1.1.4 Man-made sources

A man-made source of gravitational waves is an attractive concept, providing a test-bed for detectors before considering astrophysical sources. While *any* detection would be an extremely exciting, a examination of the simple setup described by Sathyaprakash and Schutz (3) shows that the signals produced by artificial means would be far too weak to detect. Even if we were to propose a similar project on a more ambitious scale (a centrifuge generator consisting of two equal-mass blocks joined by a stiff rod with total mass of 2×10^4 kg rotating at 100 Hz around the the middle point of the rod, say), the resulting gravitational wave strain is still $h \approx 1 \times 10^{-33}/r$, where r is the distance from the generator (at least one wavelength away to distinguish the wave from near field fluctuation). To say nothing of the difficulty in creating such a wave source, the resulting strain amplitudes remain many orders of magnitude below even the most optimistic estimates for detector sensitivities, forcing us to abandon the notion of artificial gravitational wave sources. Generating detectable strain lies firmly in the astrophysical realm.

1.2 Detecting gravitational waves

Gravitational waves signals are extremely weak, making their detection very challenging. Precision measurement is key; detectors must be incredibly sensitive lest gravitational wave signals drown in unwanted noise.

1.2.1 Bar detectors

The first gravitational wave detectors, pioneered by Joseph Weber in the 1960s (7), were resonant mass, or bar, detectors. A typical detector of this type is a large aluminium cylinder with a narrow resonant frequency, deviations from which can be caused by passing gravitational wave bursts exerting tidal strains on the bars. Nevertheless, noise (in particular, thermal noise) greatly exceeds the amplitude of vibrations caused by the burst, requiring coincident detections from two or more cylinders to make reliable claims of gravitational wave detections. Unfortunately,

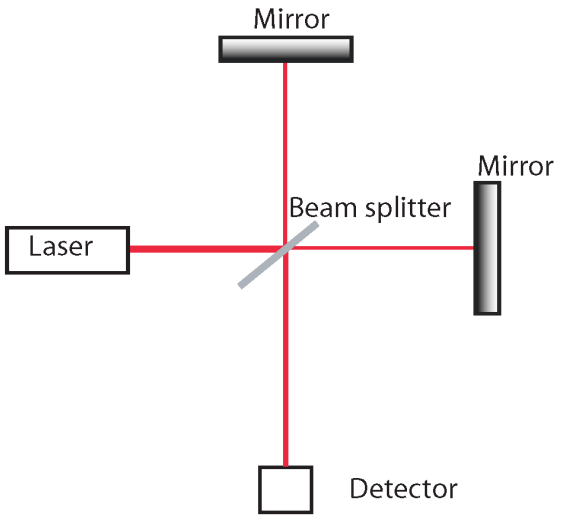


Figure 1.2: A simple Michelson interferometer. Passing gravitational waves will cause the laser arm lengths to change allowing direct detections to be made via the phase shifts in the recombined light.

despite some early candidates, and a great deal of work to resolve the problems inherent in bar detectors, no compelling evidence exists of a significant event. At this time bar detectors have fallen out of favour because of their narrow bandwidths and insufficient sensitivity, and only two groups continue to work with these devices ((W..b)[1],(W..p),(8),(9)), although they continue to collaborate with other research efforts ((10)).

1.2.2 Interferometers as gravitational wave detectors

Although bar detectors could achieve useful sensitivities within narrow bandwidths, the desire for wide-band detectors (to help observe different or frequency-changing sources) prompted several groups to explore the potential of laser interferometers. The basic concept can be easily illustrated by a simple Michelson interferometer (see Fig. 1.2), although an effective construction is much more involved ((11) provides an excellent introduction).

[1]Due to their rather temporary nature, web-based references have been included in a separate 'Webliography' after the References section at the end of the thesis, and were correct at the time of printing. Webliography entries have the prefix 'W..' or 'X..'.

Passing gravitational waves will distort the proper distances between freely-floating particles. In an interferometry setup, laser light is passed through a beam splitter, bounced off suspended mirrors, and recombined. If a + polarised gravitational wave of amplitude h propagates out of the page, the proper distance in the arms of the interferometer will be reach a maximum of $L + h/2$ and a minimum of $L - h/2$. The fractional change in the arm length is then

$$\frac{\triangle L}{L} = \frac{h}{2}. \tag{1.1}$$

The interference pattern in the recombined light, caused by the unequal arm lengths allows the fractional change to be measured and by extension the amplitude of the gravitational wave. Frequency information is gleaned from the time varying displacement of the mirrors.

The phase shifts in the light returning to the beam splitter are small, so current interferometers use Fabry-Perot cavities in the arms (12): two partially-reflecting mirrors that allows light incident on the input mirror to resonate, producing larger phase shifts and increasing the laser power in the optical system.

Noise in interferometric detectors

Because the gravitational waves will only produce a small displacement, noise threatens to obscure measurements from the interferometer. Each type of noise may be classified as technical (that is, relating to the instrument itself or the surrounding enviroment) or fundamental (imposed by theoretical limitations to measurements), and the following section examines some of the main sources. An example noise profile is presented by Fig. 1.3.

Seismic and vibrational noise is produced by vibrations in the external environment, and is prevalent at frequencies below a few Hertz (3). Nevertheless, by housing the instrument in vacuum and suspending the mirrors on sophisticated pendulums, the interferometer can be largely isolated from this noise. Ground-based detectors are also vulnerable to low frequency gravity gradient noise from changes to the local gravitational field. Although this noise falls sharply at higher frequencies (dropping below contributions from other sources), it cannot be screened from the instrument and can only be avoided by placing the interferometer in space (see 1.2.3).

At higher frequencies (12), thermal noise affects components of the detector such as the mirror, coatings and suspension systems. Typically the interferometers do

not operate near the resonant frequencies of the components (systems tend to oscillate at larger amplitudes near their natural frequency of vibration, a few Hertz for pendulum suspension systems), but careful consideration must be given to choosing materials that confine thermal vibrations to small bandwidth around their resonant frequencies, leaving measurement frequencies as free as possible (desirable materials have a high quality factor Q). Furthermore, components can absorb small amounts of power from the laser causing heating that will change their refractive properties, and time-dependent variations in these changes can occur at measurement frequencies- the behaviour of the components must be carefully characterised to determine the amount of laser power available for use in the detector.

In contrast, increased laser power will decrease the error caused by random variations in light intensity known as shot noise. This Poisson process produces an improvement in the error as $\sqrt{N}$ for mean number of N photons at the output, but the desired levels of laser power are only achieved by power-recycling, being beyond the output of available continuous lasers. Furthermore, the position sensing accuracy achieved by increasing the laser power is at the cost of displacements caused by transferring more momentum to the mirrors which can obscure gravitational wave effects. Essentially, one is in conflict with the Heisenberg uncertainty principle, but there are techniques for 'squeezing' the output signal to improve sensitivity beyond the standard quantum limit (the level of noise caused by quantum fluctuations of the laser beam's electric field) (13).

Current interferometers

The Laser Interferometric Gravitational Wave Observatory (LIGO) and Virgo gravitational wave observatories are the two largest gravitational wave interferometry projects. The US-based LIGO (W..j) has three detectors at two sites: a single detector in Livingston, LA with 4km arm lengths (designated L1), and two detectors at Hanford, WA with 4km and 2km arm lengths (H1 and H2 respectively). The French/Italian collaboration Virgo detector ((W..v),(14)) is a 3km facility at Cascina near Pisa. After 'engineering runs' designed to debug systems, LIGO began data collection on the first 'Science Run' (S1) with all three interferometers, achieving its design sensitivity in March 2005 after several more Science Runs (X..b). Virgo, having finished commissioning in 2007, joined the last part of the fifth Science Run (S5), taking data coincidentally with LIGO (the status of each facility is explored in (15) and (16)).

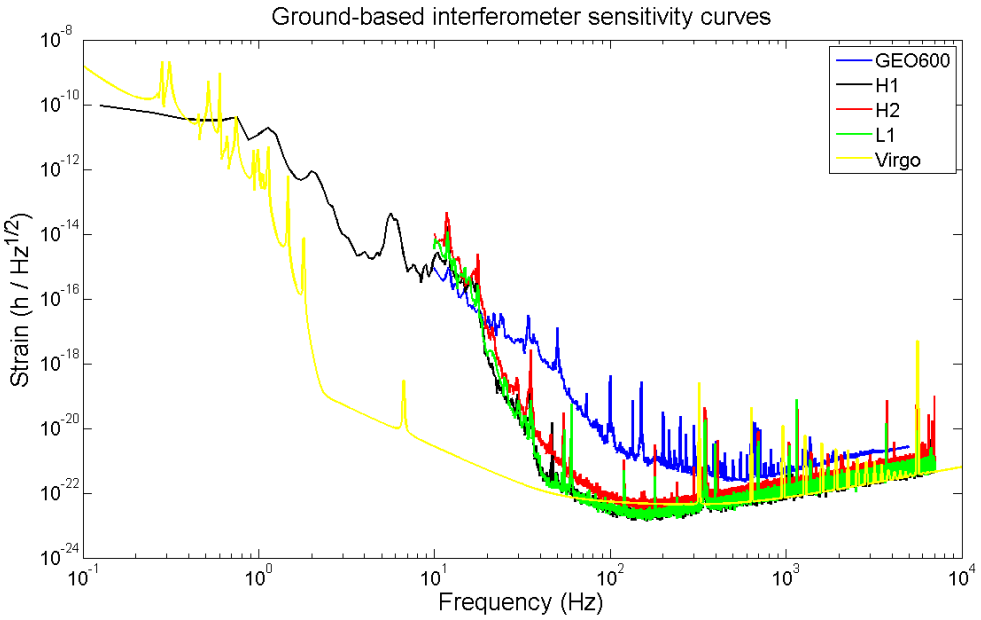


Figure 1.3: Sensitivity curves of ground-based interferometer gravitational waves detectors ((W..w),(W..k)). The designations of the detectors are given below in the main text.

In addition to these large detectors, smaller interferometers are contributing to the global network of gravitational wave observatories. The British/German GEO600 detector (W..f) (a 600m interferometer located near Hanover, Germany), cooperates with LIGO as part of the LIGO Scientific Collaboration [1] and has been a test bed for next-generation technologies due to be incorporated into LIGO in a future upgrade, Advanced LIGO. The 300m Japanese TAMA300 (W..s) has also undergone data taking sessions (17) (and was in fact the first large scale interferometer to make successful continuous measurements), and the small Australian International Gravitational Observatory interferometer (W..c) as part of the Australian International Gravitational Research Centre (AIGRC) (W..d).

The benefits of a global search goes beyond the opportunity to test different approaches and technologies within the general framework of an interferometer. A single detector is not able to measure both independent polarisations of a passing gravitational wave, receiving a combination depending on the orientation of the detector and the location of the source. Therefore, if we wish to recover the polarisation information and locate the source accurately we need multiple, well-separated detectors making coincident measurements. Collaborative efforts are a fundamental part of gravitational wave research.

Future ground-based detectors

The majority of work on future ground-based detectors at this point is to incorporate technological advances into existing detectors. The plan to create Advanced LIGO by including technology from GEO600 is already underway, with initial LIGO operations due to conclude in 2010 and the upgraded detectors to be operational by 2015. New test masses, more laser power and new suspension systems are expected to increase Advanced LIGO's limiting sensitivity by a factor of ten compared to the initial LIGO instruments (there is an enhanced readout upgrade planned before the major new installations, for use with S6 in 2009).

Similarly, Virgo has begun a two-stage upgrade: Virgo+, that began in May 2008, will improve electronics, increase laser power and install new payloads. The following upgrade to Advanced Virgo, with the goal of increasing the detector sensitivity by an order of magnitude across the detection band can begin sometime after 2010.

[1] While it pools data and analyses it jointly with the LSC, Virgo is not a member of the collaboration.

Additionally, there are benefits to be gained from new facilities, such as the Large-scale Cryogenic Gravitational wave Telescope (LCGT)(W..i) in Japan with 3km arm lengths, or the Australian Interferometric Gravitational Observatory (AIGO) proposed by the Australian Consortium for Interferometric Gravitational Astronomy (see (18) for a preliminary design and discussion). As well as allowing construction of facilities using the knowledge gained from previous efforts, new sites can help to triangulate gravitational wave sources and provide more statistical confidence in coincident events (19). Nevertheless, while new detectors may be highly desirable, they require considerable amounts of funding for their construction.

1.2.3 The LISA mission

The inescapable limits of gravity gradient noise and baseline size for ground-based detectors can be avoided by a spaced-based interferometer, and the Laser Interferometer Space Antenna (LISA) is a proposed mission jointly supported by NASA and ESA. The goal of the LISA mission is to detect gravitational waves at low frequencies ($10^{-4} - 10^{-1}$ Hz) inaccessible to existing detectors, providing data for the research of gravitational radiation, astrophysics and fundamental physics, complimenting the existing detector network rather than supplanting it.

The current design for LISA is a triplet of spacecraft housing freely-floating test masses. Each spacecraft is positioned at the vertex of an equilateral triangle formation, passing laser light between each other to create three Michelson-type interferometers that are separate although not fully independent (see Fig. 1.4). In this manner, LISA acts as an gravitational *observatory* rather than simply a detector, and should be able to provide source location information as well as direct detections of gravitational waves. The test masses are shielded from external disturbances (such as radiation pressure), and are seperated by approximately 5×10^9 m, with the centre of the formation trailing the Earth's orbit by 20°.

The engineering requirements of the spacecraft are extremely demanding; not only must the interferometers be capable of detecting the minute changes to the test mass separations, the spacecraft must survive in space for a long period of time. There will be no opportunity to repair LISA once in orbit either, and the component spacecraft must survive the launch process intact. Additionally, LISA

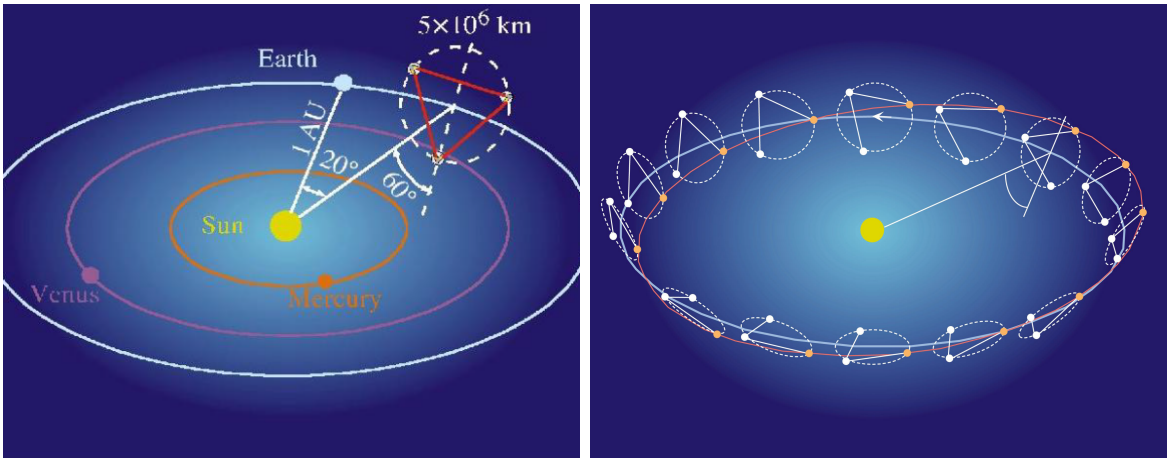


Figure 1.4: Orbital configuration of LISA and its annual orbit. *Image courtesy of NASA (X..d).*

must cope with the divergence of the laser beams over their vast separations, and other considerations such as the movement of the individual spacecraft as they orbit.

Furthermore, fluctuations in the six laser signals from the optical benches produce Laser Noise that can obscure gravitational wave sources. This can be reduced using Time Delay Interferometry (time shifting and combining individual optical bench measurements)(20), but a full solution for the required combinations taking into account the final orbital configuration of LISA is not yet available (W..e)

A great deal of effort has already gone into the conceptual and physical work required to create LISA, with a technology-demonstration mission called LISA Pathfinder (X..c) due to launch in 2011. The ultimate fate and launch date of the full LISA mission is much more difficult to estimate however, due to scale and unique challenges of the project, but at the time of writing the estimated schedule expects the earliest launch date to be 2020.

1.2.4 Gravitational wave data analysis

Not only are there intimidating technical challenges in creating instruments sensitive enough to detect gravitational waves, thoughtful data analysis is required to extract signals from the detector output.

Crucially, the large number of gravitational wave sources expected to be visible to interferometers and long periods of continuous data-taking means searches must

be performed in a large parameter space over a wide range of frequencies (thanks to the broadband nature of the detector). Not only is this computationally intensive, but sensitivity is enhanced by robust understanding of the expected waveforms for use in the search algorithms. Furthermore, the polarized nature of the gravitational radiation requires a network of detectors (discussed previously) so these algorithms must be capable of manipulating data from multiple detectors to aid in coincident analysis and event detection. Gravitational waves therefore present a rather atypical situation compared with the requirements of more conventional detection of astrophysical sources using electromagnetic radiation.

Development and discussion of data analysis techniques can be found in the annual Gravitational Wave Data Analysis Workshop (GWDAW) meetings, which provide a forum for interested parties to refine or revise their strategies. Recently, the GW Notes newsletter (W..g) has been launched in an attempt to collect efforts from a number of communities studying gravitation, including data analysis efforts. The following section details some key concepts in gravitational wave signal data analysis, and the problems hindering their successful implementation.

1.2.4.1 Matched filtering techniques

Simply put, matched filtering is the correlation of known waveforms (known in this context as templates or filters) with detector output in an attempt to detect the signal in the data, which may be buried in noise (3). A common technique in the telecommunication industry, matched filtering sets itself the task of finding the 'optimal' template; the waveform that will produce the highest signal-to-noise ratio (SNR). The technique as applied to the field of gravitational waves is most easily explained by first establishing a firm mathematical framework.

The scalar product of two functions $x(t)$ and $y(t)$ is defined as

$$\langle x, y \rangle \equiv 2 \int_0^\infty \frac{\tilde{x}(f)\tilde{y}^*(f) + \tilde{x}^*(f)\tilde{y}(f)}{S_n(f)} df, \tag{1.2}$$

(3)

where $\tilde{x}(f)$ denotes the Fourier transform of the function $x(t)$, defined as

$$\tilde{x} = \int_{-\infty}^{\infty} x(t) e^{2\pi i f t} dt, \tag{1.3}$$

$\tilde{x}^*(f)$ is the complex conjugate of $\tilde{x}(f)$and $S_n(f)$ is the one-sided PSD of the noise in the detector (the power per unit frequency in the interval between 0 Hz and the bandwidth of the detector) The noise is assumed here to be stationary, although this may not be true.

In the context of gravitational waves, our function $x(t)$ will be a time series containing a signal $h(t)$ with an arrival time t_a and noise $n(t)$ thus taking the form $x(t) = h(t - t_a) + n(t)$, while $y(t)$ is a template. The correlation of the two is

$$corr.(\tau) = \int_{-\infty}^{\infty} x(t)y(t+\tau)dt, \tag{1.4}$$

alternatively,

$$corr.(\tau) = \int_{-\infty}^{\infty} \tilde{x}(f)\tilde{y}^*(f)e^{-2\pi i f\tau}df, \tag{1.5}$$

where τ is the lag of the filter function behind the detector output. It is then possible to determine the SNR by calculating the mean value of the correlation (denoted $\overline{corr.}$), and the variance ($\overline{corr. - \overline{corr.}}$), given as

$$S \equiv \overline{corr.}(\tau) = \int_{-\infty}^{\infty} \tilde{h}(f)\tilde{y}^*(f)e^{-2\pi i f(\tau - t_a)}df, \tag{1.6}$$

and

$$N^2 \equiv \overline{corr. - \overline{corr.}} = \int_{-\infty}^{\infty} S_n(f)|\tilde{y}(f)|^2 df, \tag{1.7}$$

respectively. The mean value takes this appearance because the mean value of the noise is zero, so the mean value for the signal with noise should be the signal. The SNR ρ is defined as $\rho^2 = S^2/N^2$, which can be expressed in terms of the scalar product already defined:

$$\rho^2 = \frac{\langle he^{-2\pi i f(\tau - t_a)}, S_n(y)\rangle}{\sqrt{\langle S_n(y), S_n(y)\rangle}}, \tag{1.8}$$

and it is possible to calculate the optimal SNR, $\rho_{opt} = \langle h, h\rangle^{1/2}$.

1.2.4.2 Template bank generation

On the surface, application of matched filtering seems straightforward; a bank of templates could be generated covering the parameter space inhabited by the gravitational wave signals and determine the optimal filter, which will match the shape and

parameters of the signal. Although it is not possible to be certain that the signal is present in the data, it is possible to assign confidence levels to the results of a search, and generate estimates of the source parameters. In practice the parameter space to be searched is high-dimensional (see 3.2 for one such example), making grid-based template banks computationally intractable to create or to evaluate without being extremely coarse. Nevertheless, an effective search will require enough templates to ensure that adjacent templates are not too mismatched.

Given a template with signal parameters λ (normalised such that $\langle h, h\rangle = 1$) and following (21), one can define the match between it and another template with slightly offset parameters $\lambda + \triangle\lambda$ as

$$M(\lambda, \lambda + \triangle\lambda) = \langle h(\lambda), h(\lambda + \triangle\lambda)\rangle. \tag{1.9}$$

The match is at its maximum value when $\triangle\lambda = 0$, and the power expansion around this point produces

$$M(\lambda, \lambda + \triangle\lambda) \approx 1 + \frac{1}{2}(\frac{\partial^2 M}{\partial\triangle\lambda^i \partial\triangle\lambda^j})_{\triangle\lambda^k=0}\partial\triangle\lambda^i\partial\triangle\lambda^j, \tag{1.10}$$

defining a metric as

$$g_{ij} = -\frac{1}{2}(\frac{\partial^2 M}{\partial\triangle\lambda^i \partial\triangle\lambda^j})_{\triangle\lambda^k=0}. \tag{1.11}$$

This intuitively suggests the property of *mismatch* between the two templates, the square of the proper distance between them

$$1 - M = g_{ij}\partial\triangle\lambda^i\partial\triangle\lambda^j. \tag{1.12}$$

Traditional template banks construction addresses the notion of *completeness*: the requirement that any point in the parameter space is no further from its closest template than a chosen mismatch value, necessitating a compromise between the speed of a search and the number of templates required. An efficient, complete template bank will contain the smallest number of templates required to completely cover the parameter space.

However, setting the mismatch to be less than some desired value -m say- requires that the point in parameter space lies within the high-dimensional sphere of radius $\sqrt{m}$ centered on the template λ. Unfortunately, the inevitable overlap between

spheres at high dimensions raises the issue of the covering *thickness* Θ, where in a n-dimensional space

$$\Theta = \frac{\text{volume of an N} - \text{dimensional sphere}}{\text{volume of the fundamental region}}, \tag{1.13}$$

the *fundamental region* being the volume of the lattice defined by the centre of N-dimensional spheres for the parameter space (see (22)). This provides a convenient measure of the quality of the template bank being used: for a given mismatch we want the thickness as close to unity as possible, and with the minimum number of templates.

For two-dimensional Euclidean spaces, a hexagonal lattice is used, while at higher dimensions a hyper cubic or A_n^* lattice may be applied. More recently, 'random template banks' have been tested (23), where the criterion is the probability that a point in parameter space is covered by a template, rather than trying to achieve complete covering. In high-dimensional parameter spaces, this can reduce the number of required templates.

Nevertheless, the number of templates N_t can be given by

$$N_t = \frac{1}{m}\frac{\Gamma(n/2+1)}{\pi^{n/2}}\int_V \sqrt{\det(\mathrm{g_{ij}})}dV, \tag{1.14}$$

where V is the parameter space being searched. N_t will be large for small mismatches over large parameter spaces ((21) provides an estimate to template requirements in searching for inspiraling binaries as a measure of computational power). Furthermore, lattice template placement is usually done under the assumption that the parameter space metric is flat, an assumption that may not be true. Without an unexpected leap in computational capabilities, it seems unlikely that a straightforward search using matched filtering is a wise strategy for gravitational wave detection.

1.2.4.3 Bayesian inference

Ruling out an exhaustive search through the parameter space necessitates a more sophisticated approach to the problem. Bayesian inference is an intriguing framework in which to establish such an approach, providing a straightforward measure of the probability that a hypothesis is true. In order to demonstrate its applicability to gravitational wave data analysis, some simple definitions must be provided (for a more thorough grounding, the reader is referred to (24) and (25)).

Bayesian inference operates by considering the probability of the truth in propositions. Given two hypotheses A and B, it is possible to determine the joint probability of these being true, denoted $p(A, B)$. The joint probability is expressible using $p(A)$ and $p(B)$, the individual probability densities of the hypotheses, and the conditional probability densities $p(A|B)$ and $p(B|A)$ (that is, the probability of A being true *given* B and the probability of B being true *given* A). The product rule for joint probabilities $p(B)p(A|B) = p(A)p(B|A)$ leads to the typical form of *Bayes' theorem*

$$p(H_i|D, I) = \frac{p(H_i|I)p(D|H_i, I)}{p(D|I)} \tag{1.15}$$

where

H_i	proposition representing the hypothesis
D	proposition representing data
I	proposition representing prior information
$p(H_i\|D, I)$	posterior probability of the hypothesis
$p(H_i\|I)$	prior probability of the hypothesis
$p(D\|H_i, I)$	probability of obtaining the data if the hypothesis and the data are true (the likelihood)
$p(D\|I)$	normalisation factor to ensure $\sum_i p(H_i\|D, I) = 1$

(24)

Within the context of gravitational wave data analysis then, the proposition may be that 'the data contains a gravitational wave signal', and in fact it is a trivial task to construct a measurement known as the *odds ratio*; the ratio of the posterior probabilities of two hypotheses H_1 and H_2.

$$O_{12} = \frac{p(H_1|D, I)}{p(H_2|D, I)} = \frac{p(H_1|I)}{p(H_2|I)} \frac{p(D|H_1, I)}{p(D|H_2, I)}. \tag{1.16}$$

Bayesian methods thus allow competing theories to be compared, and can incorporate prior information. Nevertheless, although simple to define, actually calculating the posteriors may by computationally expensive. A typical application for gravitational wave data analysis is to recover astrophysical source parameters from a particular signal, with the desired output being the marginal posterior probability distribution for each parameter. It is possible to calculate these with a process

called *marginalisation*. For a set of n parameters $\{\lambda_\mathbf{i}|i = 1, 2, .., n\}$ and a desired parameter λ_k, it is possible to calculate the marginal posterior probability as

$$p(\lambda_k|D, I) = \int_{\lambda^*} p(\lambda|D, I)d\lambda^*, \tag{1.17}$$

where λ^* is the set of parameters excluding λ_k. What is clear though is that this method may require multi-dimensional integrals that may not be easy to evaluate. Moreover, the described approach has not simplified the problem if it represents only another way to present the results of an exhaustive search through the parameter space (by means of testing out multiple hypotheses).

Markov Chain Monte Carlo

Markov Chain Monte Carlo searches are an attempt to avoid the need for an exhaustive search through a large parameter space. Fundamentally, the MCMC approach is to construct a quasi-random walk through the parameter space where the posterior density in a region of the space and the probability of being in that region are proportional. The following explanation follows the terminology and approach of (24) closely, which provides a full explanation of the technique and a number of instructive examples.

The walk uses a Markov Chain, a record of the path through the parameter space where the next step depends on the current step. Although this may appear confusing at first, the principle is easily explained: given a proposal distribution of the parameters to be explored and a current position in the parameter space X_t, a candidate value -Y- for the random walk's next step -X_{t+1} can be chosen, and a decision is taken whether or not to accept the candidate based on the ratio

$$r = \frac{p(Y|D, I)}{p(X_t|D, I)} \frac{q(X_t|Y)}{q(Y|X_t)}, \tag{1.18}$$

known as the Metropolis ratio. The candidate is then accepted if the condition $r > 1$ is satisfied, or with a probability of r by generating a random value from a probability distribution between 0 and 1. The process can then simply be repeated with the path at the new position. In general the walk will move towards regions of increasing probability but is free to move away from these and explore regions of low probability (although it will soon move away from those). The benefit is that the chain will sample the parameter space with a probability density equal to the posterior probability $p(X|D, I)$.

Naturally, there are refinements of this technique designed to optimise performance, such as tempering, variable step size, and changing the burn-in period (a number of initial steps that are subsequently ignored so that the posteriors are not unduly affected by the choice of the starting point in parameter space (24), (25)), but the simple situation is still illuminating. Given time, a properly constructed Markov chain can explore the entire parameter space, but its concentration in areas of high probability will identify possible candidate parameters that can then be subjected to more exhaustive searches.

The data analysis concepts in this section are no more than a brief description, covering only a small subset of the those in use by gravitational wave data analysts throughout the international gravitational wave community. They have been included to illustrate the elaborate nature of the techniques considered, just as the range of sources described in Section 1.1.1 is an indication of the scale of the field that analysis techniques must be applied to. Accordingly, it is extremely unlikely that a single approach will produce all of the meaningful output that we desire (nor would it be expected to any more than the notion of one detection process that would cover all electromagnetic emissions from astrophysical sources).

The remainder of this thesis is an examination of alternative approaches to gravitational wave data analysis, with an initial focus on performing an analysis with a reduced amount of information. The intent is not to ignore the expertise of the wider analysis community or their methods, but to investigate whether we can produce useful information about the presence or makeup of different sources quickly or simply.

Chapter 2

A rapid search for Chirp Mass and Coalescence Time in SMBH binaries

2.1 Supermassive Black Holes

Black holes with masses of the order $10^6 M_\odot - 10^9 M_\odot$ are commonly known as Supermassive Black Holes (SMBH) and are believed to exist in the centre of most galaxies -an inference from the motion of stars or gas clouds near the centre- including the Milky Way (26), (27). The merger of two or more of these objects due to the collision of galaxies or even as isolated SMBH binaries (that is, binary systems devoid of other inspiralling material) will produce gravitational waves within LISA's frequency band. These coalescences will provide valuable information about the behaviour of spacetime in strong gravitational fields, and the high amplitude of the signals will allow the events to be detected at exceptionally large distances. Such events are therefore particularly interesting not only for the field of gravitational wave astronomy, but those of cosmography and cosmology as well.

2.1.1 The Mock LISA Data Challenges

The Mock LISA Data Challenges (MLDC) are a series of datasets generated by the MLDC Taskforce (W..l) containing simulated gravitational wave data and simulated LISA noise, released in a number of 'rounds'. With a standard encoding, and available to any interested party, the two goals of the MLDC were (W..o) "fostering

the development of LISA data analysis tools and capabilities, and of demonstrating the technical readiness already achieved by the gravitational-wave community in distilling a rich science payoff from the LISA data output"(W..o). Each challenge generally consists of training datasets accompanied by a key containing the source parameters of the gravitational wave signals, in addition to blind datasets (containing signals from similar sources to the training datasets, but which do not provide a key with the specific source parameters); this allows participants to gauge the effectiveness of their efforts to extract useful science from the training data before repeating the process without the benefit of a known solution.

Each round of the MLDC has a limited period in which groups or individuals can submit their results after which key files containing source parameters for the blind challenges are released, again allowing the performance of each analysis to be evaluated, both individually and in comparison to other participants. Although not intended specifically as contests this format encourages discourse on the correct approach to the challenges, the criteria by which success is judged and robust analysis methods. Subsequent challenges examine specific situations or increase the complexity of existing situations, with the goal of providing increasingly realistic circumstances with which to test analysis codes.

The following sections provide a brief description of the MLDC so far, and the intended focus of data analysis attempts. The specifications of individual datasets are provided by (W..e) and (W..o), while reports on the first two rounds are given by (28) and (29) respectively. A more detailed overview of the rounds 1B to 3 is provided by (30) and (31), as well as a discussion of the analysis methods used by various participating groups.

2.1.2 MLDC datasets

The first round of MLDC datasets was designed to allow the development of basic tools for the analysis of datasets containing primarily of single, or non-overlapping multiple signals, grouped in three general categories: i) white dwarf binaries ii) extreme mass ratio inspirals iii) supermassive black holes. Additionally, a dataset featuring confusion noise (wherein the overlap of multiple signals makes it difficult to extract information about the individual sources) was included. The initial challenge was active between June and December 2006 with preliminary analysis presented at the 11th Gravitational Wave Data Analysis Workshop ((W..a),(32)),

but a supplementary challenge (designated Round 1B) was issued using updated data generation codes and was concluded December 2007.

Running concurrently with some datasets from the initial challenge (specifically the extreme mass ratio inspiral datasets, in order to account for the expected difficulty of analysis on these waveforms), round two began in December 2006 and included datasets containing multiple overlapping sources and 'The Whole Enchilada'; a two-year long dataset containing gravitational wave signals from a variety of sources previously presented separately, giving a more realistic simulation of LISAs output and providing a test for analysis methods without the benefit of a single category of source types or isolated sources. The analysis result submission for the second round was June 2007.

The third MLDC round built on the previous round, including two new sources: burst signals from cosmic string cusps and a stochastic gravitational wave background. Further the challenges incorporate improvements to some of the source models used, such as spinning SMBHs and the corresponding spin-spin and spin-orbit interactions. This round ran between April and November 2008. Recently, a fourth round of MLDC Challenges (W..n) has been announced, consisting of another 'whole enchilada' dataset in order to focus the analysis on the "global-fit problem of detecting and analyzing sources of different types superposed in the LISA data". The deadline for entry to this round is intended to be late Autumn 2010.

2.1.3 SMBH signals in MLDC datasets

Correctly determining orbital evolution of the SMBH system during the merger is particularly involved, requiring consideration of the evolution of the galaxies housing each SMBH, and the interaction of the black holes with infalling gas and material, but unless explicitly stated the SMBH mergers considered throughout this chapter involve two SMBHs isolated from any disruptive effects from other objects. Even for isolated SMBH binary systems, what we require is a solution to Einstein's equations (describing spacetime being curved by the presence of the two massive objects), but as a full solution is not yet possible Post-Newtonian approximations are used instead, expressing the solution as deviations from Newton's theory of gravity.

Although higher-order Post-Newtonian (PN) effects could be calculated, the orbital evolution of SMBH binaries in the initial MLDC datasets was restricted to

the 2PN approximation without spin-spin or spin-orbit modulations, with the additional assumption that the orbits were circular. Furthermore, only the dominant mass-quadrupole harmonic is present, and the amplitude of this is not subject to PN correction; this simplification may be referred to as the 'restricted' PN approximation. The first two rounds of the MLDC therefore presented particularly simple waveforms, allowing each binary to be described by seven parameters plus sky location(see Table 2.1).

The inspiral waveform of these binaries should be visible to LISA for a significant amount of time, and the observation of the frequency of the gravitational waves at a time $t(f)$ can be related to the coalescence time of the binary t_c by the relation

$$\begin{aligned} t_c - t(f) &= 5(8\pi f)^{-8/3} \left(\mathcal{M}(1+z)\right)^{-5/3} \\ &\quad \left[1 + \frac{4}{3}\left(\frac{743}{336} + \frac{11\mu}{4M}\right) x - \frac{32\pi}{5} x^{3/2} + O(x^2)\right] \end{aligned} \tag{2.1}$$

where the *chirp mass* of the binary system, $\mathcal{M} = \eta^{3/5}\mathbf{M}$, with $\eta = \frac{M_1 M_2}{\mathbf{M}^2}$, $\mathbf{M} = (M_1 + M_2)$, the PN expansion parameter $x(f)$ is given by $x(f) = (\pi \mathbf{M}(1+z)f)^{2/3}$ and z is the binary's redshift (see (33), (34)).

To lowest order, then

$$\begin{aligned} t_c - t(f) &= 5(8\pi f)^{-8/3} \left(\mathcal{M}(1+z)\right)^{-5/3} \\ &= 3.003 \times 10^6 \left(\frac{f}{10^{-4}}\right)^{-8/3} \left(\frac{\mathcal{M}(1+z)}{10^6 M_\odot}\right)^{-5/3} \end{aligned} \tag{2.2}$$

and a simple rearrangement of this expression yields

$$f = \left(\frac{t_c - t(f)}{3.003 \times 10^6 \left(\frac{\mathcal{M}(1+z)}{10^6 M_\odot}\right)^{-5/3}}\right)^{-3/8} \left(10^{-4}\right) \text{Hz}. \tag{2.3}$$

Note: In the remainder of this chapter, the reshift of a source is ignored when determining chirp mass, since it is unknown.

Because the final moments of the merger and the ringdown are not part of this frequency evolution, a taper is applied to the waveform to prevent spectral leakage

M_1	Mass of first black hole.
M_2	Mass of second black hole.
t_c	Time of coalescence.
Φ_0	Initial orbital phase.
θ	Inclination of orbital angular momentum to observer's direction.
D	Luminosity distance.
ψ	Polarisation angle.

Table 2.1: The physical parameters for a SMBH binary.

(due to the large dynamical range resulting from suddenly terminating the waveform suddenly in the time domain).

The third round of the challenges included spin effects and the resulting modulations due to the now precessing binary orbits, marking a departure from the simple waveforms present in the first two rounds, and increasing the parameters required to describe the binaries; both the magnitude and orientations of the spins are required. While rigorous and accurate parameter estimation of SMBH binary signals have been implemented with some success, we have tried to explore a rapid search for the signals and the information to be gained from such a method.

2.2 The rapid search

The following section explores our somewhat naive search for SMBH binary signals in MLDC datasets, and an attempt to extract the chirp mass and coalescence time of these signals, referred to as the 'excess power' method. We believed there was merit in sacrificing the accuracy of results possible from techniques such as MCMC searches or matched filtering, for speed. Our motivation was that our estimates could then be passed on as initial data for an MCMC search, providing prior information that would allow us to home in on accurate parameter values under this more thorough analysis.

We considered a hierarchical approach to the gravitational wave data analysis. Our contribution would be an initial stage in such an approach; to quickly, if inaccurately, provide estimates of parameter values. This way, we could incorporate of research into the efforts of the wider data analysis community, but hopefully without recreating or overlapping those efforts. Again, this was not intended to be a full

solution to the problem but at this stage was rather an initial investigation into the difficulties of gravitational wave data analysis and an exercise in manipulating and gaining familiarity with the MLDC datasets.

Unless otherwise stated, the MLDC Challenge 1.2.1 training dataset will be used throughout the following section to illustrate this method and its application to further datasets is given in section 2.2.2.

Generating spectrograms

The following method is used to generate the SMBH spectrograms discussed in later sections, using the notation of (35).

The gravitational waveform is represented by a vector of M real numbers x_k, $k = 1...M$ representing the strain amplitude channel ($X(t)$ or $Y(t)$ or $Z(t)$) with a sample frequency of f_s. The power spectral density of this waveform is then

$$PSD = \frac{2 \cdot |y|^2}{f_s \cdot S_2} \tag{2.4}$$

where y_m is the short-time Fourier transform (of length N) of x_k, w_j represents the vector of real numbers of length N that makes up the window function, and

$$S_2 = \sum_{j=1}^{N} w_j^2. \tag{2.5}$$

Throughout, a sample frequency of $f_s = 1/15$ Hz is used, and the windowing function is a Hanning window, defined as

$$w_j = \frac{1}{2}\left[1 - \cos\left(\frac{2\pi \cdot (j-1)}{N}\right)\right]. \tag{2.6}$$

Windowing was necessary to remove discontinuities in the timeseries being converted into a spectrogram. Discrete Fourier Transforms assume that the timeseries repeats itself infinitely in a periodic manner and if the frequency of the input signal is not an exact multiple of the sample frequency (and therefore does not fall in the exact centre of a frequency bin), then there is a discontinuity between the first and last samples, spreading power across the spectrum. Window functions generally start and end at or close to zero, removing this discontinuity. There are a number of defined and named window functions, although it is perfectly possibly to design custom window functions. Each will involve a compromise between the reduction

of the spectral leakage into other frequency bins, the amplitude accuracy and the width of the peak in the frequency domain when applied to a signal.

The Hanning window was an appropriate choice for our work in this thesis because it causes small amounts of spectral leakage, and narrow peak widths, at the cost of some amplitude accuracy. However, our analysis methods (further explored in this chapter and in Chapter 4 onwards) relied more on recovering information about the shape of spectral features than recovering the amplitudes of signals, making this an attractive choice. Further, the Hanning window was a built-in function of the Matlab package used in our analysis, making it simple to incorporate into our algorithms.

Alternative window functions (such as flat-top windows which could recover the amplitude accurately) would have contributed more spectral leakage or wider peak width, distorting the shape of the signal in the spectrograms. Heinzel (35) assesses a number of window functions, giving their strengths and weaknesses.. Unfortunately, there was not sufficient time to test our analysis techniques using spectrograms created with different window functions. It is possible that there exists an optimal window which has eluded us, but the Hanning window appeared to be sufficient for our purposes.

2.2.1 The 'excess power' method

The frequency evolution of SMBH inspirals in MLDC dataset signals produced by the relation in Eqn. 2.3 produces a sort of 'characteristic' shape, changing only slightly when the chirp mass of the binary is altered. Naturally, this will also change depending on the coalescence times as well, but if these are equal the shape is apparent (see Fig. 2.1).

In some cases, the presence of such signals could be identified by eye within a spectrogram of the dataset (see Fig. 2.2.1) but, the signal is usually drowned out by noise above and below a small range of frequencies. Within this range however, the signal stands out. Crucially, the power in the waveform increases as the coalescence time approaches; by isolating the time of maximum power we will be able to identify the coalescence time.

The initial step is to select a range of frequencies that isolate a region of high SNR in a spectrogram of the dataset; in practice this involves ignoring extremely

low-frequency information and anything higher than a few mHz, and the refinement of this method identified a desirable range of $0.1 - 1$ mHz (shown in Fig. 2.3(a)).

The total power in each time bin across this range of frequencies is calculated (see Fig 2.3(b))- the increase in SNR as the inspiral progresses towards the coalescence time reveals itself as a corresponding increase in these measurements; there will be a contribution of power from the noise too, but the signal should produce most of the power. It is this 'excess power' (that is, above the noise contribution) that we wish to guide us.

Crucially, because the MLDC dataset does not include the final merger and ringdown the coalescence time of the signal is then followed by a sharp drop in the sum power measurement, a feature which we can exploit. The rise and fall in sum power measurements can be turned into distinct spikes if a sensible threshold is applied, and the peaks of these spikes are then used as the first estimates for the coalescence time of the SMBH signals. The peaks are found by identifying those time bins for which the sum power measurement lies above the threshold but precedes a time bin that falls below the threshold (thus, the sharp drop really is key to the first estimates). Each estimate is considered evidence of a separate *candidate* SMBH signal.

This stage is simple enough in practice, but we then examine subsets of the original dataset surrounding each of the candidates, slicing these subsets into several sections and constructing PSDs from each slice, determining the most powerful frequency bin in each slice. In theory, this highlights the frequency evolution of the SMBH signal leading up to the coalescence (see Fig 2.5(b)). We refer to these as *extracted* evolutions.

Specifically, we looked at the timeseries data from two days preceding the initial candidate coalescences up until those values, creating five slices. We discovered that our results were more accurate if the slices were of unequal length; each was half the length of the preceding slice (that is, the first of the five slices contained one days' worth of timeseries data, the second half a day, and so on). Furthermore, we actually took the average value of the three most powerful frequency bins rather than the most powerful outright. Thus, our extracted frequencies were not regularly spaced, but closer together as we approached the candidate coalescence time (during which the frequency of the gravitational waves was increasing ever more rapidly).

It is then possible to extract a chirp mass and coalescence time using a simple least-squares fit. A bank of templates for different chirp masses and coalescence

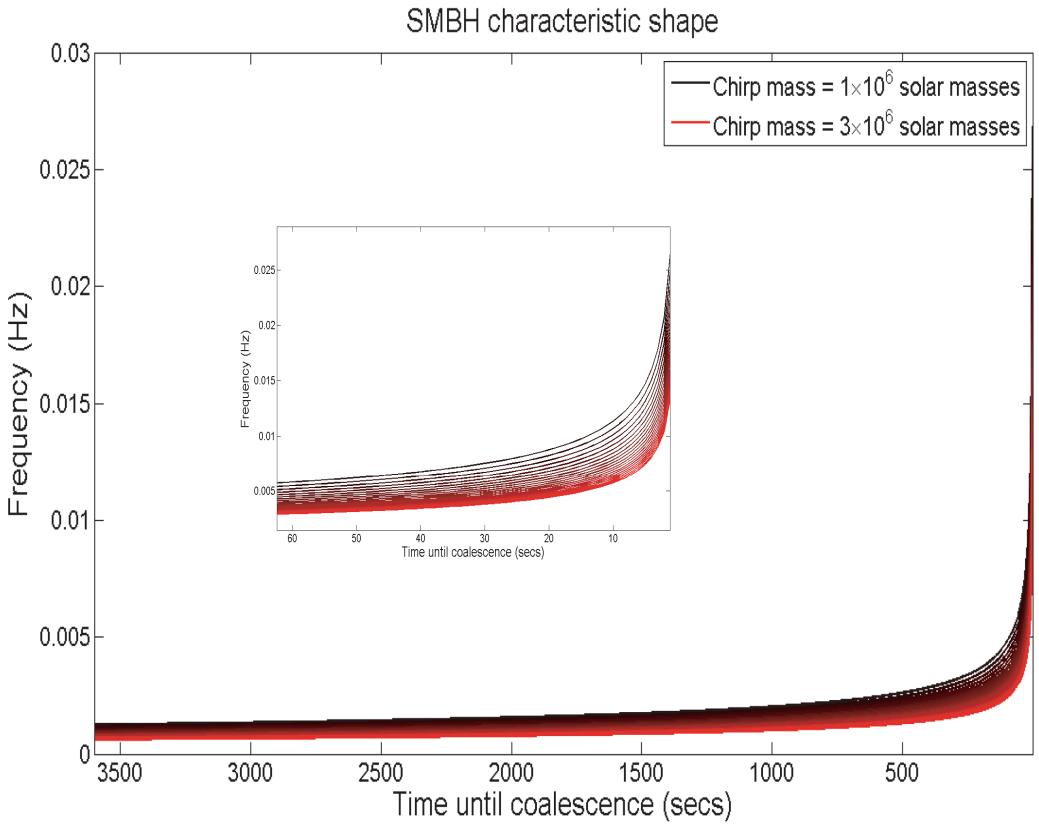
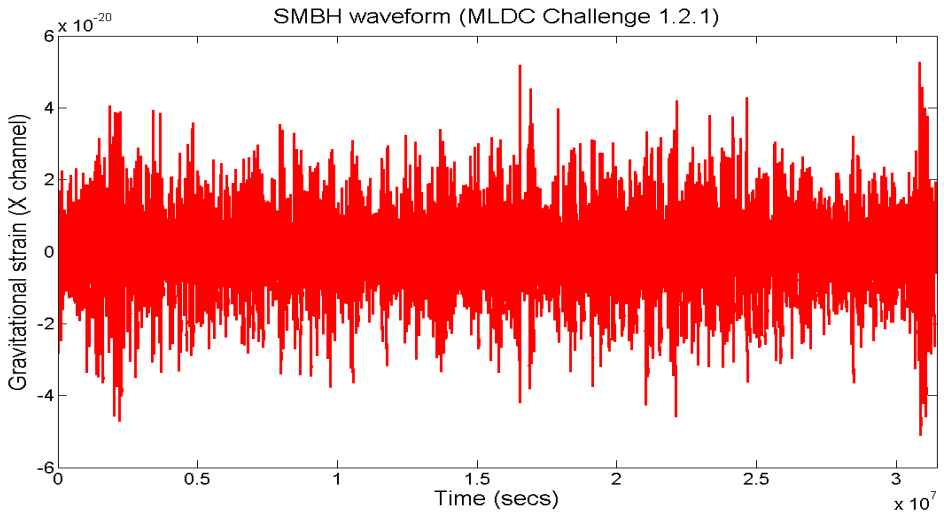
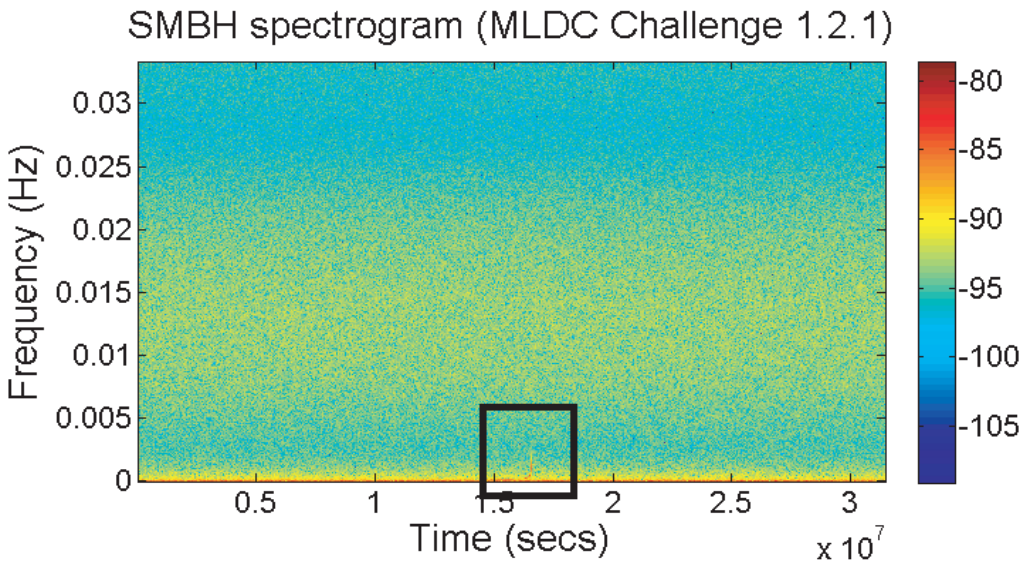


Figure 2.1: The 'characteristic shape' of SMBH inspiral signal frequencies in the final hour before coalescence. The chirp mass for the binaries is in the range $1 \times 10^6 - 3 \times 10^6\ M_\odot$ with a resolution of $1 \times 10^5\ M_\odot$, and as the chirp mass increases (depicted with a change from black to red), so does the sharpness of the curve. An insert showing the last minute before coalescence is also included.

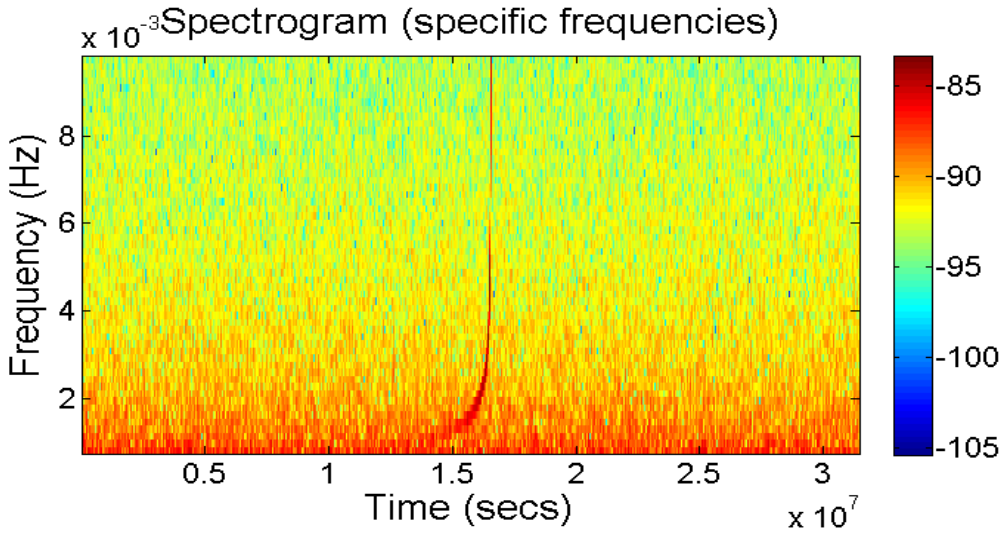


(a) SMBH waveform from the MLDC Challenge 1.2.1 dataset X channel

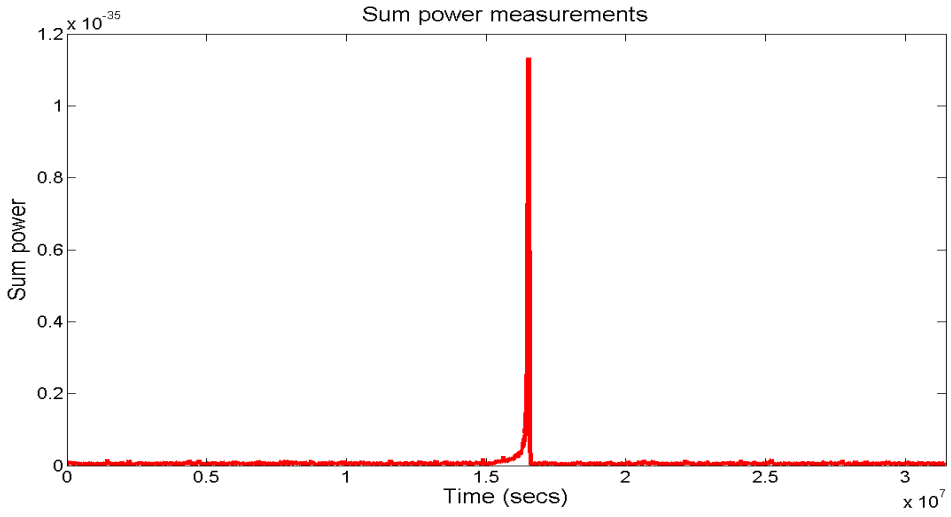


(b) The spectrogram constructed from the waveform above.

Figure 2.2: A SMBH inspiral waveform with noise, and the corresponding spectrogram. The signal cannot be seen in the waveform timeseries, but evidence of the characteristic shape is visible in the spectrogram at low frequencies with considerable difficulty (the relevant region is highlighted as a guide). The natural logarithm of the power (indicated by the colour bar) is shown to make the feature easier to see.



(a) The selected frequency region of the spectrogram in Fig. 2.2(b). The presence of the SMBH signal is much easier to see here.



(b) The sum power measurements from the spectrogram above.

Figure 2.3: The isolated frequency region of the spectrogram and the sum power measurements. The approximate coalescence time corresponds to the peak of these measurements. The logarithm of the power is displayed.

times (we refer to these as *trial* coalescence times and chirp masses)can be generated covering a desired parameter range with a particular coarseness, dependent to a large extent on available computing resources. For a given trial chirp mass and coalescence time, the *expected* evolution according to Eqn.2.3 (that is, the expected frequency at the same times as the extracted frequencies are calculated) can be compared to the extracted evolution and the best fit selected.

In other words the fitting procedure algorithm was simply

1. Extract the strongest frequencies from PSDs around the time of the candidate coalescences, and the corresponding time. These are the *extracted* times and frequencies.

2. Select a trial chirp mass and coalescence time. Determine the frequency of the gravitational wave at the extracted times according to the trial values. These are the trial frequencies.

3. Perform a least-squares fit on the five extracted and trial frequencies, and record this value as well as the trial values.

4. Repeat stages 2 and 3 for different trial values.

5. Select the smallest least-squares value; the corresponding trial values are the best fit parameters.

Naturally, there are a number of subtleties involved in implementing this approach. As the expected quadrupole frequency heads towards infinity as the coalescence time approaches, it is important not to attempt to fit templates to an extracted frequency evolution including data from after the sharp drop off in total power- this would establish a different shape to which the templates would not match well. Therefore by examining the frequency evolution preceding the original estimated coalescence time the method risks ignoring or missing the point of highest SNR in favour of robustness in tackling different datasets. Conceivably, contribution from the noise could produce a peak later than the coalescence time, although the high SNR of the SMBH signal has thus far prevented this, allowing the initial estimates to be considered an upper bound for the data considered. It is worth emphasising that the search for the best fit parameters can cover times beyond these

estimates; it is simply that the fit is made only to extracted frequencies from before this point.

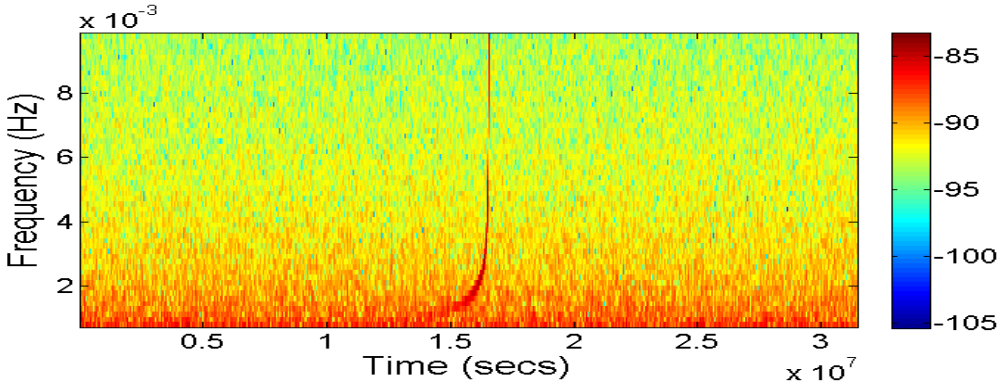
Furthermore initial implementation of this method was simplified by the availability of noise-only datasets, allowing noisy datasets to be whitened by accounting for the power present in the expected LISA noise. The mean power of the noise at different frequencies, $\langle P_{\text{noise}}(\nu)\rangle$is calculated, as well as the power in the noisy time series at different frequencies and times $P_{\text{n}}(\nu, t)$. The whitened power $P_{\text{w}}(\nu, t)$ is simply the ratio of the two measurements.

$$P_{\text{w}}(\nu, t) = \frac{P_{\text{n}}(\nu, t)}{\langle P_{\text{noise}}(\nu)\rangle}$$

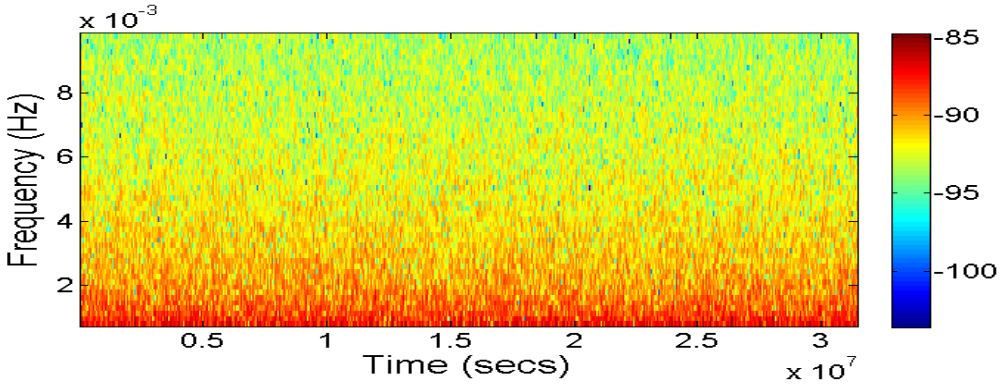
This is a simple approach but highlights the signal (as seen in Fig. 2.4(c)). Nevertheless, its success depends to some extent on the gaussian, stationary nature of the noise used during the challenge, which will not be an appropriate approach for a more realistic representation of the data. One immediate change is that the sum power measurements are altered by the whitening, and are henceforth referred to as normalised sum power measurements.

The extent of the data used to extract the frequency evolution is another subtlety; ideally the search should look as far back as possible without attempting to search before the beginning of the dataset, or so far back that it is confused by other sources. It must also refrain from pursuing the signal frequency as it drops down into the strong extremely-low-frequency noise. A more sophisticated approach may contain conditional statements allowing the search to avoid these problems elegantly, but in its current form it consistently looks at the data only in the two days preceding the initial coalescence time estimates.

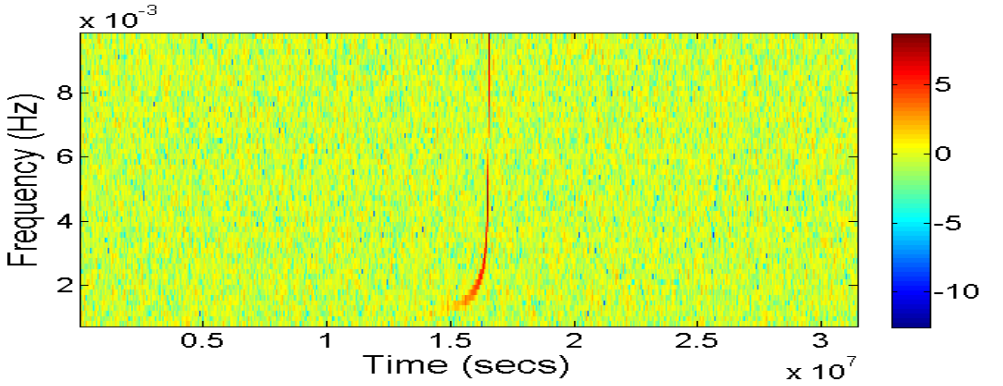
Some thought must also be given to the possibility of false alarms. These would manifest as a spike corresponding simply to a small fluctuation in the sum power close to the threshold set to isolate the SMBH signals. The search would assume this to be a coalescence candidate and attempt to extract the frequency evolution from spectrogram data preceding it; in some sense this is mostly harmless but the resulting information would be useless. In order to reduce the possibility of this occurring, the search has a 'resolution' of one week, only allowing one spike to become an estimated coalescence time (corresponding to the time bin associated with the maximum sum power within this period; by this process, candidates close enough in time are merged). Although this has solved the problem with regards to



(a) The original spectrogram



(b) The spectrogram of the noise



(c) The whitened spectrogram

Figure 2.4: The original spectrogram (within the selected frequency range), the noise spectrogram, and the resulting whitened spectrogram. In all cases, the natrual logarithm of the lower is displayed (indicated by the colour bar).

current MLDC datasets, it nevertheless has the disadvantage of limiting the ability of the rapid search to distinguish between multiple inspiral signals coalescing in a shorter period of time. However, the demands of the search for frequency bins of *maximum* power during the frequency evolution extraction would not be able to cope with this situation anyway.

Finally, ignorant of a preferred distribution of chirp mass values and free to choose the coalescence time arbitrarily, it is simplest to conduct the 2-dimensional search in a grid (the coarseness of which is dependent on available computing resources). It would be fairly trivial though to extract to incorporate distribution information into the search.

2.2.2 The results of the rapid search

The excess power method was applied to a number of MLDC datasets, and the resulting extracted parameters are discussed below. A number of points of note raised by the different datasets are examined in further detail in section 2.3. For the purpose of running multiple iterations of the search, it was necessary to recalculate the noise present in the dataset- details of the method used are given in appendix A. In total, the tests were repeated 100 times for each dataset, and individual spectrograms and sum power measurements throughout the section are taken from one iteration of the rapid search.

Furthermore, each rapid search was conducted on a 2-dimensional grid with a chirp mass resolution of $2000 M_{\odot}$ and a time resolution of 30 seconds; this was adopted for practical reasons as we wished to justify the rapid nature of the search and to be able to conduct the search on a standard desktop computer. The trial coalescence times covered a period five days preceding and following the candidate coalescence times, while the chirp mass range was changed depending on the dataset in use (this is discussed in Section 2.3). Finally, a uniform threshold for the normalised sum power measurements of 1×10^4 was used throughout; this value was determined empirically by examining the MLDC datasets and normalised sum power measurements in the provided challenge datasets.

2.2.2.1 Challenge 1.2.1

Challenge 1.2.1 contained a single SMBH binary inspiral with a high SNR ($450 < \mathrm{SNR} < 500$), and was easily identified by the excess power method search (see Fig.

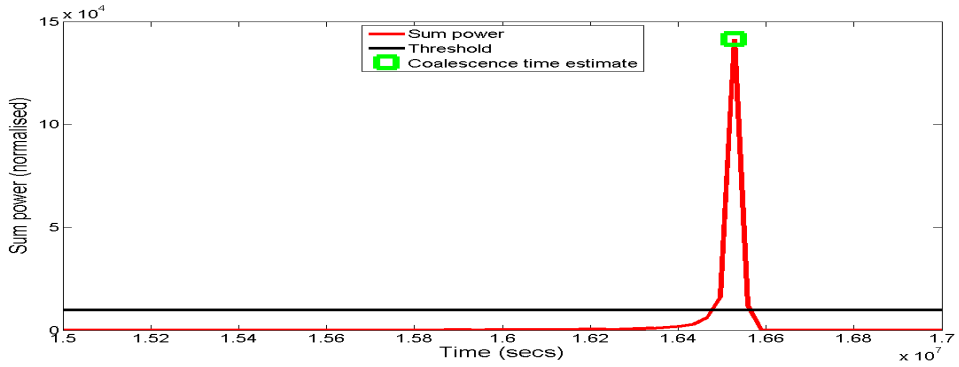
2.5(a)). A best-fit trial coalescence time and chirp mass were identified, and Fig. 2.5(b) shows the original spectrogram of the MLDC dataset around the time of the coalescence, with the extracted frequencies used overlaid on this (and a line showing the expected evolution). Also included is the 'true' evolution; the expected frequencies given the true coalescence time and chirp mass. However, even a close look at the time surrounding the candidate signal coalescence, the two lines are extremely difficult to tell apart.

A histogram of the 100 iterations of the search provides further information (see Fig. 2.6(a) and Fig. 2.6(b)). On a scale that shows the full extent of the dataset (blue lines denote the range of the chirp mass searched, and the extent of the dataset time, although the actual search for the best-fit time only covers the 10-day period stated earlier), the best-fit coalescence time appears to be same for all iterations, while there is a spread in the chirp mass. However, a closer look at the results, in Fig. 2.7(a) and Fig. 2.7(b), reveals that there is a spread in both parameters, with no strong contender for the most likely value for either.

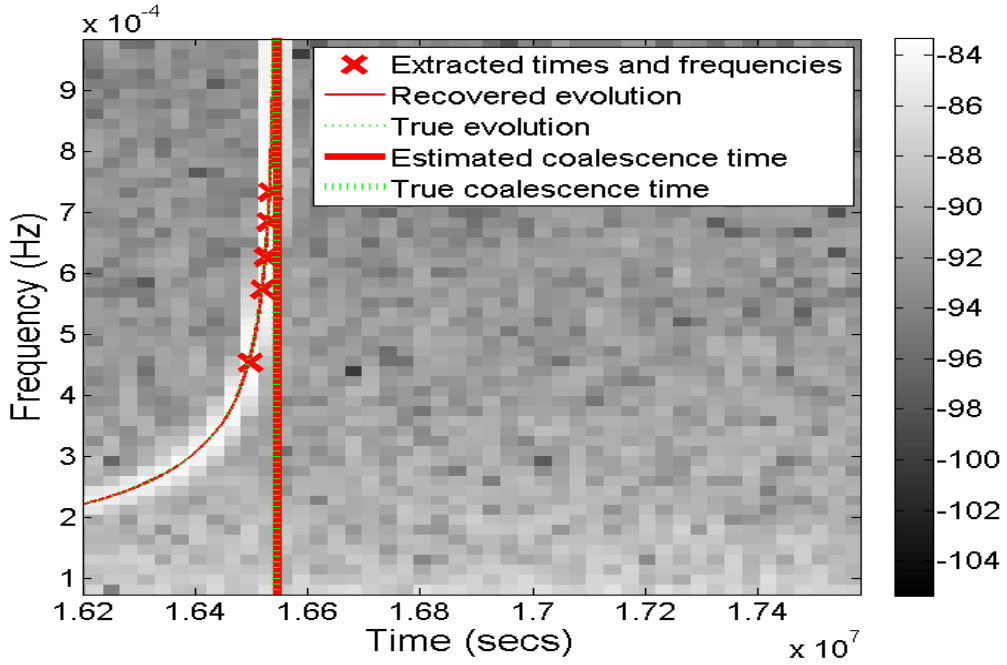
2.2.2.2 Challenge 1.2.2

Challenge 1.2.2 contained a single SMBH binary inspiral as well, albeit with a greatly reduced SNR ($20 < \text{SNR} < 100$). In addition, the SMBH binary did not coalesce during within the time of the dataset. Therefore, there was no sharp drop off in the sum power, and as a result the excess power method failed to identify a candidate time. Given this, the attempt to find best-fit parameters was stymied, so no result were recorded. Fig. 2.8(a) shows the original waveform, along with the whitened spectrogram (Fig. 2.8(b)), and the normalised sum power measurements for the dataset (Fig. 2.8(c)). Although it is possible to see an increase in the sum power measurements towards the end of the dataset, there is certainly no visible characteristic shape

In this dataset, no normalised sum power measurement is actually higher than the threshold, and attempts to lower this threshold began to produce numerous false alarms quickly. This did show that the algorithm was able to deal with large numbers of candidate signals, but revealed no further useful information (the extracted evolutions were very poor fits to any expected evolution).

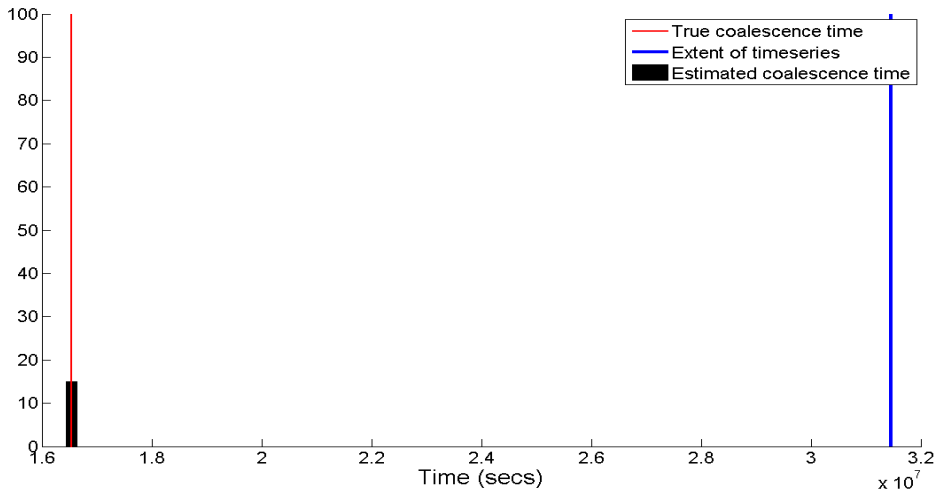


(a) Normalised sum power measurements. The initial estimate of the coalescence time, marking the peak of these measurements, is shown as well (green square).

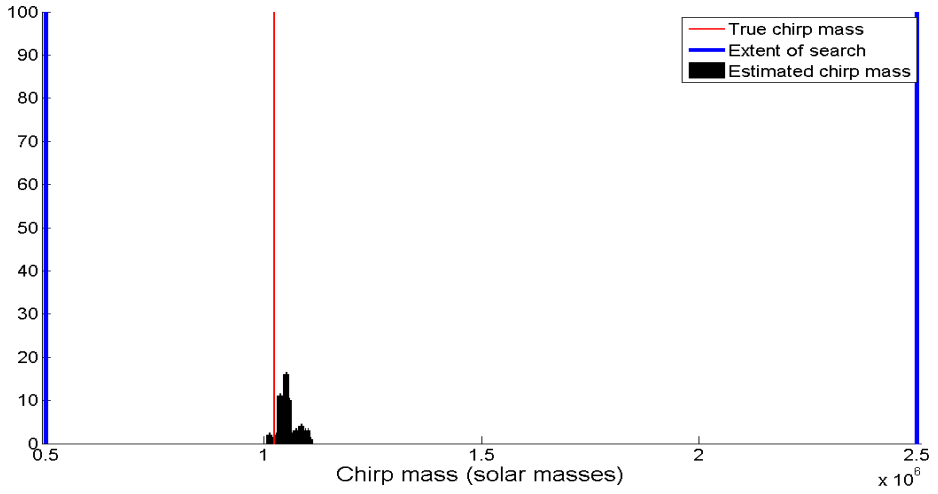


(b) Whitened spectrogram and evolution overlays. The natural logarithm of the power (indicated by the greyscale bar) is used to make the structure of the features easier to see.

Figure 2.5: The normalised sum power measurements of the MLDC Challenge 1.2.1 dataset spectrogram, and a close up of the spectrogram around the time of the candidate coalescence time. The extracted times and frequencies are shown, along with the expected evolution of the signal if the best-fit trial values were used in Eqn.2.3 (labelled here as 'Recovered evolution'; red line), and the expected evolution according to Eqn.2.3 using the correct chirp mass and coalescence time (green dashed line). At this resolution, they are almost completely indistinguishable.



(a) Best-fit trial coalescence times



(b) Best-fit trial chirp mass

Figure 2.6: Histogram of the MLDC Challenge 1.2.1 best-fit parameters according to the rapid search. There is an visible spread in the chirp mass, but the coalescence times selected are not as uniform as they appear.

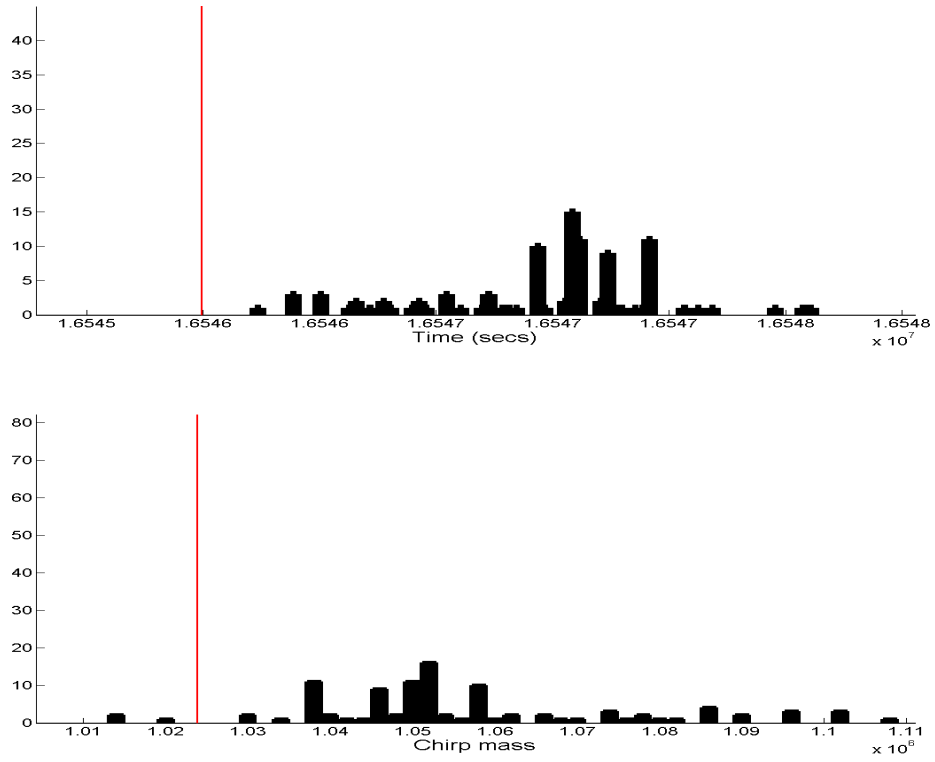
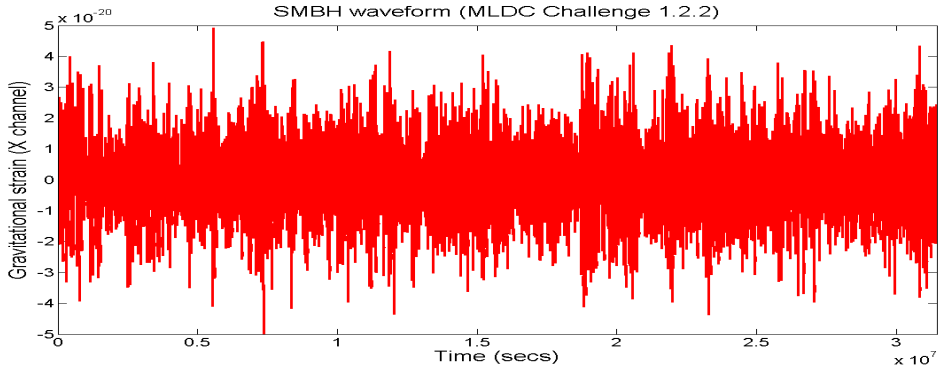
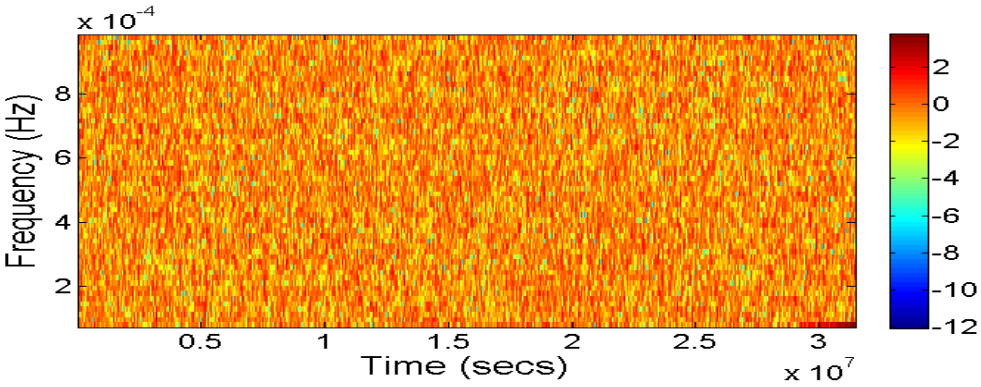


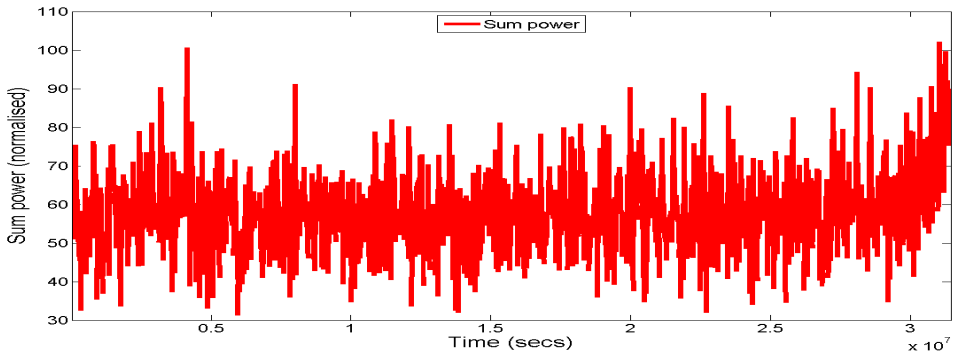
Figure 2.7: A closer look at the results from Fig. 2.6(a) and Fig. 2.6(b). The spread in the best-fit values of the parameters is evident, and there does not appear to be a clear choice for the most likely value of either.



(a) The original waveform



(b) The whitened spectrogram. The natural logarithm of the power (indicated by the colour bar) is used to make the structure of the features easier to see.



(c) Sum power measurements

Figure 2.8: The MLDC Challenge 1.2.2 dataset waveform, whitened spectrogram and normalised sum power measurements. The SMBH signal is not distinguishable in either the timeseries or the spectrogram, and there is no peak sum power measurement to act as a candidate coalescence time.

2.2.2.3 Challenge 2.2: 'The Whole Enchilada'

Known as 'The Whole Enchilda', Challenge 2.2 contained a wealth of information. As well as five of SMBH inspiral signals, there were gravitational wave contributions from Extreme Mass Ratio Inspirals, and 30×10^6 galactic white-dwarf binaries. This did not change our approach though; the search proceeded as normal.

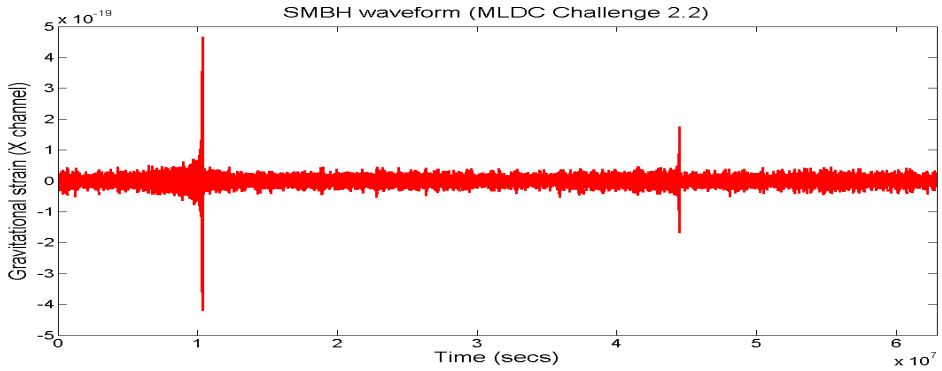
The rapid search identified all three SMBH signals that coalesced within the time period covered by the dataset (there were two signals present that coalesced some time later; their coalescence time and chirp masses have been included in the histograms of the recovered best-fit values). These candidates were sufficiently far apart that they were not merged, and their distinct evolution could be extracted. Fig. 2.9(a) shows that at least two of the signals are visible in the original data, while the whitened spectrogram in Fig. 2.9(b) reveals three (one very faintly). The wide range in sum power measurements from the different signals requires that they are displayed on a logarithmic scale (see Fig. 2.9(c)), but a single threshold still isolates all three spikes.

Histograms of the best-fit parameter values shows that the best-fitting chirp masses were not particularly accurate, although the coalescence times are fairly accurate estimates of the true values (as indicated by the red line in Fig. 2.10(a) and the corresponding close-up plots). The spreads in the values is also different for each signal.

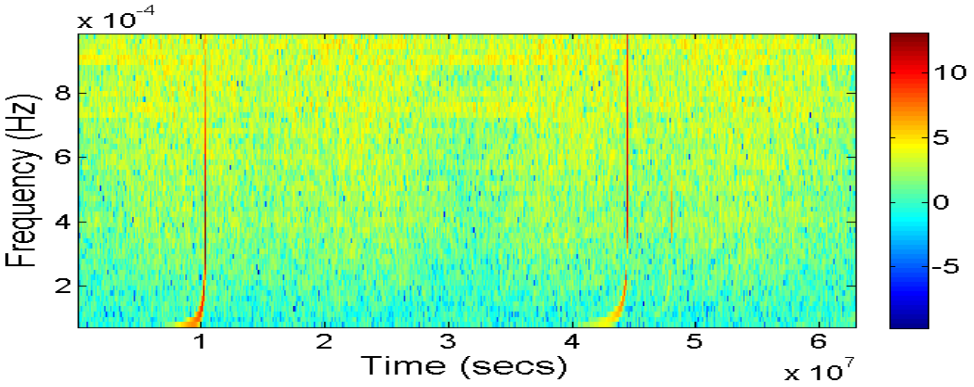
Fig. 2.12 shows an example of the recovered results overlaid on the original dataset spectrogram as well as the true evolution. Again, it is extremely difficult to tell them apart unless we take a much closer look at them individually (Fig. 2.13(a) – Fig. 2.13(c)).

2.3 Remarks about the rapid searches

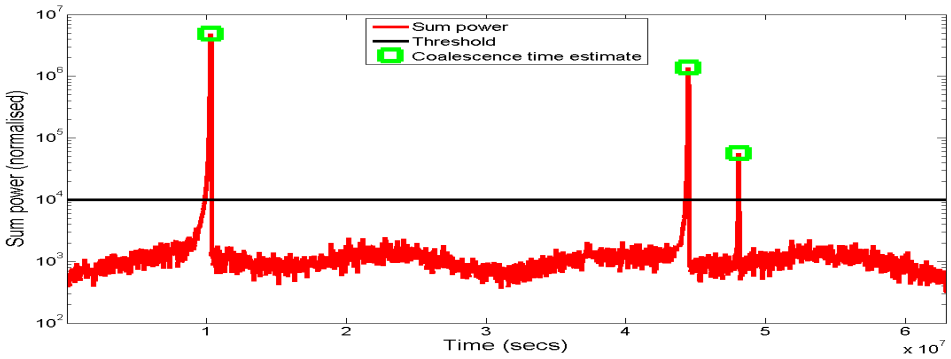
The results of the rapid search for SMBH do not seem particularly promising, and the approach is not sufficiently robust for us to be truly confident about. Nevertheless, it raises a number of points that can prove instructive when dealing with similar challenges and highlight the difficulties in applying the rapid search method to more complicated problems. The following section addresses these issues.



(a) The original data

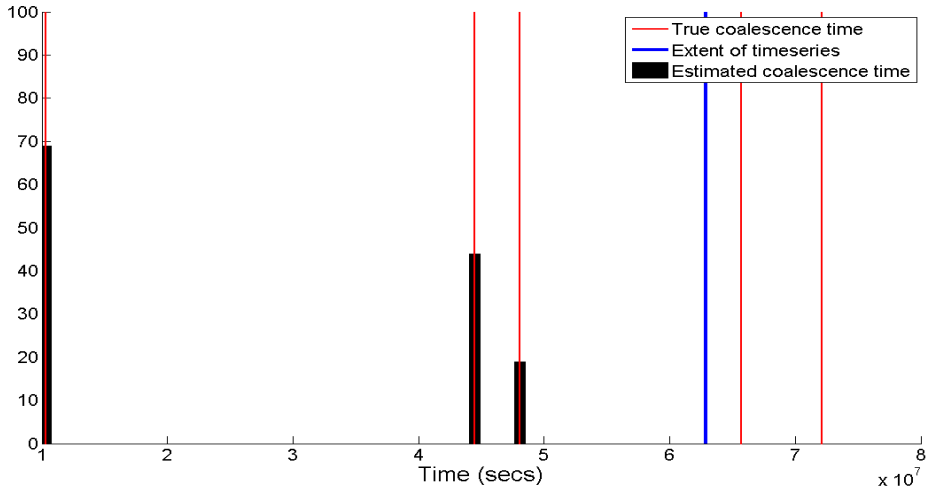


(b) The whitened spectrogram. The naatural logarithm of the power (indicated by the colour bar) is used to make the structure of the features easier to see.

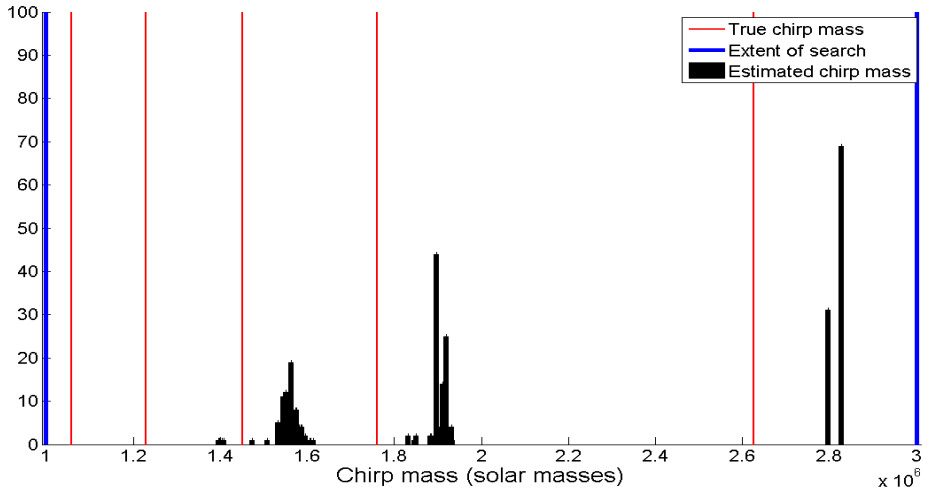


(c) Normalised sum power measurements. The range throughout the dataset is so large a logarithmic scale has been adopted here.

Figure 2.9: The MLDC Challenge 2.2 dataset waveform, whitened spectrogram and normalised sum power measurements. There are three distinct signals visible (although the third is considerably more difficult to see).



(a) Trial coalescence time histogram for MLDC Challenge 2.2



(b) Trial chirp mass histogram for MLDC Challenge 2.2

Figure 2.10: Histograms of the best-fit trial parameter values of MLDC Challenge 2.2 dataset according to the rapid search. The parameter values of those signals that coalesced after the dataset finished have also been included, although there was no possibility that they would be detected by the rapid search method since it relied on the sharp drop in power associated with the coalescence.

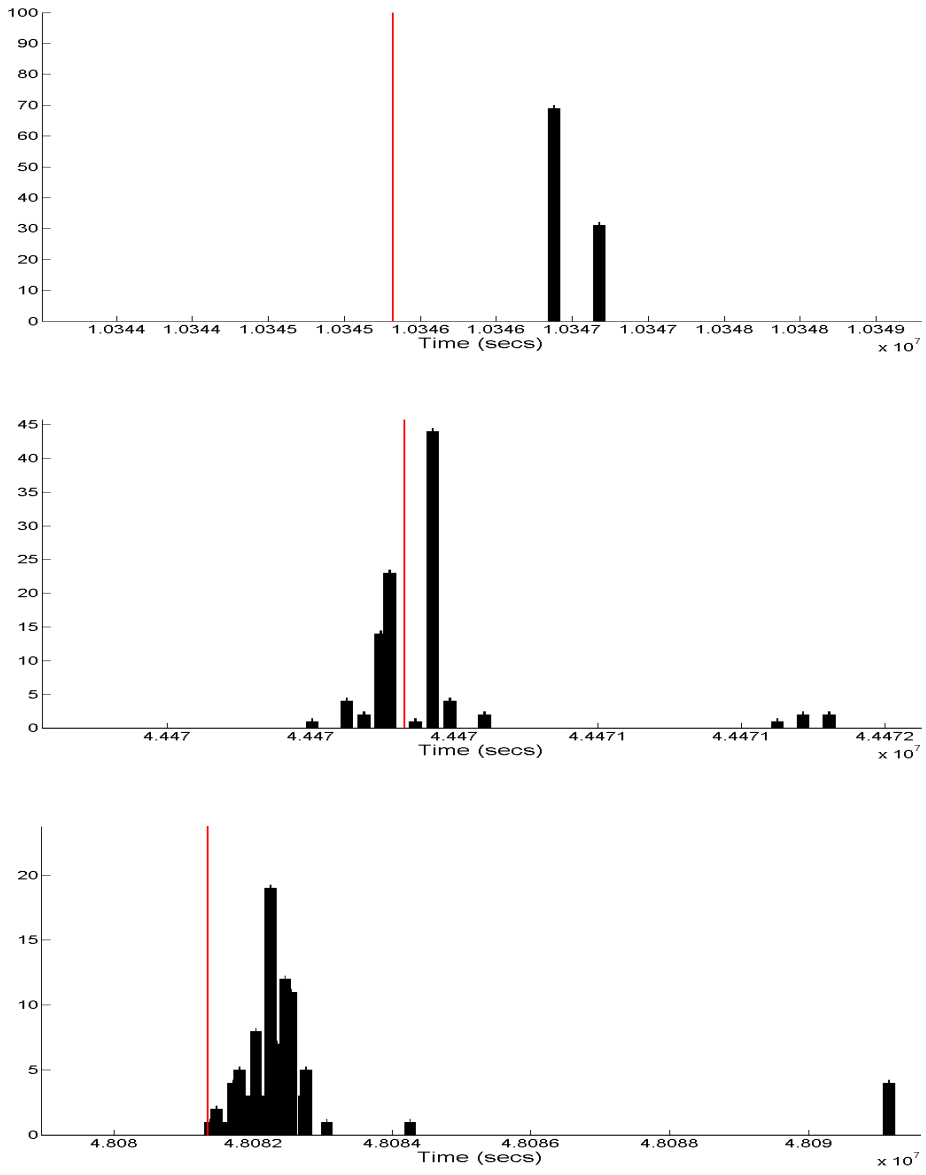


Figure 2.11: A closer look at the best-fit coalescence times MLDC Challenge 2.2 dataset according to the rapid search.

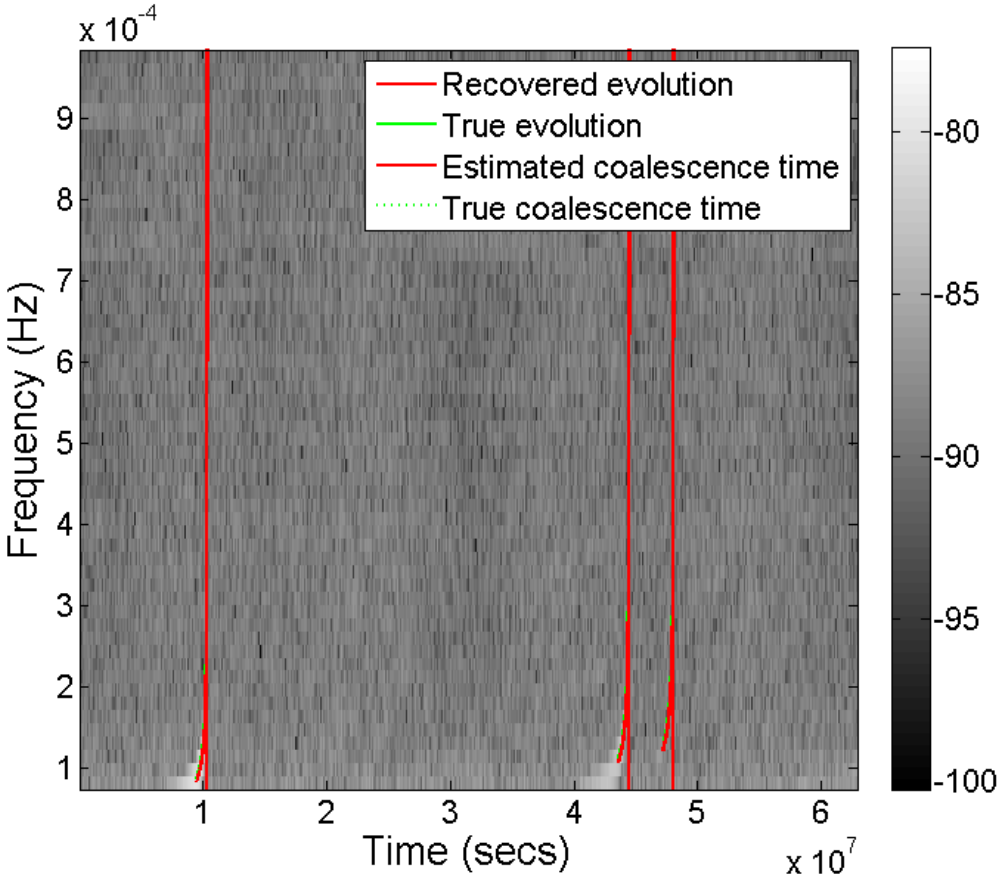


Figure 2.12: Spectrogram of the MLDC Challenge 2.2 dataset, with the true evolution and recovered evolution of the detected SMBH signals added. At this scale it is almost impossible to tell them apart, the lines representing them are too close. The natural logarithm of the power is displayed (indicated by the greyscale bar).

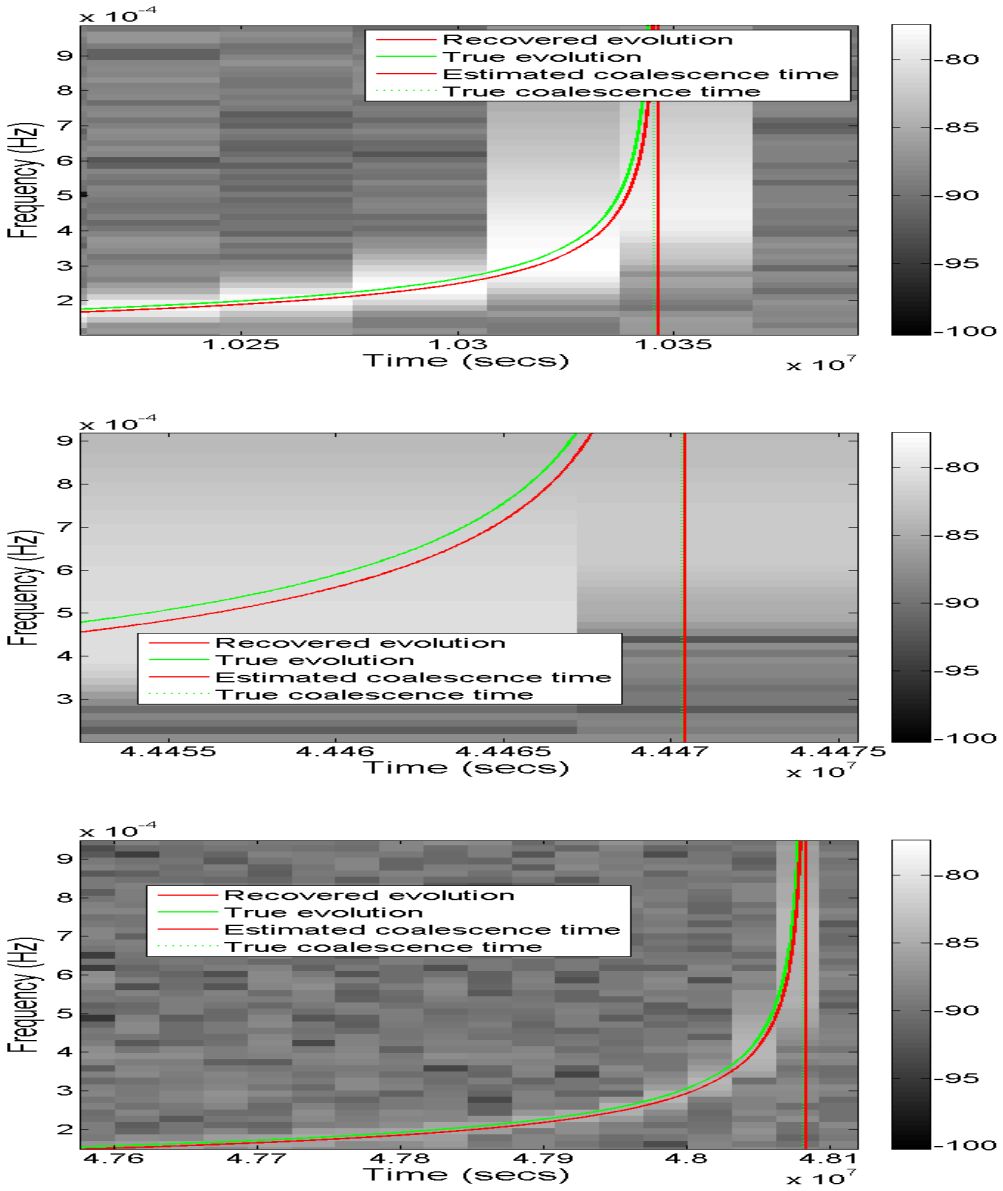


Figure 2.13: A closer look at the three SMBH signals in the MLDC Challenge 2.2 dataset with the recovered and true evolutions. The difference is easier to see, but the match is still good. The natural logarithm of the power is displayed (indicated by the greyscale bar).

2.3.1 Why SMBH signals are amenable to rapid searches

Principally, we have relied on the high SNR possible from SMBH binary gravitational wave signals. As seen previously this is occasionally high enough to identify some signals by sight in the MLDC data, and even when it is not, such as the third signal identified by the rapid search in the MLDC 2.2 dataset, the sum power measurements can identify the coalescence without user input. Unfortunately, without the high SNR the rapid search method struggles (as demonstrated by the MLDC 1.2.2 dataset). Thus far, the threshold applied to the sum power measurements required is unable to result in meaningful coalescence time estimates if it cannot extract the correct frequency evolution in the run-up to coalescence. Furthermore, because one threshold is applied to an entire dataset, we limit its robustness as a means of isolating different-SNR signals and requires us to look at the sum power measurements before choosing the threshold, distancing the method from a completely automated search. However, under the correct circumstances, this approach does work in the sense that we can pick out the evolution of the signal from background noise, something that would clearly not be possible without the high SNR.

We also rely on our assumption that the signal's quadrupole frequency within the spectrograms fits the characteristic shape we observed in the noise-free frequency evolution. As well as the stated assumptions about the frequency evolution, there was an implicit assumption that when we added the SMBH signals to a noisy time-series that this shape would not be distorted, and that a simple least-squares fit was the best approach to determine the extracted parameters. We assume too that the effects of the LISA antenna pattern and the source location information for each signal will similarly keep the frequency evolution in the desired shape. In its current incarnation, the rapid search cannot accommodate these possibilities, and cannot provide an estimate for any such parameter. Nevertheless, we must not lose sight of how we approached this problem; we decided to attempt a *rapid*, almost cursory, search for SMBH gravitational wave signals at the expense of accuracy, to use as estimates for more thorough searches. The attempt was deliberately simple, and our expectation was that we would produce simple and not necessarily informative results. By reducing the situation to a two-parameter search, the computational resources required were small; each search of an MLDC dataset took less than one minute on a standard desktop computer, and were re-run very easily (the results shown in the histograms of the 100 searches of each dataset took a little over an

hour to produce). A search of a larger parameter space (as discussed in section 3) would increase the computational demands, and risk the accuracy of describing the search as 'rapid'.

Crucial to the current working of the search is the advantage of a rapid drop in power after the coalescence time, without which the mergers will not manifest as spikes in the same manner. One approach is to address this advantage simply in terms of the data as presented in the mock challenges and to exploit known features. However, it is worth remembering that the datasets are designed to simulate LISA output and that the goal for analysis techniques is their application to this output for the purpose of extracting useful science. Although the features of the signal during the merger is not yet incorporated into the MLDC datasets, this may be a necessary step to ensure confidence in any search technique used, and would be particularly important for our excess power method.

Lastly, the formation and growth processes of SMBHs are not well understood and the coalescence event rate is highly uncertain. The number of these coalescences that are then detectable using the excess power method is therefore even more uncertain. Nevertheless, estimates of the coalescence rate is fairly low (36),(37), a necessity for the excess power method (although Sesana et al (38) suggest ≈ 100 during a three year mission is possible). The rapid search depends on being able to distinguish individual spikes and is vulnerable to mergers that coalesce within a short time interval. Should SMBH mergers (or other sources capable of producing sufficiently high SNR signals as to significantly alter the sum power measurements) prove to be more numerous than currently believed, the excess power method may not be useful even as a rapid search, continually producing false alarms and struggling to determine the frequency evolution of overlapping or otherwise competing sources.

2.3.2 Limitations of the rapid searches

The rapid search was a simple approach to a complicated situation, so the inaccuracy of the results produced are not unexpected. Nevertheless, as an initial foray into gravitational wave data analysis (as stated previously, the true purpose of this exercise) it provided a wealth of information that can be used to inform a more sophisticated attempt.

The search was indeed rapid at the resolutions considered, but becomes ever slower as these resolutions are increased. We are not increasing the parameter space itself (although select mass ranges were used, and the searches would be even slower if this range was increased) but simply the number of templates required for each search. It became apparent that we could quickly reach a state where the search could no longer be justifiably deemed 'rapid'. This would be unacceptable; we desired an alternative to any sort of exhaustive search and were willing to trade accuracy for speed.

Additionally, our approach was ad hoc; a solution designed specifically for the MLDC Challenges we had easy access to. While this was a sensible choice, there were drawbacks to being inextricably linked to these datasets. The robustness of the rapid search would be vulnerable to changes to the MLDC data, and such changes could take a number of forms.

Firstly, updates to the simplified waveforms could produce datasets that did not provide the feature exploited by the rapid search. Similarly, the noise (assumed to be stationary and gaussian) may not be an appropriate reflection of LISA's real noise output, and a more accurate simulation may obscure the characteristic frequency evolution feature. The whitening stage of the excess power method, while not strictly necessary in the presence of high SNR signals, does rely on being able to determine the average noise at different frequencies with some accuracy, and if this is no longer appropriate a more sophisticated account of the noise would be needed. Although not inherently problematic to incorporate, changes in the noise must be examined carefully and their effect on the frequency evolution extraction tested.

Additionally, the antenna pattern and the corresponding modulations to the gravitational wave signals as LISA's orbits were not included either (LISA was considered to be located at the solar system barycentre). While minor fluctuations may not effect the excess power method in the presence of a high SNR signal, weaker signals risk disappearing in the noise. Further, the process of whitening the spectrogram would have to be modified to take the orbital effects into account even if the form of the noise itself remained unchanged. Once again, the demands of gravitational wave measurements are evident- quite aside from the precision required, the long periods of observation introduce complications that must be taken into account. Thankfully there is no reason that the antenna pattern could not be introduced to the rapid search, but it does come at the cost of requiring datasets to be

accurately timed with respect to LISA's orbit rather than being applicable to any dataset ignorant of the precise time period under scrutiny.

Finally, as presented, the rapid search only makes use of the X channel in the MLDC datasets, ignoring data in the Y and Z channels; therefore the rapid search does not make use of all available information. Though it would likely come at an increased cost in terms of speed (although this can be countered by using faster, or multiple, CPUs), it might be possible to improve the accuracy of the search by looking for coincidental coalescence estimates in different data channels, or combining them in a manner that improved the SNR of the signal for frequency evolution extraction. There is nothing inherently problematic in using these other channels, simply that the rapid search method was designed only to consider a single data stream. By the time we had produced a working version of the procedures involved and had tested this on the X channel of MLDC datasets, the shortcomings of the method had become clear and we had resolved to abandon the notion of using excess power and evolution extraction in favour of a more sophisticated approach.

The excess power method rapid search fails to be a particularly useful tool for parameter extraction, and similar approaches are not the focus of current data analysis techniques; a complicated situation certainly seems to demand a complicated, or difficult-to-implement, solution (likely of the type suggested in section 1.2.4). It is clear that the rapid search, despite its strengths in taking advantage of some of the simple assumptions and situations in the challenge datasets, is not utilising all the information possible and pays a price for this in terms of the accuracy of its results. Nevertheless, implementing this method has proved instructive in highlighting the difficulties of analysing gravitational wave datasets as well as the particulars of MLDC datasets.

Chapter 3

EMRI problem space

The shortcomings of the rapid search exposes the dangers of over-simplifying the analysis problem and failing to take into account all of the necessary information. Nevertheless, it seems clear that a full solution to a complicated system may be unfeasible and we must consider how to approach such a situation. This chapter explores Extreme Mass Ratio Inspiral (EMRI) systems and their gravitational wave output as waveforms and spectrograms, and an introduction to the data analysis challenge posed by exploring EMRI 'problem space'- the parameter space inhabited by this output.

3.1 EMRI systems

An Extreme Mass Ratio Inspiral is the inspiral of a stellar mass compact object (CO) into a supermassive black hole (SMBH), and such systems are a fascinating potential source of gravitational waves for the LISA mission. The compact object, stellar remnants such as stellar black holes or neutron stars, are expected to have masses in the range $1M_{\odot} \lesssim \mu \lesssim 10^2 M_{\odot}$ while the supermassive black hole will be $10^5 M_{\odot} \lesssim M \lesssim 10^7 M_{\odot}$ (39): thus the mass ratio of the two is *extreme*. EMRI binaries are expected to occur as COs sink towards the centre of galaxies, perturbing their path until they pass close to the SMBH and become bound to it. The binary orbit then decays via gravitational wave emission at frequencies within LISA's band, before the CO plunges into its massive companion.

There are several reasons that make EMRIs particularly interesting. The extreme mass ratio means the CO can be treated as a perturbation of the spacetime of the

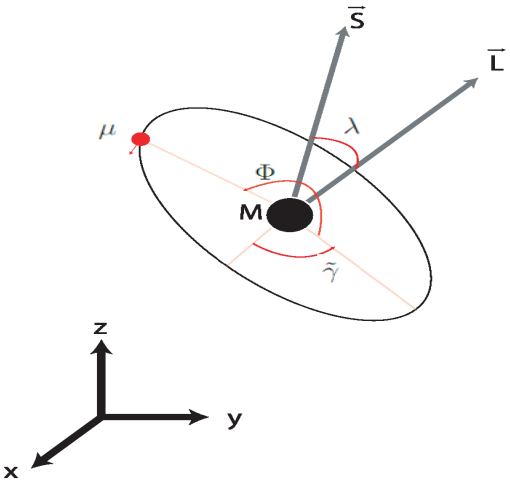


Figure 3.1: An EMRI binary system. The physical parameters are described in Table 3.1.

SMBH, and the gravitational waves emitted will encode information about this spacetime, allowing it to be mapped out and to determine whether these are in fact Kerr black holes as described by General Relativity. Additionally, LISA is expected to detect a few thousand inspirals out to a redshift $z \lesssim 1$ during the lifetime of the mission (40), (41).

However, extracting useful information from EMRI gravitational waves will prove a formidable technical challenge. The signal amplitudes are expected to be below that of LISA's instrumental noise, and further obscured by galactic binary confusion noise. Likewise, the antenna pattern for the spacecraft will modulate the signal over time. Most significantly, a fully coherent matched filter search would require $\sim 10^{40}$ templates (a rough estimate, explored in (41)), a computational task beyond even the most optimistic predictions of computer power available at the time of the LISA mission. Therefore, sub-optimal detection strategies must be considered, such as stack-slide search algorithms (42).

Nevertheless, successful extraction requires theoretical understanding of EMRI gravitational waveforms and a method for generating these. Almost inevitably, there is a conflict between the resulting accuracy of the waveforms and the ease with which they can be generated. A number of different approaches are examined briefly below.

MiSaTaQuWa waveforms

The waveforms calculated by Mino, Sasaki, Tanaka, Quinn and Wald (see (43), (44)) are significant for incorporating the first-order gravitational self-force influence of a test particle in a curved spacetime and the corresponding deviation from geodesic motion, representing a significant step towards what Drasco (39) characterises as 'Capra' waveforms; those of even greater ambition which require the derivation of higher-order parts of the self-force to create extremely accurate gravitational waveforms and in turn detailed predictions about gravitational wave measurements.

Indeed, Poisson sets out what he calls the *Capra scientific mandate*

- To formulate the equations of motion of a small body of mass m in a specified background spacetime, beyond the test-mass approximation.
- to concretely describe the motion of the small body in situations of astrophysical interest, including generic orbits of a Kerr black hole
- to properly incorporate the equations of motion into a wave-generation formalism

but this only acts as a framework of desirable, thus-far elusive, goals.

Teukolsky waveforms

Saul Teukolsky was able to derive a description for first-order gravitational field perturbations of a rotating black hole -as well as electromagnetic and neutrino field perturbations (45), (46)- encapsulated in the 'Teukolsky master equation'; Teukolsky waveforms are generally considered to be those based on solving this equation. The most common strategy is to iteratively extract snapshot waveforms by computing the geodesic orbits of a point particle on a background spacetime. These snapshots approximate the true waveform, and the rate of change of orbital constants inform the next iteration. By stitching together a sequence of the snapshot waveforms it is possible to approximate the inspiral waveform, and recent challenges have been to apply this technique to generic orbits (eccentric, evolving and inclined, with non-zero black hole spin angular momentum).

'kludge' waveforms

Kludge waveforms describe those waveforms based on different, potentially conflicting formalisms and assumptions. Less accurate than more rigorous approaches, they have the advantage of retaining significant features present in more realistic waveforms while being computationally cheap. Further, kludge waveforms are often far easier to adapt by incorporating or removing physical effects. This flexibility and availability makes them immensely valuable for examining data analysis techniques.

In the MLDC datasets, EMRI waveforms are calculated from the 'analytic kludge' method introduced by Barack & Cutler (see (47)) whereby the orbit is approximated as a Newtonian orbit at a given instant, but which evolves according to post-Newtonian equations including the effects of radiation reaction, pericentre precession and Lense-Thirring precession of the orbital plane on the orbital decay. The EMRI system produces a Peters-Matthews (48)waveform corresponding to the instantaneous orbit. The waveforms produced display the main features expected from true waveforms while being simple and quick to generate, with the expectation that search strategies successfully implemented on analytic kludge waveforms can be adapted to accommodate true GR waveforms once they become available.

Therefore, subsequent description and analysis on gravitational waveforms in this thesis will concern only those generated by the analytic kludge method. Analytic techniques will also be applied to simple situations (such as in Section 4.2), but it will be made clear that this is to test the robustness of the technique and the data used is not designed to be a realistic depiction of a gravitational waveform. Application of the analysis to different waveform types is discussed in Chapter 7.

3.2 MLDC waveforms, spectrograms and parameter space

The two-body EMRI systems presented in the MLDC datasets are described by 14 parameters [1], and for clarity this section (and the remainder of the thesis) follows the notation given in (W..e).

The SMBH and CO masses are M and μ respectively, while the magnitude of the spin angular momentum is S and the angle between the orbital angular momentum $\hat{L}$ is λ. The eccentricity and mean anomaly of the orbit are e and Φ, and the direction

[1] A complete description requires 17 parameters, but the spin of the CO is ignored, although it can be marginally relevant (see (47)).

t_0	The time where the orbital frequency sweeps through a fiducial value ν_0
μ	The mass of the CO
M	The mass of the SMBH
S/M^2	Magnitude of specific spin angular momentum of SMBH
e_0	$e(t_0)$ where $e(t)$ is the orbital eccentricity
$\tilde{\gamma}_0$	$\tilde{\gamma}(t_0)$ where $\tilde{\gamma}(t)$ is the angle in the orbital plane between $\hat{L} \times \hat{S}$ and the pericenter
Φ_0	$\Phi(t_0)$ where $\Phi(t)$ is the mean anomaly angle
θ_s	Source's direction polar angle
ϕ_s	Azimuthal direction to source
λ	$\cos^{-1}(\hat{L} \cdot \hat{S})$
α	$\alpha(t_0)$ where $\alpha(t)$ is the the direction of $\hat{L}$ around S.
θ_k	Polar angle of SMBH's spin
ϕ_k	Azimuthal direction of SMBH's spin
D	Distance to source

Table 3.1: The physical parameters of the EMRI system and their meaning. The notation is designed to inform the description of the waveform generation method.

of pericentre with respect to $\hat{L} \times S$ is $\tilde{\gamma}$. Together, these parameters are 'intrinsic'; they govern the evolution of the system independent of an observer's location or orientation.

In contrast, the 'extrinsic' parameters are those that locate the binary in time and space, and describe its orientation with respect to an observer; a Cartesian system based on ecliptic coordinates is used (wherein the Earth's orbit around the sun is the $x-y$ plane). The polar angle and azimuthal direction to the EMRI system are θ_s and ϕ_s, while θ_k and ϕ_k represent the same properties for the SMBH's spin. The angle α is the direction of $\hat{L}$ around S, and D is the source's distance. Lastly, many of these parameters vary in time as the orbit evolves, requiring the variable t that specifies when the other parameters hold their particular values.

Generating MLDC EMRI datasets

In practice, generating EMRI waveforms is a straightforward process, a strength of the analytic kludge approximations. A point is selected in the 14-dimensional parameter space representing the point of the CO plunge- in particular, t_0 defines

the time when the inspiral phase comes to an end and the unmodelled final plunge is said to occur. The selections are however, subject to some limitations: in the Schwarzchild metric, a point particles plunges at a frequency

$$\nu_{max} = (2\pi M)^{-1} \left[\frac{1-e^2}{6+2e}\right]^{3/2}, \tag{3.1}$$

and this is used to calculate the 'fiducial' frequency value ν_0 (that is, the radial orbital frequency at the time t_0). This value is the maximum allowed radial orbital frequency, beyond which the evolution is cut off.

The orbital evolution is determined by solving the ODEs 3.2 - 3.6 for $\nu(t)$, $\Phi(t)$, $\tilde{\gamma}(t)$, $e(t)$ and $\alpha(t)$ for the desired period of time . We used Matlab's built-in ODE solver to make these calculations. Fig. 3.2 shows the evolution of these parameters over a short period of time. Peters and Mathews (48) determined how to relate these variables to the second time derivative of the inertia tensor in a Newtonian binary system, and in turn this is used to explicitly express the n-harmonic components of amplitude strain for the two gravitational wave polarisations.

Finally, the gravitational wave strain h_+, $h_\times$ at the detector can be calculated from the strain amplitude coefficients of each polarisation. At this point, LISA Simulator and/or Synthetic LISA would be used to create the LISA TDI responses (see 1.2.3) $X(t)$, $Y(t)$ and $Z(t)$. However, we concentrated solely on the h_+ and $h_\times$ polarisations assuming an observer positioned at the solar system barycentre (SSB).

It should be noted that the most common approach is to solve the orbital evolution ODEs for distinct timesteps, each with constant seperation in time (MLDC datasets usually have this cadence set at 15 seconds). Each timestep will be initially given with respect to the time t_0, but these can be relabeled for the creation of MLDC datasets (which may contain multiple overlapping signals). A realistic calculation of LISA's response will depend on the times considered with respect to LISA's position and corresponding antenna pattern, but we omitted the effects of the spacecrafts' orbits. This omission is discussed further in Chapter 6.

$$\begin{aligned}\frac{d\nu}{dt} = {} & \frac{96}{10\pi}(\mu/M^3)(2\pi M\nu)^{11/3}(1-e^2)^{-9/2} \\ & \{[1+(73/24)e^2+(37/96)e^4](1-e^2)+ \\ & (2\pi M\nu)^{2/3}[(1273/336)-(2561/224)e^2-(3885/128)e^4-(13147/5376)e^6]-(2\pi M\nu) \\ & (S/M^2)\cos(\lambda)(1-e^2)^{-1/2}[(73/12)+(1211/24)e^2+(3143/96)e^4+(65/64)e^6]\},\end{aligned} \tag{3.2}$$

$$\frac{d\Phi}{dt} = 2\pi\nu, \tag{3.3}$$

$$\begin{aligned}\frac{d\tilde{\gamma}}{dt} &= 6\pi\nu(2\pi M\nu)^{2/3}(1-e^2)^{-1}[1+\frac{1}{4}(2\pi M\nu)^{2/3}(1-e^2)^{-1}(26-15e^2)] \\ &\quad -12\pi\nu\cos(\lambda)(S/M^2)(2\pi M\nu)(1-e^2)^{-3/2},\end{aligned} \tag{3.4}$$

$$\begin{aligned}\frac{de}{dt} &= -\frac{e}{15}(\mu/M^2)(1-e^2)^{-7/2}(2\pi M\nu)^{8/3}[(304+121e^2)(1-e^2)(1+12(2\pi M\nu)^{2/3}) \\ &\quad -\frac{1}{56}(2\pi M\nu)^{2/3}((8)(16705)+(12)(9082)e^2-25211e^4)] \\ &\quad +e(\mu/M^2)(S/M^2)\cos(\lambda)(2\pi M\nu)^{11/3}(1-e^2)^{-4} \\ &\quad [(1364/5)+(5032/15)e^2+(263/10)e^4]\end{aligned} \tag{3.5}$$

$$\frac{d\alpha}{dt} = 4\pi\nu(S/M^2)(2\pi M\nu)(1-e^2)^{-3/2}. \tag{3.6}$$

In general, we are not attempting to generate MLDC datasets. Our analysis will concentrate first on isolated EMRI signals before dealing with multiple signals, and will deal primarily with the $h_\times$ polarisation of the gravitational waveforms without the effect of the antenna pattern or looking at TDI responses. The necessity and wisdom of this approach will be addressed throughout the analysis, but for the moment we can consider this to be the 'output signal'.

Generating spectrograms

The 14-dimensional parameter space of EMRI systems poses significant computational challenges, and some of these difficulties can be dealt with by considering the spectrograms of the resulting gravitational waveforms of the EMRI signals. As Fig. 3.3 shows, the waveforms are not immediately revealing judged simply by eye, and the spectrograms (see Fig. 3.4) do provide a more instinctive (to the author's eye, at least) way of viewing the evolution of the waveform. However, the true advantage of examining spectrograms rather than the waveforms themselves is more fully explored in Section 4.1.2.1.

The method used to generate EMRI spectrograms from the waveform timeseries follows the notation of (35). In practice, it is identical to that presented in 2.2 except that the timeseries being considered is the gravitational wave strain amplitude rather than the LISA data channels.

The gravitational waveform is represented by a vector of M real numbers x_k, $k = 1...M$ representing the strain amplitude (h_+ or $h_\times$) with a sample frequency of

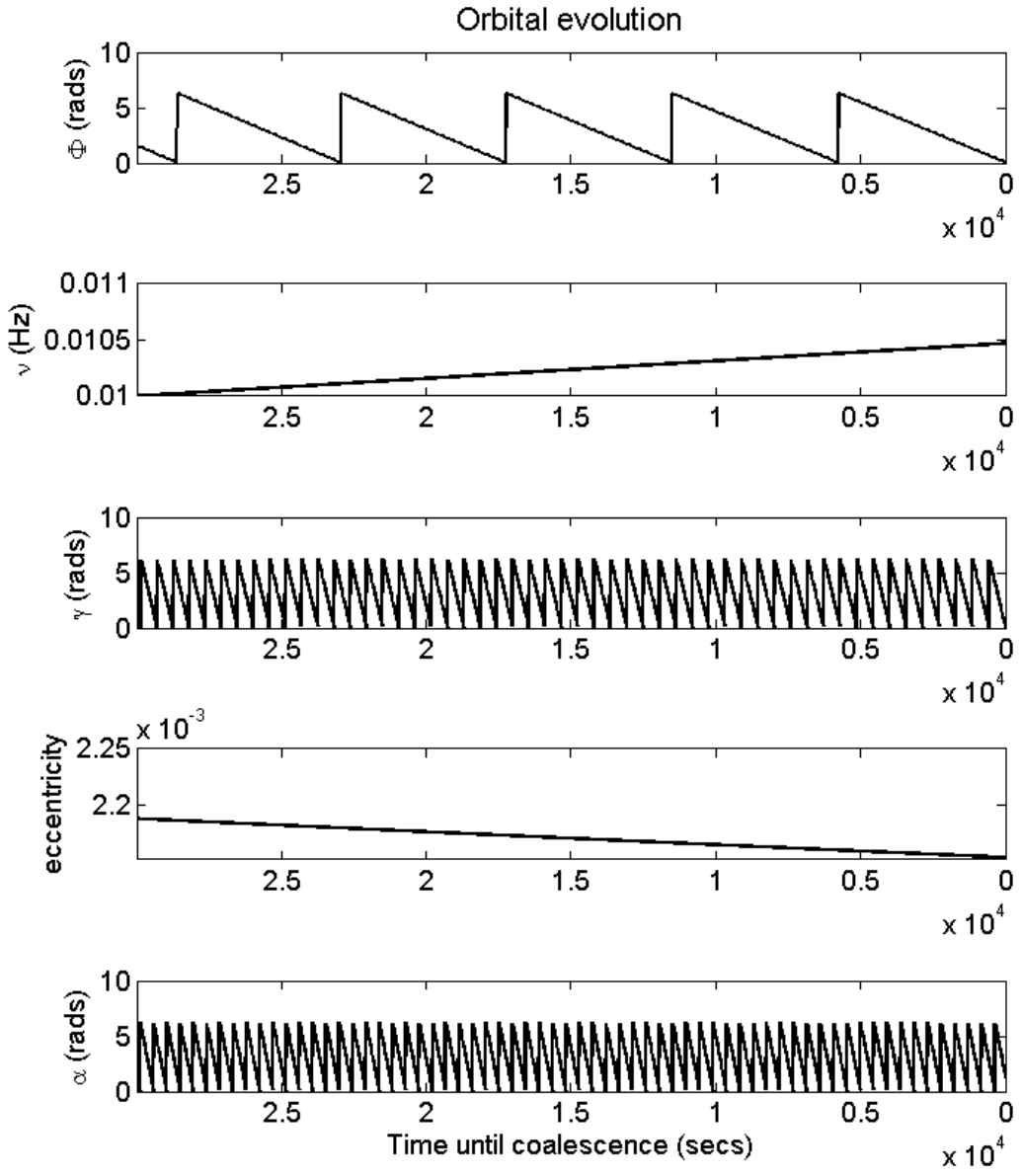


Figure 3.2: The 5 orbital evolution parameters evolved over a short period of time, using a randomly selected list of initial parameters. The saw-tooth appearance in some is the result of the angles being calculated modulo 2π.

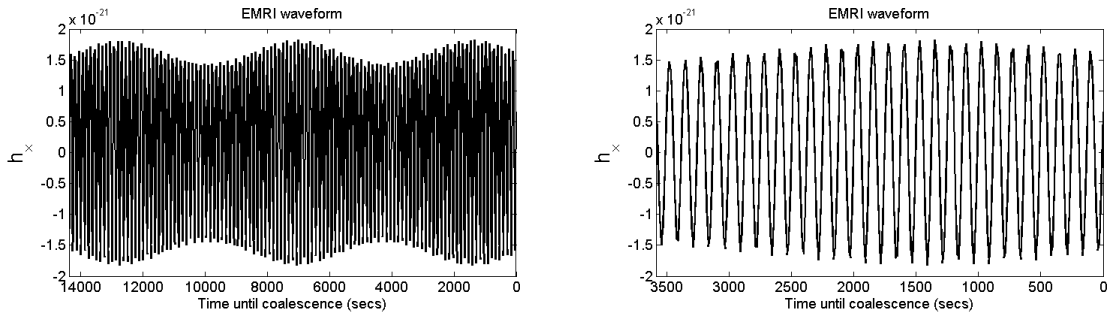


Figure 3.3: An analytic kludge EMRI waveform in the four hours preceding the coalescence, and a closer look at the hour before the final plunge. The evolving sinusoidal nature of the waveform can be seen, but specific details are hard to make out by examining the timeseries alone. The parameters were chosen from a small subset of the whole EMRI parameter space (matching those of the waveform ξ described in Section 6.1), but is intended simply as an example of a 'typical' EMRI waveform before its coalescence.

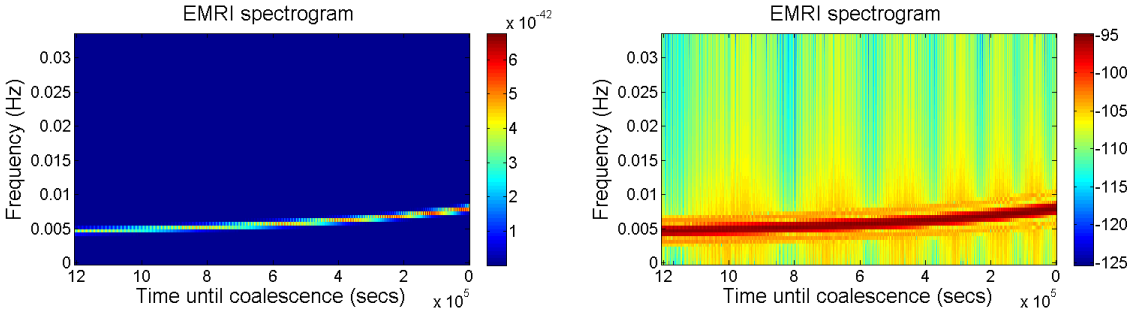


Figure 3.4: A spectrogram of the $h_\times$ polarisation an analytic kludge gravitational waveform. Also included is a plot displaying the natural logarithm value of the power indicated by the colour bar, revealing some fine structure in the signal (that is the higher frequencies have some power in them, which is not immediately clear from the left-hand spectrogram). The waveform used to generate the spectrogram is the same as in Fig. 3.3, but covers a period of approximately two weeks preceding the coalescence.

f_s. The power spectral density of this waveform is then

$$PSD = \frac{2 \cdot |y|^2}{f_s \cdot S_2} \tag{3.7}$$

where y_m is the short-time Fourier transform (of length N) of x_k, w_j represents the vector of real numbers of length N that makes up the window function, and

$$S_2 = \sum_{j=1}^{N} w_j^2. \tag{3.8}$$

The specific values of the above properties are stated in the relevant sections (see 4.2 and 6), but unless otherwise stated the MLDC convention of a $f_s = 1/15$ Hz sample frequency is used, and the window function is a Hanning window, defined as

$$w_j = \frac{1}{2}\left[1 - \cos\left(\frac{2\pi \cdot (j-1)}{N}\right)\right]. \tag{3.9}$$

Chapter 4

Mapping the EMRI parameter space

4.1 Principal Component Analysis

The extent of the EMRI parameter space and the computing cost of generating even the simple analytic kludge waveforms makes an optimal search for EMRI signals unfeasible, requiring a non-optimal approach to the problem.

A sample of EMRI spectrograms (see C.3) reveals that they do not always differ significantly when generated with distinct parameters, suggesting that the parameter space with regards to pattern matching might have a reduced dimensionality. The following chapter discusses the procedure of Principal Component Analysis (PCA) to map this reduced parameter space, an its application to EMRI spectrograms in a small region of the original parameter space.

4.1.1 PCA Theory

Principal Component Analysis (also known as the Hotelling transform or Karhunen-Loève decomposition) is an attempt to account for the variance in a dataset of observations of a set of correlated variables with a smaller set of uncorrelated variables called *principal components*. It is often helpful to visualise the observations as a cloud of points in a high dimensional space spanned by the variables, and the first principal component defining line in the variable space of closest fit to these points, accounting for as much of the variance in the dataset as possible. The second principal component is a line perpendicular to the first accounting for as much of the

remaining variance as possible and defining a plane of closest fit with the first principal component. Further principal components define hyperplanes in the variable space while accounting for yet more of the dataset's remaining variance.

The number of principal components cannot exceed the dimensionality of the variable space, but in the situation that there are linear dependencies in the variables there may be fewer. Thus, it may be possible to account for all of, or the majority of, the variance in the dataset by using a subset of the principal components. By doing so, the *redundancy* of some of the variables can be used to reduce the parameter space, making it more amenable to automated searches.

For clarity, it is easiest to describe PCA with respect to EMRI spectrograms from the outset. Motivated in part by the work of Turk and Pentland (49) (whose mathematical conventions are followed closely throughout this chapter), each spectrogram may be considered for the moment simply to be one point in a high-dimensional space encompassing every possible EMRI spectrogram: *the problem space*. The spectrograms have N time and frequency bins, so the problem space has N dimensions; each point is defined by a vector of length N.

A set of *training* points $\Gamma_1, \Gamma_2, \Gamma_3, ...\Gamma_M$, and their mean-subtracted counterparts $\Phi_1, \Phi_2, \Phi_3, ...\Phi_M$ (where $\Phi_n = \Gamma_n - \Psi$ and $\Psi = \frac{1}{M}\sum_{n=1}^{M}\Gamma_n$), are used to create a set of M vectors u_i chosen so that the variance

$$\lambda_k = \frac{1}{M}\sum_{n=1}^{M}(u_k^T\Phi_n)^2 \tag{4.1}$$

is maximised, constrained by the demand that

$$u_k^T u_k = \delta_{lk} = \begin{cases} 1, & \text{if k=l;} \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

That is, the vectors are orthonormal, and account for the maximum amount of the training points' variance.

The variance of a linear combination $u_k^T\Phi_n$ is $u_k^T C u_k$, where

$$C = \frac{1}{M}\sum_{n=1}^{M}\Phi_n\Phi_n^T \tag{4.3}$$

is the covariance matrix of the mean subtracted training set. To maximise the variance subject to the orthonormal constraint, Lagrange multipliers are utilised

(see (50), (51)), requiring a solution to the equation

$$\frac{d}{du_k}\left(u_k^T C u_k - \lambda(u_k^T u_k - 1)\right) = 0 \tag{4.4}$$

giving

$$Cu_k = \lambda u_k \tag{4.5}$$

or

$$(C - \lambda I_N)u_k = 0, \tag{4.6}$$

where I_N is the $(N \times N)$ identity matrix. Therefore, λ is an eigenvalue of the covariance matrix C, and u_k is the corresponding eigenvector.

It may not be obvious which eigenvector of the covariance matrix gives $u_k \Phi_n$ the maximum variance, but the value to be maximised is $u_k^T C u_k$, and thus

$$u_k^T C u_k = u_k^T \lambda u_k = \lambda u_k^T u_k = \lambda, \tag{4.7}$$

meaning that the eigenvalue λ should be as large as possible. By arranging the eigenvalues in order of decreasing magnitude, the associated eigenvectors are arranged in order of the amount of variance they account for. The first eigenvector is termed the first principal component and accounts for the most variance in the training set, the second accounts for the largest portion of the remaining variance and is the second principal component, and so on until all the variance is accounted for[1]. It is a straightforward exercise to calculate the variance in the dataset contained by the first p principal components as a fraction of the total variance by

$$\frac{\sum_{p=1}^{k} \lambda_i}{\sum_{p=1}^{M} \lambda_i}. \tag{4.8}$$

4.1.1.1 Avoiding the covariance matrix

The covariance matrix is symmetric, and therefore orthogonally similar [2] to a diagonal matrix whose entries are the eigenvalues of the covariance matrix (see (52)) and there is an orthogonal matrix whose columns are the corresponding eigenvectors. In short, it is always possible to find a set of principal components of the EMRI spectrograms by performing the *spectral decomposition* of the covariance matrix.

[1] Jolliffe (51), provides the proof with respect to the size of the eigenvalues.

[2] matrices A and B are orthogonally similar if there exists an orthogonal Q such that $B = Q^T A Q$.

Unfortunately it will be computationally expensive to determine the N eigenvectors of the covariance matrix, but in the situation where the training set of points is smaller than the problem space dimensionality there will only be $M-1$ eigenvectors with an associated non-zero eigenvalue (or fewer, if linear dependencies exist).

The covariance matrix can be expressed as

$$C = AA^T, \tag{4.9}$$

with $A = [\Phi_1 \ \Phi_2 \ ... \ \Phi_M]$, and will be $N \times N$. Turk and Pentland (49) showed that by considering the matrix $L = A^T A$, where $L_{mn} = \Phi_m \Phi_n^T$, and by determining the M eigenvectors of this matrix ($M < N$), it is possible to determine the eigenvectors of the covariance matrix.

$$(A^T A)v_l = \mu_l v_l \tag{4.10}$$

can be premultiplied by A

$$(AA^T)Av_l = \mu_l Av_l \tag{4.11}$$

revealing that $Av_l = u_l$ are in fact the eigenvectors of the covariance matrix of the dataset. This method is not as computationally expensive as manipulating the covariance matrix directly, and at this stage the eigenvectors are normalised. This process forms a basis set of vectors that span the space inhabited by the training set of spectrograms, and potentially the entire problem space.

4.1.2 Application of PCA to EMRI spectrograms

In practice, spectrograms are presented as an $A \times B$ image (marking specific time and frequency bins, and the power contained in each bin), which can be reshaped as a column vector of length N. The PCA method detailed previously applied to each of these vectors produces a set of orthogonal 'eigenspectrogram' column vectors that span the space inhabited by the training spectrograms. Each eigenspectrogram is a principal component of the training set space, and the two terms are almost completely interchangeable[1]. However, it is unlikely that this space covers the entire problem space, requiring the capacity to review the quality of the eigenspectrograms generated and to change them as needed.

[1]In the remainder of this work it may be assumed that the eigenspectrogram have been normalised, unless otherwise stated.

A *test* spectrogram Γ can be 'projected' into the space spanned by the eigenspectrogram basis set (henceforth the *eigenspace*) with the operation

$$\omega_k = u_k^T(\Gamma - \Psi) \tag{4.12}$$

creating a vector $\Omega = [\omega_1, \omega_2, ..., \omega_M]$ describing the contribution, or 'weight', of each eigenspectrogram to the spectrogram. Two methods of assessing test spectrograms immediately present themselves;

Spectrogram reconstruction

A test spectrogram may be reconstructed to some extent by summing the eigenspectrograms according to its weight vector. Formally then, the reconstruction of the spectrogram is

$$\Phi_R = \Sigma_{i=1}^{M'} \omega_i u_i \tag{4.13}$$

noting that $M' \leq M$. The distance between the reconstruction and the point in the problem space inhabited by the (mean-subtracted) test spectrogram, the 'residual distance', is then

$$D = \parallel \Phi - \Phi_R \parallel . \tag{4.14}$$

A simple measurement of what we might think of as the 'quality' of the reconstruction, is to express the difference between the original spectrogram and the reconstructed version as a fraction of the power contained by the original spectrogram. We will henceforth refer to this measurement simply as the **residual**, denoted R and calculated as

$$R = \frac{\parallel \Phi - \Phi_R \parallel}{\parallel \Phi \parallel}. \tag{4.15}$$

Spectrogram classes

Spectrogram classes may be defined using a number of known spectrograms and their weight vectors. The distance ϵ between a test spectrogram and kth spectrogram class is calculated as

$$\epsilon^2 = \parallel \Omega - \Omega_k \parallel^2 \tag{4.16}$$

using the test spectrogram weight vector Ω and the weight vector of the kth spectrogram Ω_k.

With these two assessment methods, there are four possible categorisations of test spectrograms:

1. **Near the eigenspace and near a spectrogram class** In this situation, the spectrogram is close to or inside the space spanned by the eigenspectrograms, and its projection vector is similar to a previous spectrogram. It will not contribute much new information with which to better define the problem space, although this depends on the threshold for defining 'near'. The most extreme case of this category is where there is zero distance to a particular spectrogram class and a perfect reconstruction, indicating that the test spectrogram belongs to the training set used to create the principal components (assuming each of these defined a spectrogram class).

2. **Near the eigenspace but far from a spectrogram class** A spectrogram in this category is also close to or within the subspace already defined by the principal components, but is not 'known' (that is, it doesn't appear to project into the eigenspace like any other spectrogram recorded). Once again, there is an associated threshold that must be considered defining 'far'.

3. **Far from the eigenspace and near a spectrogram class** Most likely a false-positive, the spectrogram appears to belong to a spectrogram class but does not belong to the eigenspace. This situation arises when the quality of reconstruction must be high, but the distance to different spectrogram classes is allowed to be large. This category requires special care and is explored further in 4.1.3.

4. **Far from the eigenspace and far from a spectrogram class** This indicates that the test spectrogram is significantly different from others, inhabiting a region of the problem space far from the eigenspace, and projecting into the eigenspace to create an unfamiliar weight vector. A similar result may be gained by substituting the test spectrogram for something distinctly alien column vector, such as that created from an entirely different waveform. In the application of PCA to face recognition (see (49)), images fall into this category when the system is presented an image that is not a face at all.

4.1.2.1 Why a spectrogram-based eigenspace?

Spectrogram-based PCA may seem counter-intuitive: observations of a problem space defined by 14 variables have been replaced by observations of an N-dimensional problem space (where $N \gg 14$), and the time resolution has been reduced from

seconds to hours. It is therefore wise to examine what makes spectrograms amenable to PCA and the best choice.

The most straightforward approach might appear to be to simply use the EMRI gravitational wave timeseries. Fundamentally, the method would remain unchanged from that described previously, but the phase of the waveform can radically change the principal components produced. Two waveforms that differed only in phase would produce two distinct principal components despite their similarities. Constructing spectrograms destroys this phase information and would only produce one principal component (the second spectrogram could be reconstructed from the information in the first). Naturally, redefining the situation should not be done thoughtlessly, but COs are expected to undergo thousands of orbits of the SMBH, suggesting that the initial orbital phase is marginally relevant.

4.1.3 A method for spanning EMRI parameter space using eigenspectrograms

A simple iterative algorithm for a principal component-based attempt to span EMRI spectrogram parameter space takes the following form:

1. A selection of training spectrograms are generated, and the corresponding eigenspectrograms created, defining the eigenspace.

2. Generate a batch of test spectrograms, and project them into the eigenspace, creating a set of weight vectors.

3. Reconstruct the test spectrograms according to their weight vectors and the calculated eigenvectors. A threshold R_T for the residual is considered:

   (a) $R > R_T$: spectrogram cannot be well reconstructed and therefore contains new information which should incorporated into the eigenspectrogram set. The spectrogram is considered to be a *candidate*.

   (b) $R < R_T$: spectrogram is well reconstructed, and does not include enough new information to expand the eigenspectrogram set.

4. Any spectrogram designated as a candidate is incorporated into the set of eigenspectrograms, enlarging the eigenspace. The mean spectrogram is updated to include the contribution from the new spectrograms.

5. Steps 2, 3 and 4 are repeated until a termination condition is met.

Each part of this framework requires attention.

Generating training spectrograms

To some extent, the spectrograms used to create the training set are unimportant. If, as is likely, the training set does not contain all relevant information to create an eigenspace that accounts for most of the variance in the problem space, test spectrograms will be poorly reconstructed by the eigenspectrograms generated and the dataset will be expanded on the next iteration. Should the training set of spectrograms actually contain all relevant information on the other hand, the algorithm will have met any sensible termination condition and have run to completion.

Nevertheless, the training set must contain only example EMRI spectrograms (within the desired parameter space, in the event that only a subspace of the problem space is being examined) since they are not assessed like test spectrograms, and are used to define an eigenspace straight away. PCA does not 'know' what an EMRI is, or any astrophysical meaning behind its input or output. For example, the inclusion of a SMBH spectrogram in the training set would create an eigenspace that incorporated some part of SMBH parameter space, and prevent similar spectrograms from being rejected when using the eigenvectors to detect EMRI signals (see chapter 6).

In practice, finite computing resources means that a limited number of iterations can be performed. On a standard desktop computer (3 Gb of RAM, 2.61 GHz), it was possible to store around 700 eigenspectrograms in memory at once (the precise number varied from test to test, dependent on the information content of each eigenspectrogram), allowing the mathematical operations required to determine residual values to be processed quickly. It was also possible to store individual eigenspectrograms in memory in order to make the desired calculations rather than store one large array, allowing us to retain far more that 700, but at the cost of increased processing time (as each eigenspectrogram would have to be read from memory and then discarded before the next one was loaded). We decided to prioritise speed, knowing that we would have to run several tests multiple times as we developed our algorithms. However, the requirement of large eigenspaces revealed in some of these tests (see later) suggested that without a substantial increase in available

RAM memory, the latter method of organising the data would be necessary to span problem spaces well. In its current incarnation at least, the PCA examined here was beyond the capabilities of a standard laptop computer at the time of writing. It is reasonable to expect that increases in available processor speed over the next few years -since we have time to develop our PCA before the LISA mission begins- would mitigate the problems of loading individual eigenspectrograms, but a more rigourous comparison between the two possible methods of storage (and retrieval) was not explored. There is a balancing act between how quickly the algorithm can go through each iteration, and how much information can be held in memory at any given time.

Another issue here is how to generate useful spectrograms- obviously an exhaustive progression through the original EMRI parameter space is unfeasible (and the motivation for this approach), but any prior information about the structure of the problem space can be easily incorporated into the algorithm from the generation of the training set, and continued with the generation of test spectrograms. There is nothing to prevent a distribution changing over time by incorporating new information either. Therefore, it would be possible to incorporate Baye's theorem and an MCMC approach to selecting points in a parameter space in order to sample from regions that produce high residual values. However, it is crucial to understand that the PCA algorithm will simply transform the data it is provided in an attempt to reveal its underlying structure, but does not itself incorporate prior information about this structure.

Generating test spectrograms

Again, a static or evolving (by which we mean one that incorporates new information after an iteration of the PCA) distribution may be consulted for the purpose of selecting the parameters used to generate the spectrograms. While it is possible simply to generate a single test spectrogram for every iteration of the algorithm, it is more computationally efficient to generate batches of test spectrograms and assess them all using the existing set of eigenspectrograms. The number of test spectrograms generated per batch is usually dependent on computational resources, with between 30 and 50 being an efficient balance between processing time per iteration and memory requirements.

Furthermore, it is easier to take candidate spectrograms, find the eigenspectrogram set of these using the existing average spectrogram for mean subtraction,

and then join it to the existing eigenspectrogram set. The principal components of this enlarged set, once normalised, are the eigenspectrograms for all of the training set and candidate spectrograms. Although it may initially seem confusing, this process still ensures that only those spectrograms sufficiently different from those already within the defined eigenspace will be used to expand it. In the event that several spectrograms in one test batch are designated candidates despite being identical, their inclusion will simply result in one or more additional zero-eigenvalue eigenspectrograms in the expanded set, which are automatically removed. A more complicated situation arises if several candidates are similar but not identical: had the algorithm been iterated one test spectrogram at a time, the inclusion of the first candidate may have prevented the next test spectrogram from being designated a candidate depending on the residual threshold. However, calculating the expanded set of eigenspectrograms will simply result in one or more additional eigenvectors with small eigenvalues, which can be eliminated at this stage (and are, by default). Nevertheless the mean spectrogram will be slightly different once updated, but this is unlikely to be significant unless the residual threshold is particularly low and all of the eigenspectrograms are kept (which is counter-productive in an attempt to reduce the parameter space).

Like selecting appropriate training spectrograms, it is assumed that the test spectrograms will in fact be constructed from analytic kludge waveforms in the manner previously discussed. The algorithm above is designed to map out the problem space, not to search within it (this is considered in Chapter 6), and PCA adheres to the notion of 'rubbish in, rubbish out'; unless the residual threshold R_T is extremely high, non-EMRI spectrograms will appear radically distant from the established eigenspace and the information they contained will be used to expand it.

Quality of reconstructions: the residual

The reconstruction residual threshold R_T gives an intuitive measure of the distance of the test spectrograms from the eigenspace, and can be easily expressed as a percentage of the original spectrogram if desired. In the above method it is the only measure used, but it is not immediately obvious what constitutes a sensible threshold, or even whether the threshold should remain unchanged or evolve over the course of many iterations.

Despite this, it is easy to imagine two situations that are best avoided. First, that the threshold is set too high, in the sense that every test spectrogram is considered 'well reconstructed' and the eigenspace is never expanded because no new information is incorporated. Second, that the threshold is too low, and even very similar spectrograms are used to expand the eigenspace by minute amounts each iteration. The ideal situation strikes a balance between these two extremes, producing meaningful output in a reasonable time, but how to achieve this balance is not clear without preliminary runs of the algorithm.

The algorithm above does not include the effects of, or attempt to define, spectrogram classes, since at this point there is no clear way to approach such classification. Simply judging by eye is not useful, since there is no obvious feature that might be the signature of a particular type of EMRI (at least not with the limited numbers of spectrograms it is possible to generate and examine), and clustering in the high dimensional problem space would be impossible to visualize. PCA, blind to the underlying structure of the dataset used until it is presented to the algorithm, does not suggest any classes itself. Further, at this stage PCA is not being used to examine a known signal, but this notion is explored in 6.2.

A more extensive exploration of a problem space consisting of all gravitational wave signals might benefit from classes, if different signal types produce spectrograms different enough to be group in distinct clusters- this is discussed further in chapter 7 concerning the applicability of a similar approach to PCA, *linear discriminant analysis*.

Termination conditions

The termination conditions are likely to be motivated primarily by computing resources; the processing time allowing a certain number of iterations to be performed or the hardware available dictating the size of the matrices that can be manipulated. Alternatively, the algorithm might be designed to terminate based on the PCA output in a variety of ways before these hard limits. For instance, it could be programmed to stop when the variance in the eigenvalues falls below a given limit, or when one thousand (say) consecutive test spectrograms consistently fall below a certain residual threshold. What constitutes a sensible termination condition must be given considerable thought, but early choices when largely or entirely ignorant of the expected PCA output are likely to be fairly simple.

What is certain, is that to be useful the mapping process must stop at some point so that its performance can be analysed. The eigenvectors produced are simply a reinterpretation of the data generated- what is important is if, and how, it aids comprehension of the original problem.

4.2 Initial tests on sinusoids

Having established an iterative method for spanning the EMRI spectrogram problem space using eigenspectrograms, the following section provides details of the initial efforts to implement it under a variety of circumstances in order to evaluate its performance. Therefore it is worth reiterating the expectations from the proposed principal component analysis method: a 'successful' implementation will provide the principal components of the EMRI spectrograms; a set of orthogonal vectors which span the problem space and can therefore be used to construct any EMRI spectrogram. Further, it will be possible to identify non-EMRI spectrograms since these will not inhabit the problem space.

Properly developed, these abilities will provide a basis for a method that searches LISA data for EMRI signals. Nevertheless, ignorant of the extent of the EMRI spectrogram problem space, and aware of the complicated structure of the spectrograms generated, it is sensible to test the PCA method on much simpler spectrograms first, followed by tests of the EMRI spectrograms within restricted regions of parameter space. Not only does this afford a greater level of control over the situation and thus allows errors to be spotted more easily, but is less computationally expensive than tackling the entire EMRI problem space.

4.2.1 PCA performed on sinusoids

The analytic kludge EMRI waveforms appear to be slowly-evolving sinusoids. Consequently, the timeseries of simple sinusoids are the obvious choice from which to construct spectrograms to test the proposed principal component analysis method. There are several aspects of these spectrograms, and their relevance to EMRI spectrograms, that must be examined separately. For this reason, and to highlight differences between them, it is important to keep the construction of the spectrograms as consistent as possible.

Therefore, the sinusoid spectrograms being considered are all created from $M = 2^{13}$ -sample long waveforms, with a sample frequency of 800 Hz. The fourier transform and Hanning window applied to the timeseries have lengths of 512 samples, with a 50% overlap between segments. The spectrograms produced have a frequency resolution of 1 Hz and covers the range $1 - 100$ Hz[1] while the time resolution is 0.32 seconds.

The spectrograms are deliberately significantly different from those produced from MLDC datasets. As stated previously, this is primarily to make the analysis less computationally expensive and easier to manipulate, but has a further benefit; to act as a transition between the theoretical desciption of the PCA and its application to EMRI spectrograms. By performing tests on simple sinusoid spectrograms, we are able to demonstrate the robustness of our approach and clearly show its effect on the data it is provided.

4.2.1.1 Simple monochromatic spectrograms

The most straightforward contruction is a set of spectrograms created from monochromatic waveforms, with frequencies *randomly* chosen from the frequency range defined above. Furthermore, only a single waveform amplitude is permitted. As expected, the spectrograms display lines centred at the chosen frequency that have constant power over the entire time period in question (examples are given in Fig. 4.1). Initially, the training set is composed of ten spectrograms, and the initial eigenspectrograms are calculated from this set.

The PCA method is then iteratively applied by producing a number of test spectrograms and determining the principal components' ability to reconstruct them. Each iteration only introduces one new monochromatic spectrogram with a randomly chosen frequency, and a threshold for the residual measurment of the reconstruction is set at $R = 0.001$. Any reconstruction that is above this threshold is used to expand the set of principal components, while those that fall below it are disregarded. Spectrogram classes are not used to evaluate test spectrograms in this situation, but will be used to examine the resulting eigenspace.

There are 100 unique spectrograms that can be generated under these conditions. Although all of the power is not contained at one frequency (some power leaks into

[1]Matlab allows the user to specify a vector of frequencies and calculates the spectrogram using the Goertzel algorithm (53). Specified frequencies are rounded to the nearest discrete fourier transform bin commensurate with the signal's resolution (see (W..m)).

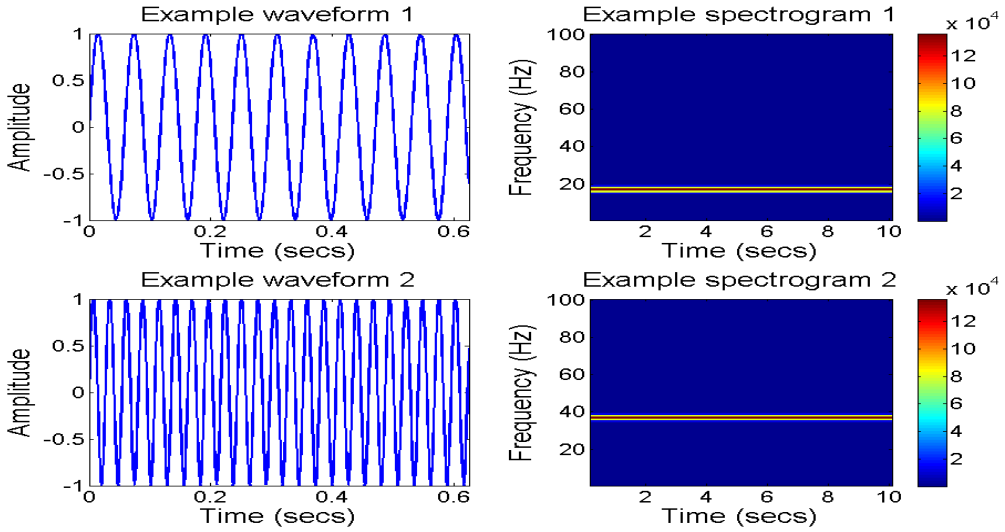


Figure 4.1: The timeseries of two monochromatic sinusoids (17 Hz and 37 Hz) and the corresponding spectrograms. Only the first 500 data points of the waveforms are displayed in the left-hand plots.

adjacent frequency bins), spectrograms generated from waveforms with significantly different frequencies have almost no overlap in the distribution of power, and the PCA method therefore creates 100 principal components; each distinct spectrogram cannot be reconstructed with contributions from the others, and must be incorporated into the set of eigenvectors. After this new test spectrograms are inevitably copies of previous spectrograms and can therefore be reconstructed from the eigenspectrograms almost perfectly (with extemely small amounts of residual power due to rounding errors): the eigenspectrograms span the problem space. The quality of the reconstructions over 500 iterations is displayed in Fig. 4.3, and a complete set of eigenspectrograms is included for reference in Appendix C.1.

With the defined eigenspace covering the entire problem space, some additional test spectrograms were created as examples of one of the categories described in section 4.1.2. A definition of spectrogram classes is required, but this can be provided unambiguously by having each of the 100 spectrograms form its own class. The weight vector of each of the 100 unique spectrograms projected into the 100 principal components can be used to define how close the additional test spectrograms are from each spectrogram class using their projections.

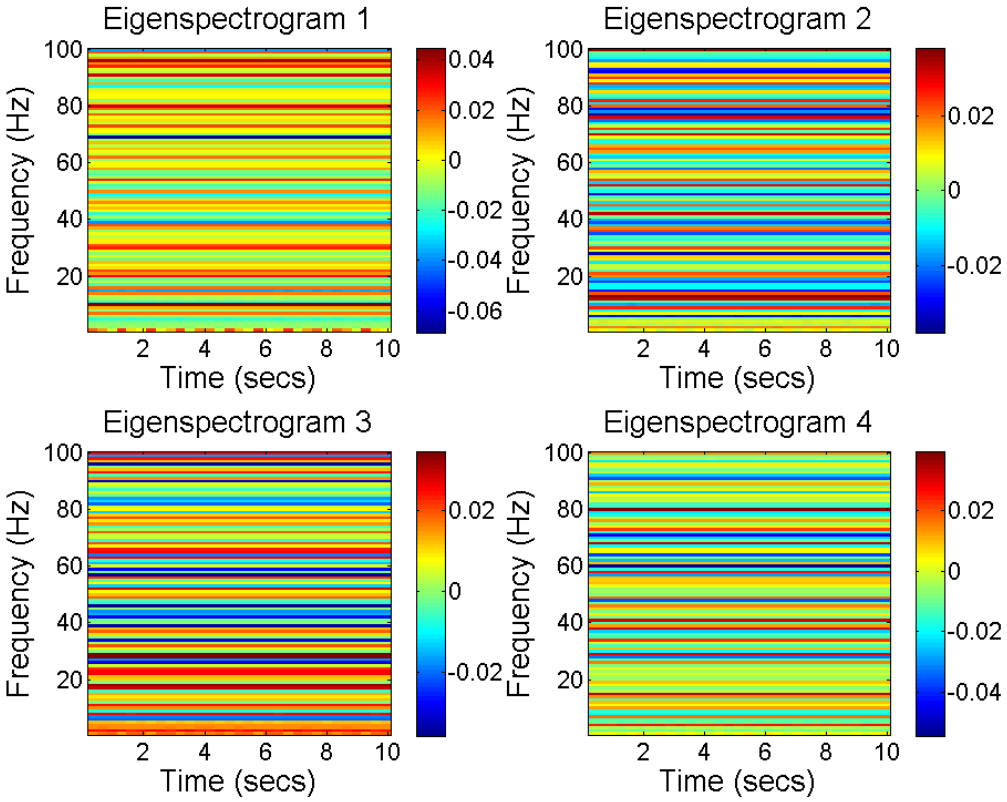


Figure 4.2: Four randomly chosen eigenspectrograms from the eigenspectrogram set that spans the monochromatic sinusoid eigenspace. A complete set of eigenspectrograms is given in C.1.

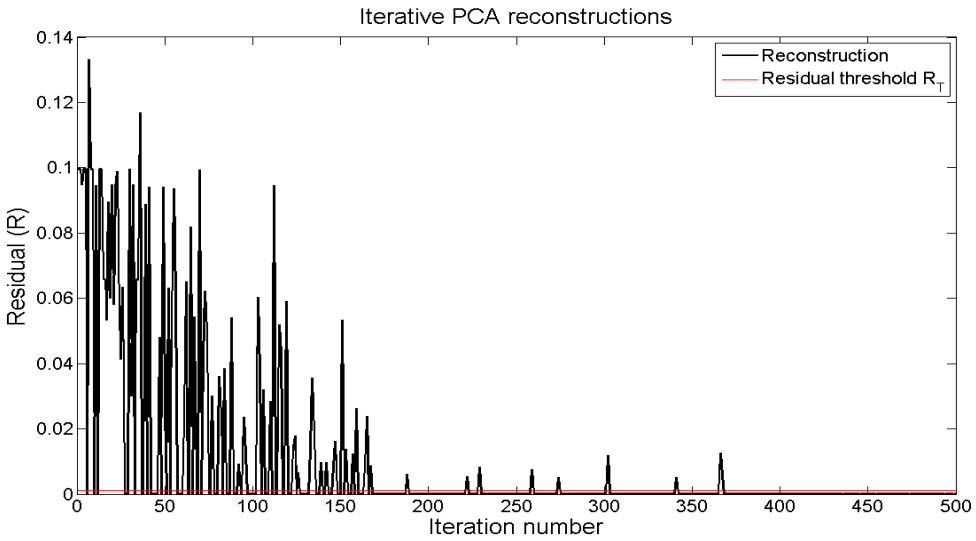


Figure 4.3: The residual measurements of reconstructions during 500 iterations of the PCA method. As unique spectrograms are added, the problem space is better defined, improving the quality of the reconstructions. Eventually, all 100 unique spectrograms are included and the principal components define the problem space completely. Subsequent spectrograms can be reconstructed almost flawlessly.

The four spectrograms **a-d** are displayed in Fig. 4.5, and details regarding their construction are given below. The simplicity of the monochromatic spectrograms permits straightforward definitions of 'near' and 'far', given in Table 4.1. The spectrogram class distance required to define 'far' was determined empirically during some initial tests, while the threshold for residual measurements was more open to interpretation. We decided to demand low residual measurements to be classified 'near', because we were performing the initial tests in situations we had a significant amount of control over and would therefore not present the PCA algorithm that was very different from the training spectrograms (except in those cases where this was specifically the purpose behind the test spectrogram). We had no such prior expectations when we examined EMRI spectrograms later on, using our experience gained by examining the sinusoid spectrograms and an appreciation of the complicated evolution of the waveforms to inform our choices for threshold values of R.

a) Near the eigenspace and near a spectrogram class The test spectrogram is simply created from another monochromatic waveform in the predefined frequency range (specifically, 50 Hz); as expected, the spectrogram can be reconstructed flawlessly (see Fig. 4.6. Further, the frequency of the test spectrogram can be determined by calculating the distance between its weight vector and those of each of the 100 spectrograms used to define the eigenspace (this is demonstrated in Fig. 4.7 and Fig. 4.8). The correct spectrogram will have an identical weight vector, whereas the others will be distinct and hence more distant.

b) Near the eigenspace but far from a spectrogram class Here, the test spectrogram is created from a 50 Hz monochromatic waveform, but the amplitude is doubled. As a result, the contributions from each eigenspectrogram is larger, but again, the spectrogram can be reconstructed extremely well ($R \ll 0.001$, see Fig. 4.9). However, the difference in the weight vectors (Fig. 4.12) means that this spectrogram is very distant from the defined spectrogram classes (Fig. 4.11). Despite this, the spectrogram class closest to the weight vector of the new spectrogram belongs to the unique spectrogram constructed from the 50 Hz waveform.

	R	ϵ^2
Near	≤ 0.001	$\leq 4 \times 10^{12}$
Far	> 0.001	$> 4 \times 10^{12}$

Table 4.1: Distance from the eigenspace and spectrogram classes in the context of monochromatic spectrograms.

c) Far from the eigenspace and near a spectrogram class It is difficult to engineer a spectrogram that satisfies these two conditions. A waveform that changes frequency halfway through (but remains within the defined frequency range) will project into the eigenspace in such a way that its weight vector will not be distant from those of the monochromatic spectrograms, but the quality of the reconstruction is poor ($R \approx 0.71$). The two frequencies in this example are 50 Hz and 65 Hz, but the effect is observed in any similar combination.

d) Far from the eigenspace and far from a spectrogram class Finally, this spectrogram is significantly different from the others, and is therefore is poorly reconstructed (see Fig. 4.16, $R \approx 0.7$) and far from any spectrogram class (see Fig. 4.17 and 4.18). The waveform used to generate the spectrogram is calculated from the a vector of random amplitudes $(1 - 20)$ and frequencies (within the defined frequency range), resulting in the chaotic distribution of power throughout its duration.

What does the PCA reveal?

Firstly, the majority of the eigenvectors appear as we would expect; linear combinations of the monochromatic spectrograms, with different contributions from each. It is difficult to extract much information simply by looking at them, but it can be seen that the power at each frequency remains constant over time. Similarly, the weight vectors of the unique spectrograms do not provide much visual information (see Fig. 4.4), but we can see that each eigenspectrogram contributes to the makeup of each spectrogram to some extent, again as expected. Although the individual eigenspectrograms (and hence weight vectors) will be different, repeated trials of this situation will provide essentially the same results (see Fig. 4.19); the order in which the spectrograms are used to expand the eigenspace does not alter the effectiveness of the end result.

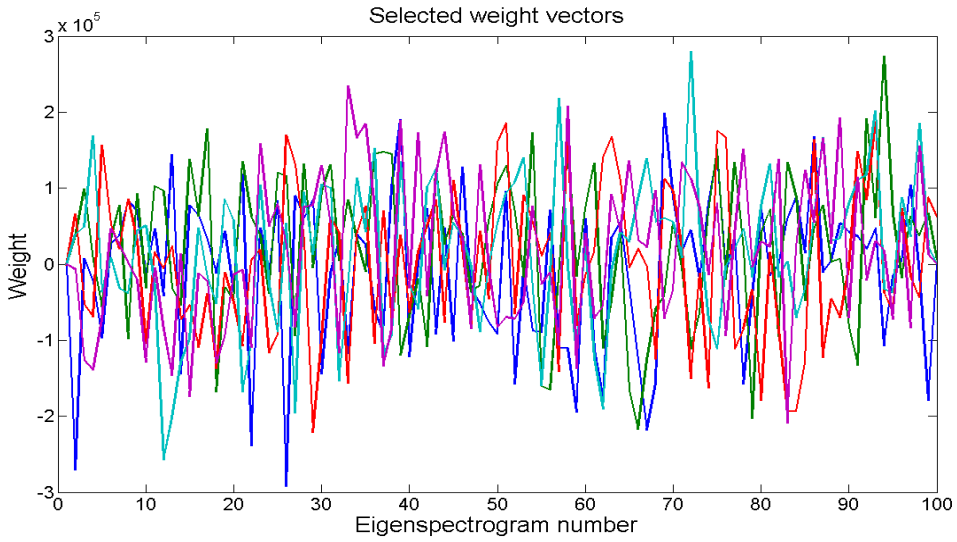


Figure 4.4: Weight vectors for 5 randomly selected spectrograms according to the complete set of 100 eigenspectrograms. Displayed in this manner, it is difficult to extract useful information, but they define the spectrogram classes and the typical distances between them.

Even so, when sorted according to increasing eigenvalue (and hence by the amount of variance in the dataset they account for) the first and last eigenspectrograms look closer to unaltered spectrograms than a combination of 100 different frequency components. The precise method by which Matlab determines the eigenvectors and eigenvalues is dependent on the LAPACK algorithm library (see (W..h) and appendix B for details), but the power in the unusual eigenspectrograms is centred on the frequency of the final spectrogram added during the iterative process to define the eigenspace. Further, the weight vectors of the spectrograms according to the eigenspectrograms calculated iteratively show that there is minimal contribution from the first and last eigenspectrograms, except in those whose frequency matches, or is close to, the final spectrogram added.

During every iteration that incorporates a new spectrogram to expand the eigenspace, we are determining the eigenvectors of a set consisting of several orthogonal vectors plus one non-orthogonal vector, rather than from a set of non-orthogonal vectors. This has the effect of scrambling the output produced somewhat; creating an eigenspace from a complete set of 100 spectrograms non-iteratively (that is, starting with a training set of the 100 unique spectrograms) creates eigenvectors that are

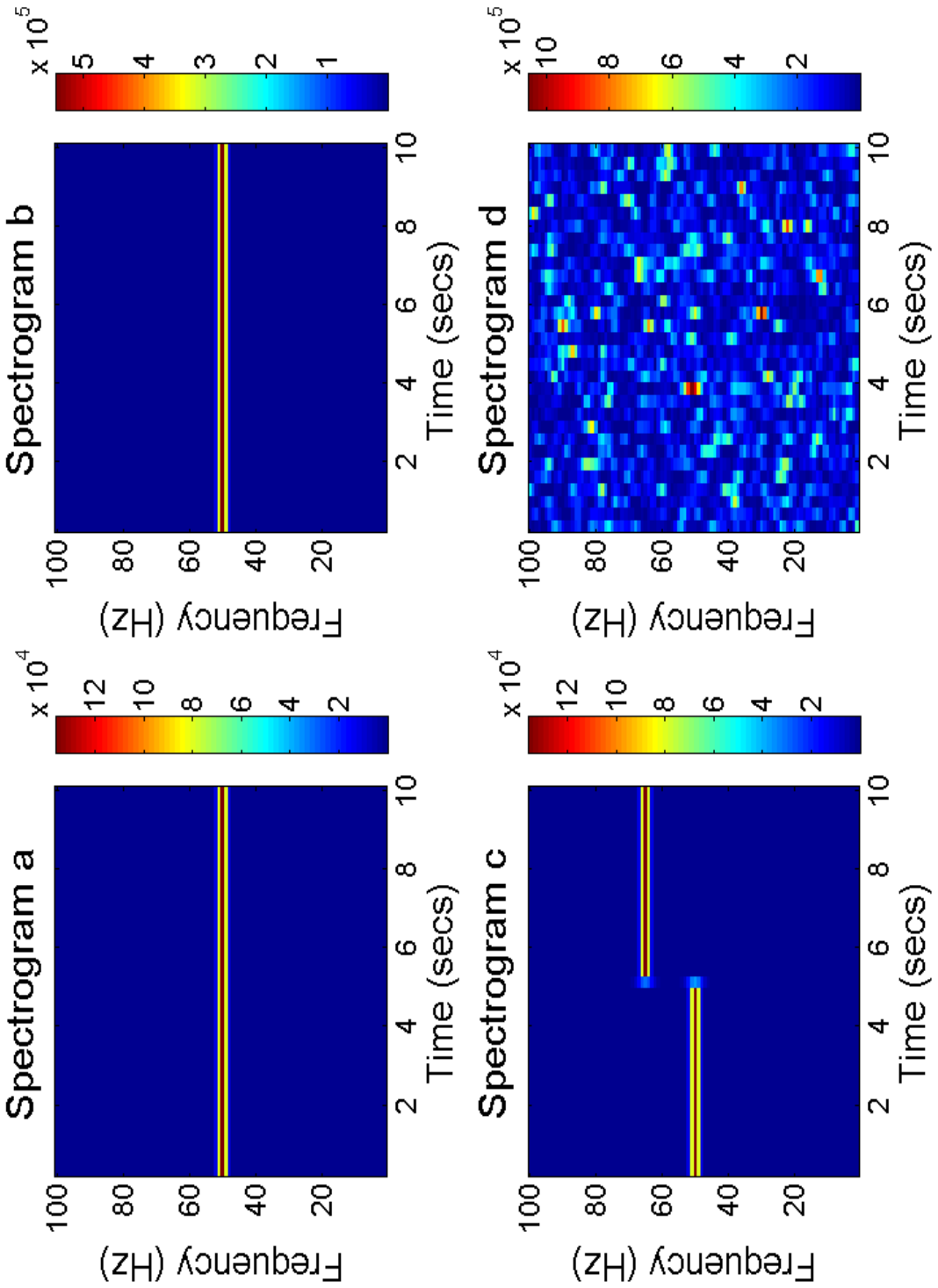


Figure 4.5: Four spectrograms **a-d** representing different categories. Each one provides a different insight into the behaviour of the PCA method.

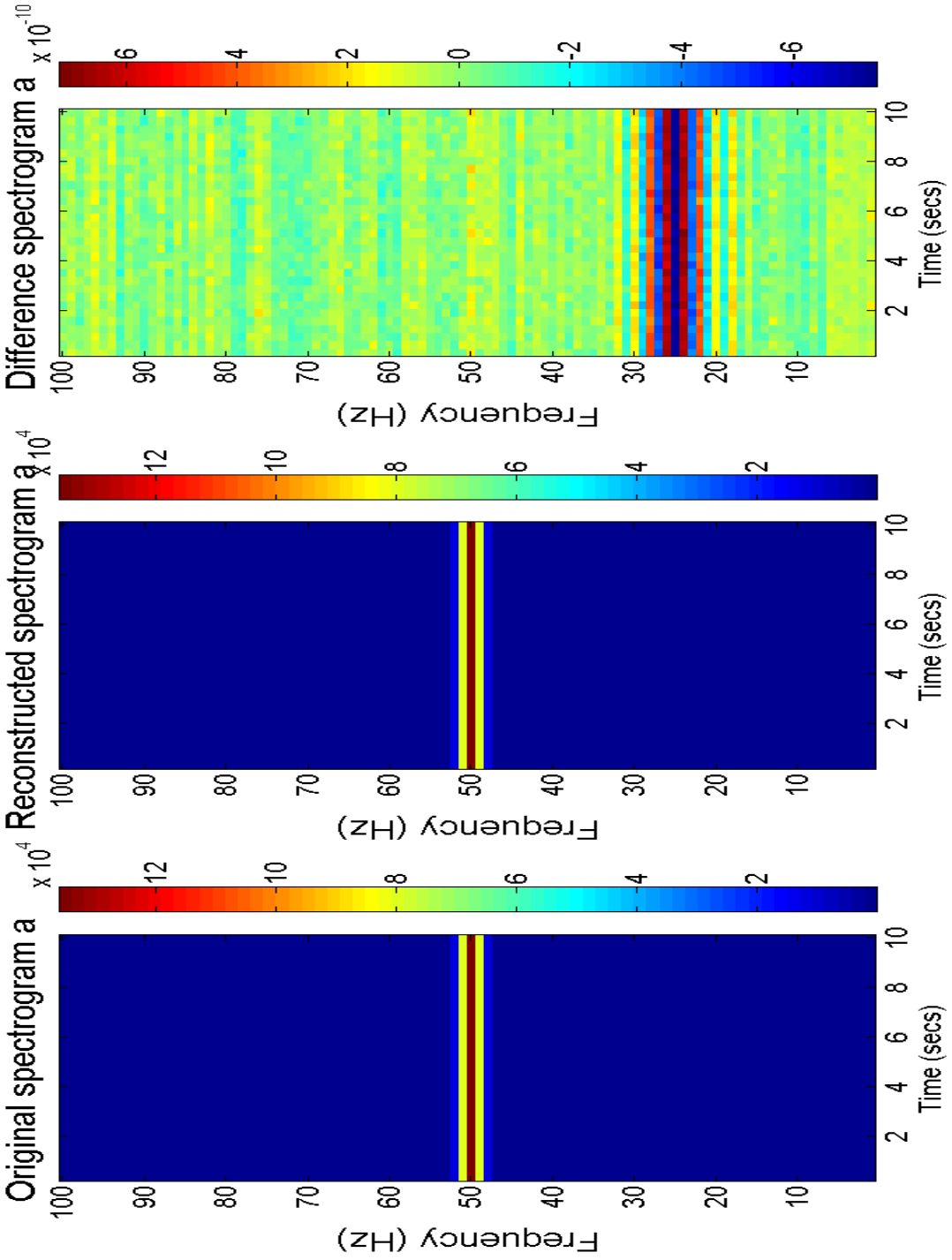


Figure 4.6: Spectrogram **a** and its reconstruction with a complete set of eigenspectrograms. The 50 Hz spectrogram belongs to the set used to construct the eigenspace and can therefore be reconstructed almost perfectly. A spectrogram showing the difference in the power between the original and the reconstruction is also included.

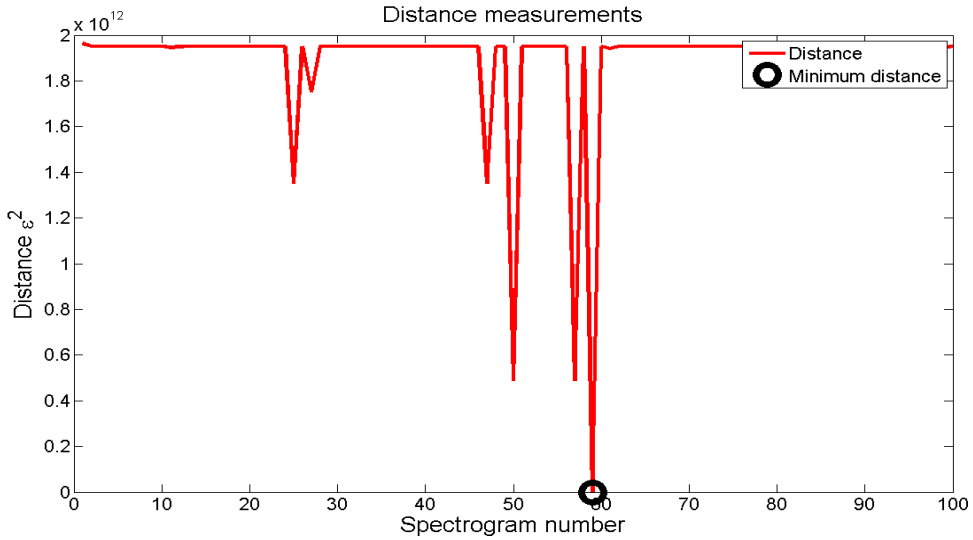


Figure 4.7: The distance between test spectrogram **a** (see Fig. 4.5) and the 100 unique spectrograms, according to their weight vectors. The spectrograms are ordered according to their inclusion in the set used to create the eigenspace, and the minimum distance is highlighted.

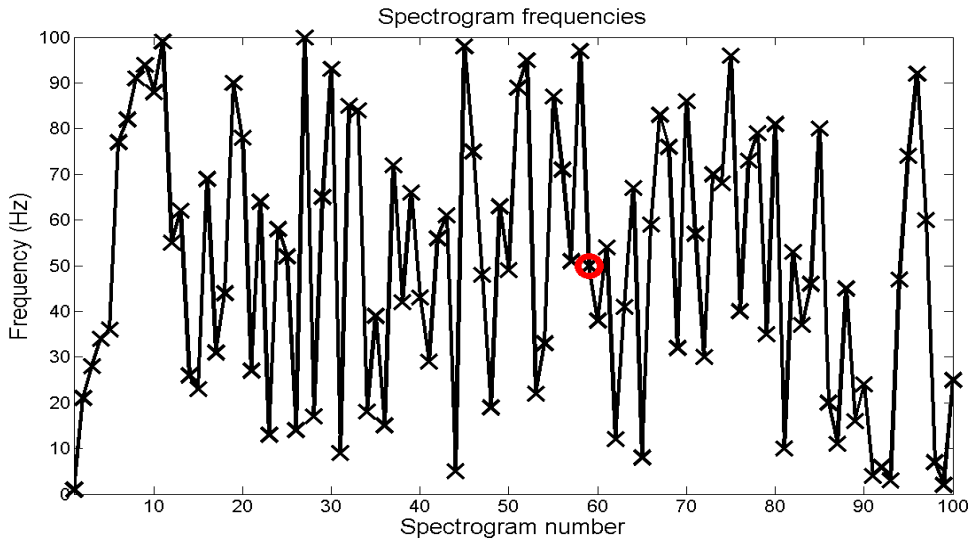


Figure 4.8: The frequencies of the unique monochromatic spectrograms, according to the order of their generation. The most likely frequency of the test spectrogram **a** (see Fig. 4.5), as determined by the minimum distance calculated from weight vectors is highlighted. The two match exactly (50 Hz).

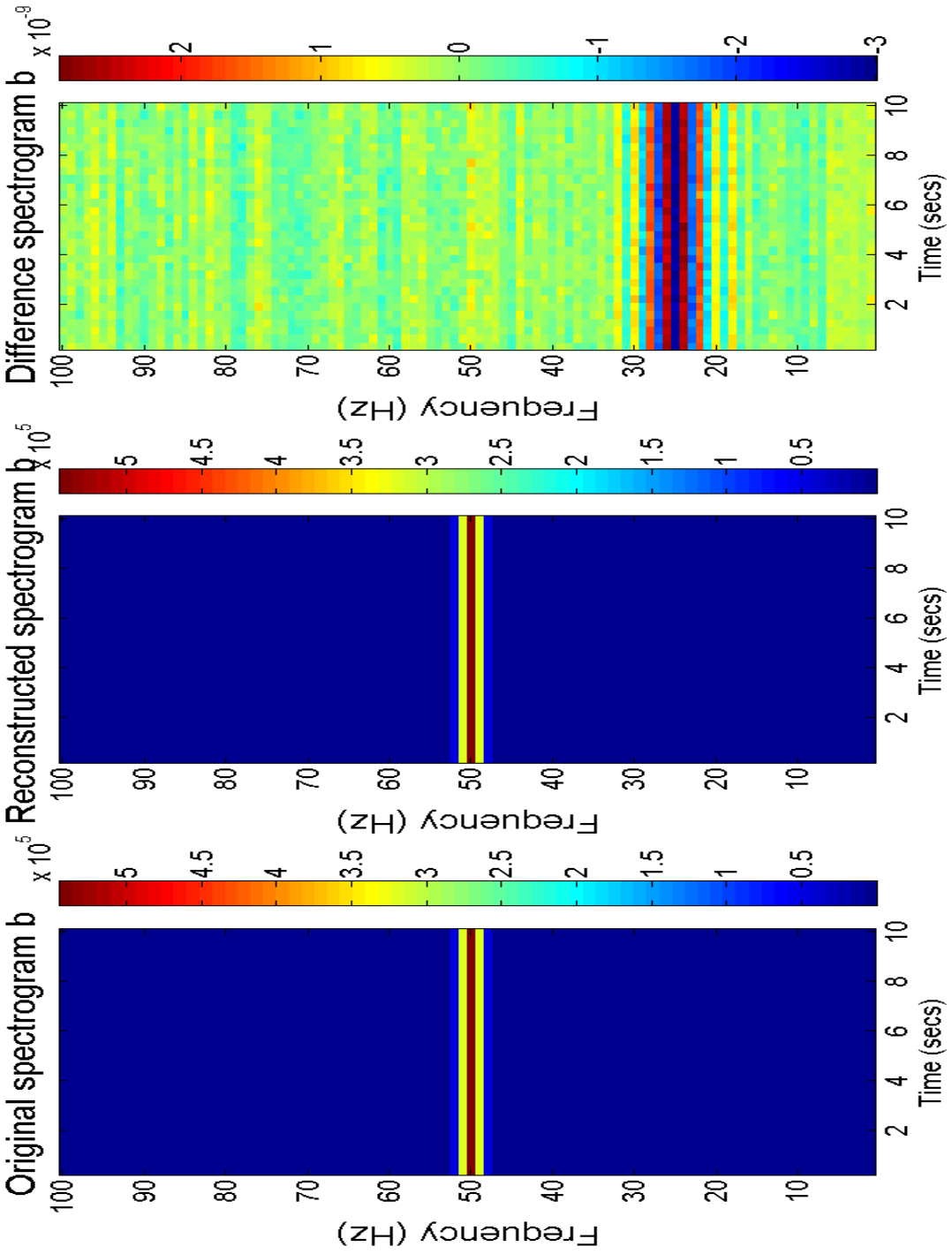


Figure 4.9: Spectrogram **b** and its reconstruction with a complete set of eigenspectrograms. The 50 Hz spectrogram does not belong to the set used to construct the eigenspace, but can still be reconstructed almost perfectly. A spectrogram showing the difference in the power between the original and the reconstruction is also included.

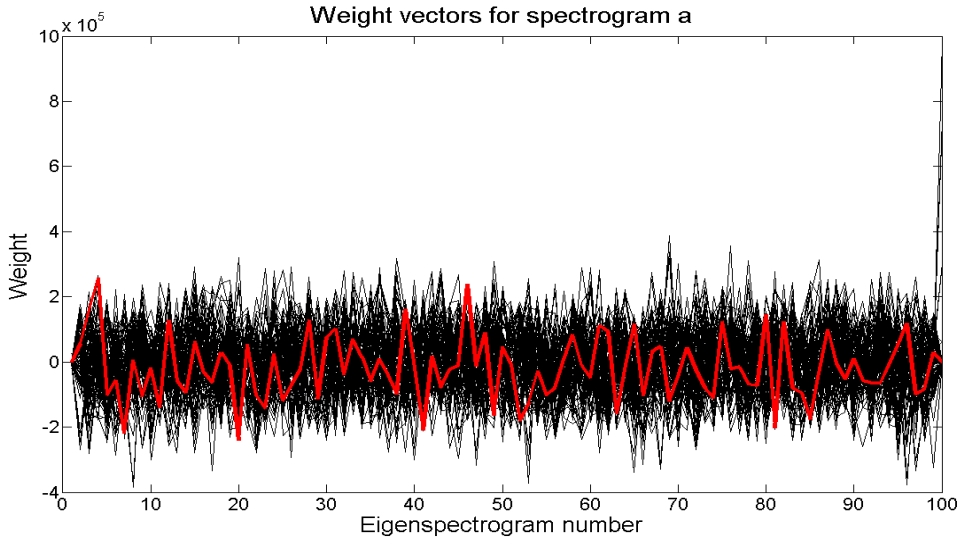


Figure 4.10: The weight vectors for the 100 unique monochromatic spectrograms (in black) as well as the weight vector of spectrogram **a** according to the same set of eigenspectrograms (in red), which matches the weight vector of the 50 Hz spectrogram.

rather different looking. There is no cause for concern though, since the same small errors in the reconstruction of a type **a** spectrogram are present in both situations. It appears to be impossible to eliminate this source of error completely.

In addition, a spectrogram added iteratively is nearly orthogonal to the set of existing eigenspectrograms before expansion, even though it may be radically different in appearance. This allows it to be incorporated into the set of eigenspectrograms almost unaltered, with the smallest-eigenvalue eigenspectrogram being a 'correction' vector that accommodates the orthogonalisation algorithm's inability to make completely error-free orthogonal eigenvectors. The precise relation between the power in the smallest-eigenvalue eigenspectrogram and the overlap between adjacent-integer-frequency spectrograms is unknown as yet, but the error in the resdiaul of a spectrogram of type **a** from a complete set of eigenspectrograms is of the same order of magnitude as a rounding error ($R \approx 10^{-14}$). In addition, most of the power contribution to the error is centred at the frequency of the final spectrogram added (see Fig. 4.6).

If we attempt to reconstruct spectrograms using a subset of the eigenspectro-

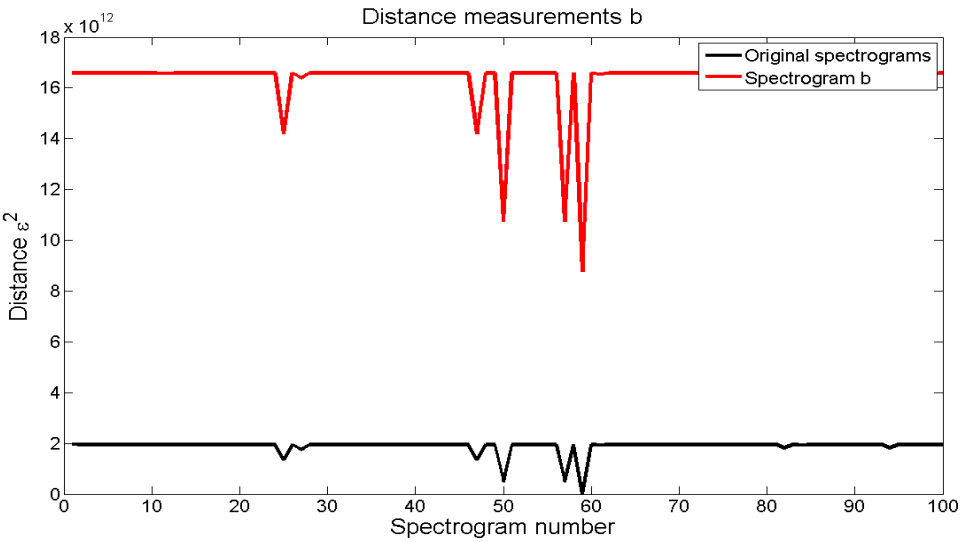


Figure 4.11: The distance between test spectrogram **b** (see Fig. 4.5) and the 100 unique spectrograms, according to their weight vectors (in red). Also included is the distance between the 50 Hz spectrogram and each of the 100 unique spectrograms (in black), giving an indication of how far spectrogram **b** is from the spectrogram classes compared to the typical distance between them.

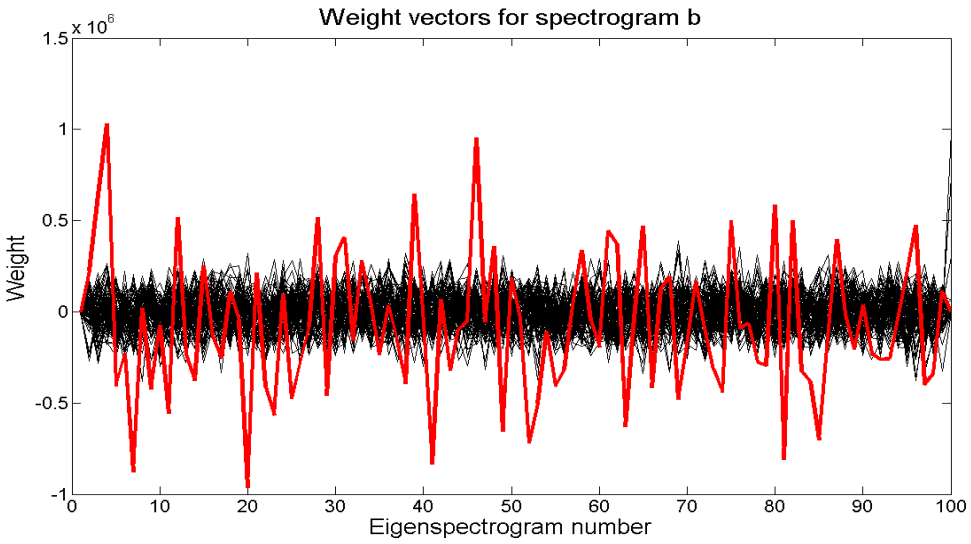


Figure 4.12: The weight vectors for the 100 unique monochromatic spectrograms (in black) as well as the weight vector of spectrogram **b** according to the same set of eigenspectrograms (in red). It shows that the magnitude of the contributions for spectrogram **b** is generally much larger than those of the original spectrograms.

grams, we can see that all 100 are always required for a perfect reconstruction. While this shows that the space is spanned completely by the eigenspectrograms, it also confirms that there is no redundancy in the problem space. Even if we wanted to reconstruct each of the spectrograms such that the residual measurements are less than $R = 0.5$, around 90 eigenspectrograms are required (most can be reconstructed with fewer, but in order to guarantee success we take the maximum number, see Fig. 4.20).

Despite its robust way of constructing the set of eigenspectrograms, the PCA method has taken some effort (and computational power) to do so. Here though, we are reaping the benefits of prior knowledge of the problem space. We know that it will require 100 principal components to span the problem space because each observation of the problem space is so distinct. In the monochromatic spectrogram space, the same effect can be achieved by creating a training set of the 100 unique spectrograms and running the PCA method once. It would also be possible to further simplify the generation by creating 'spectrograms' that were with all of the power contained at the desired frequency: completely lacking any overlap, these

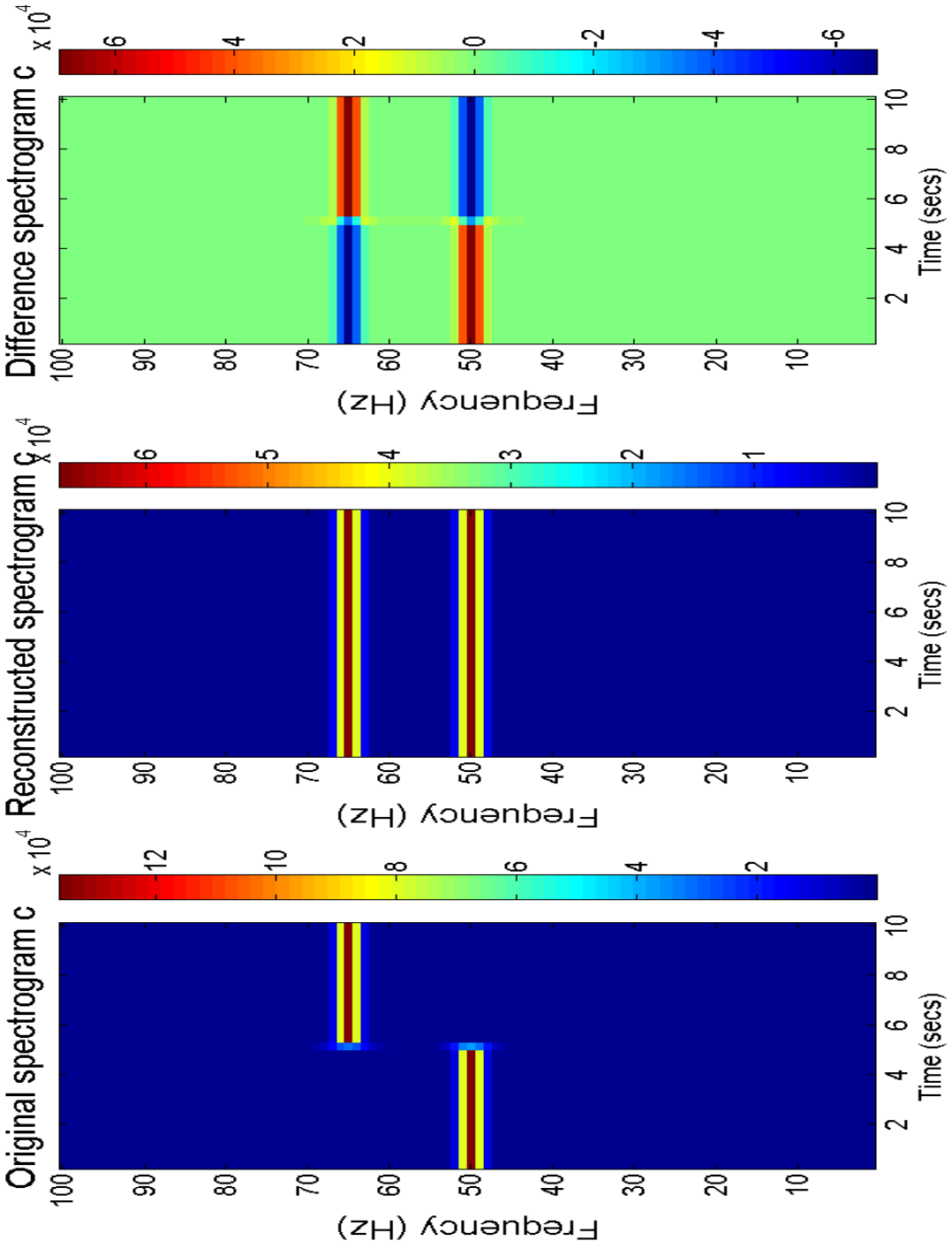


Figure 4.13: Spectrogram **c** and its reconstruction with a complete set of eigenspectrograms. The reconstruction is poor because the original spectrogram is unlike those that define the eigenspace, but it contains components of two of these (a 50 Hz and 65 Hz spectrograms) and so projects into the eigenspace close to defined spectrogram classes. A spectrogram showing the difference in the power between the original and the reconstruction is also included.

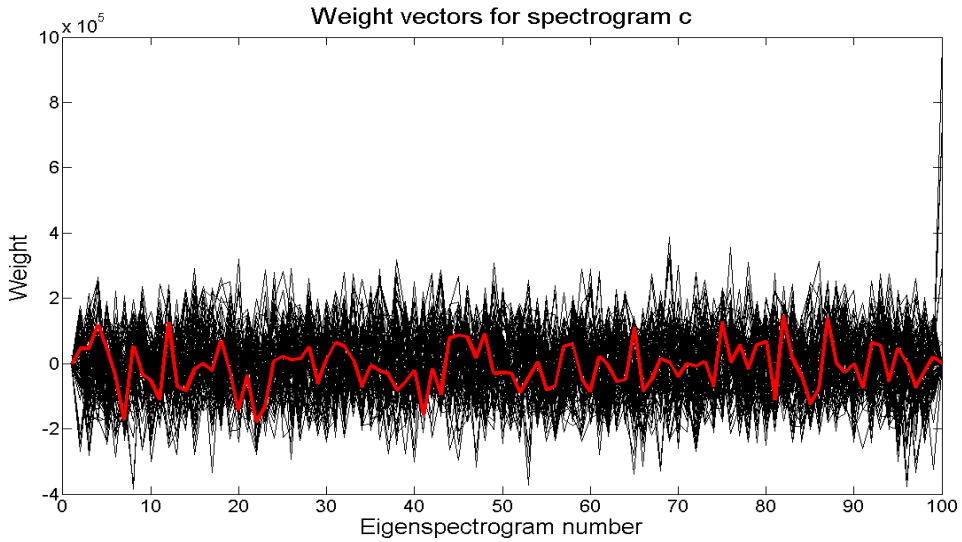


Figure 4.14: The weight vectors for the 100 unique monochromatic spectrograms (in black) as well as the weight vector of spectrogram **c** according to the same set of eigenspectrograms (in red). The magnitude of the contributions for spectrogram **c** is of the same magnitude as those of the original spectrograms, but does not match any of them.

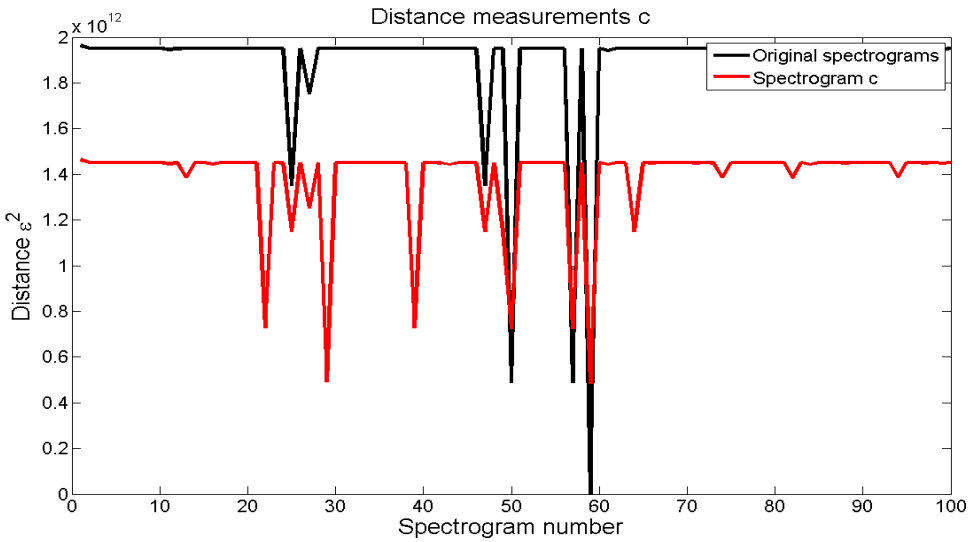


Figure 4.15: The distance between test spectrogram **c** (see Fig. 4.5) and the 100 unique spectrograms, according to their weight vectors (in red). Also included (in black) is the distance between the 50 Hz spectrogram and each of the 100 unique spectrograms, giving an indication of how far spectrogram **c** is from the spectrogram classes compared to the typical distance between them.

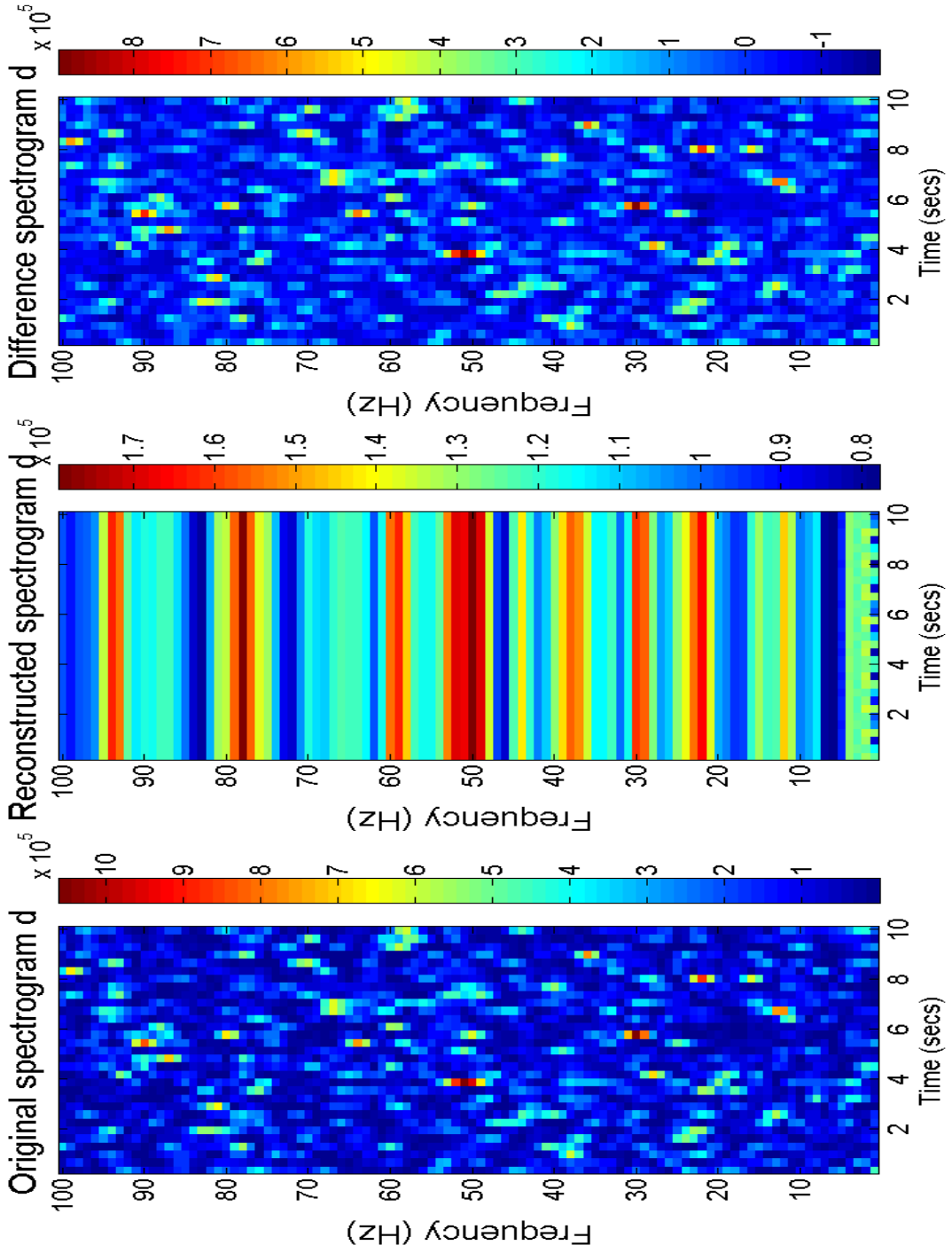


Figure 4.16: Spectrogram **d** and its reconstruction with a complete set of eigen-spectrograms. The reconstruction is poor because the original spectrogram is unlike those that define the eigenspace because of its significantly distribution of power and so projects into the eigenspace in a very different manner from defined spectrogram classes. A spectrogram showing the difference in the power between the original and the reconstruction is also included.

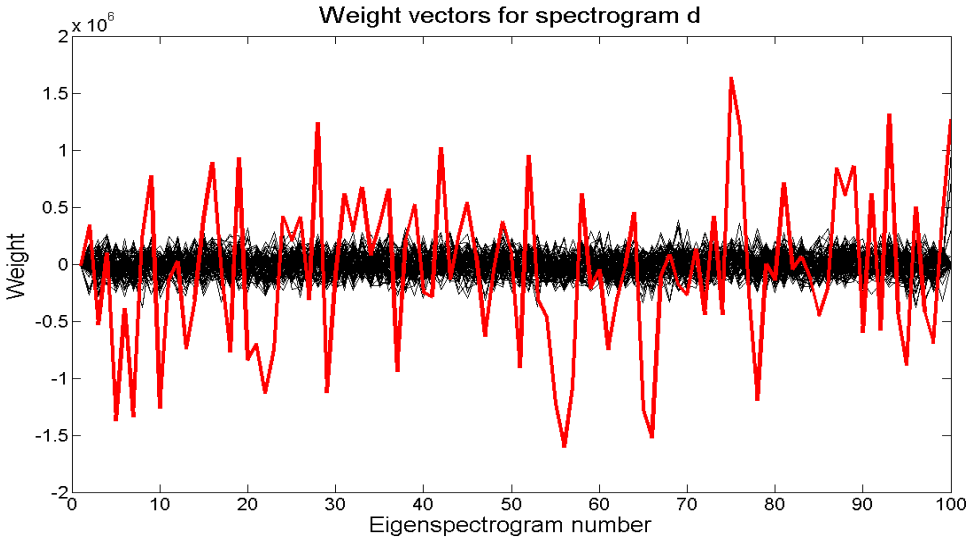


Figure 4.17: The weight vectors for the 100 unique monochromatic spectrograms (in black) as well as the weight vector of spectrogram **d** according to the same set of eigenspectrograms (in red). The magnitude of the contributions for spectrogram **d** is larger than those of the original spectrograms, and does not match any of them.

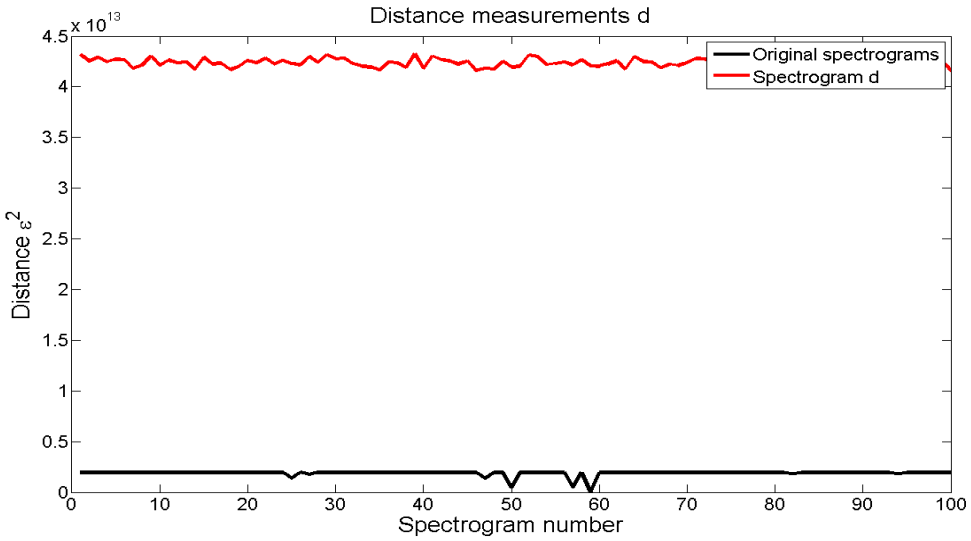


Figure 4.18: The distance between test spectrogram **d** (see Fig. 4.5 and the 100 unique spectrograms, according to their weight vectors (in red). Also included is the distance between the 50 Hz spectrogram and each of the 100 unique spectrograms, giving an indication of how far spectrogram **b** is from the spectrogram classes.

'spectrograms' would each be an eigenvector, and their weight vectors would only contain a non-zero contribution from themselves.

Our knowledge of the problem space also helps to set the residual threshold required to include new spectrograms particularly low, since two different spectrograms would contain distinct information. Likewise, having each spectrogram represent a distinct spectrogram class. On the other hand, the typical distance between the spectrogram classes was established empirically after a successful generation of the eigenspectrograms. In this situation then, they are not used to influence the construction of the eigenspectrogram set, and the manner by which we determine the qualities 'near' and 'far' are applied retrospectively.

What can we learn from the four spectrograms?

The characteristics of the spectrograms that cause them to be placed in distinct categories reveal information about the eigenspace that has been constructed and the behaviour of the iterative PCA method itself. This information will suggest what may be expected when a similar approach is applied to more complicated spectrogram and different parameter spaces, mindful of the intended subject of the analysis, the EMRI spectrograms.

Spectrogram **a** confirms that the eigenspectrograms generated do indeed span the required space by showing their ability to recreate any of the original spectrograms. This is to be expected of course, since we are only rearranging the information supplied by the spectrograms, rather than reducing it (again, this is due to the lack of redundancy in the problem space). Though it is not particularly illuminating provided the PCA is operating properly, testing using **a**-type spectrograms is a simple method of checking for errors in the PCA processes. With regards to EMRI spectrograms, the situation is the same; we should expect that if the iterative PCA method is supplied an EMRI spectrogram it has been shown previously (and determined to be sufficiently different to be worth including in the dataset used to construct the eigenspectrograms), it will not be included a second time because it contains no new information.

The test spectrogram **b** is more interesting, since it reveals that the residual is not sensitive to the amount of power in the spectrogram, only the distribution. The weight vector showing the contribution of each eigenspectrogram has the same shape as that of spectrogram **a** (compare the weight vectors of Fig. 4.10 and Fig. 4.12), but

scaled up, since the spectrogram is fundamentally a rescaled version of spectrogram **a** (in terms of the power in each time and frequency bin). Again, this is expected, since the eigenvectors form a basis set for the eigenspace and any point that is a scalar multiple of a point in the eigenspace is also in the eigenspace. This may be particularly significant when analysing EMRI spectrograms because the signal strength of the gravitational waveforms will depend on the source distance. This is the first indication that there is redundancy in the EMRI problem space that can be exploited by principal component analysis.

Spectrogram **c** has two separate components, each belonging to the eigenspace (or would, if they continued for the entire duration of the spectrogram). In combination, however, they create a spectrogram quite unlike the set used to construct that eigenspace. Despite this, they project into the eigenspace with a weight vector that is close to others in the set, making it impossible to reject it based on spectrogram class distances alone. Thankfully, the high residual measurement from this reconstruction attempt using the eigenvectors would allow it to be rejected by a search based on this measurement, but the presence of multiple signals may result in similar situations. The focus of the technique thus far has been on isolated signals (multiple and overlapping signals are discussed in 6.3), but we must be careful to distinguish between detecting no signals and detecting too many. Further, this situation confirms that the continuity of the data in the spectrogram is an important feature; a loss of the signal at specific frequencies or for short periods of time will create erroneous structure in the spectrogram that will cause it to seem like a distinct entity.

Finally, spectrogram **d** shows that the PCA method will not be able to reconstruct every spectrogram it is presented with, and that such spectrograms project into the eigenspace in a dissimilar manner to those that belong. Such behaviour was anticipated, but this test was a useful confirmation.

Governing each classification was our definition of 'near' and 'far'. As stated previously, when we controlled the process of generating the training and test spectrograms in these simple situations, it seemed appropriate to determine the spectrogram classes definition empirically and settle on a low residual measurement to deem a reconstruction 'good' quality. Nevertheless, later tests (see Fig. 4.24 and Fig. 4.25) showed that it was possible to reconstruct a test spectrogram in a manner that seemed extremely similar to its original form visually, but was deemed to have too high a residual to be considered so by the PCA algorithm. Therefore,

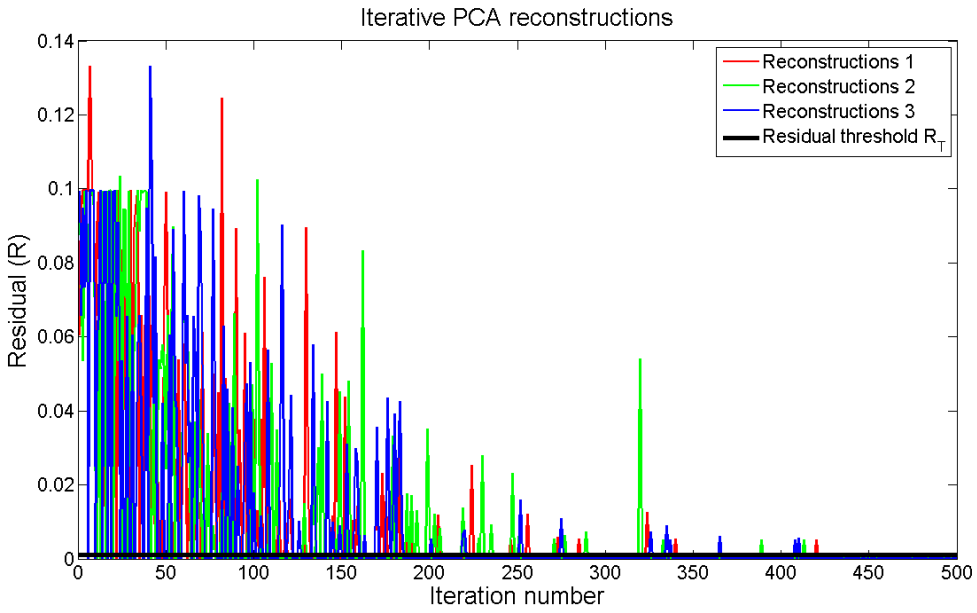


Figure 4.19: Multiple application of the PCA method to monochromatic spectrograms using random generation of the frequencies to be included. After 500 iterations, the eigenspace has been completely mapped in each run, and the behaviour of the set of eigenspectrograms is consistent despite the eigenspectrograms that make up these sets being different.

we could not be confident that our thresholds were particularly wise choices. At this stage however, it seemed they were sufficient to show the four categories in a straightforward manner.

4.2.1.2 Simple polychromatic spectrograms

Having established that the PCA can be used to define the eigenspace of a very simple situation, the next step should be to complicate matters and see how the PCA performs. EMRI spectrograms are clearly not monochromatic, so the logical next step is to perform PCA on polychromatic spectrograms of some description.

This time, spectrograms are constructed from sinusoids with randomly chosen frequencies in the same manner as the monochromatic spectrograms. However, the sinusoids are now given multiple frequency components under a number of different circumstances.

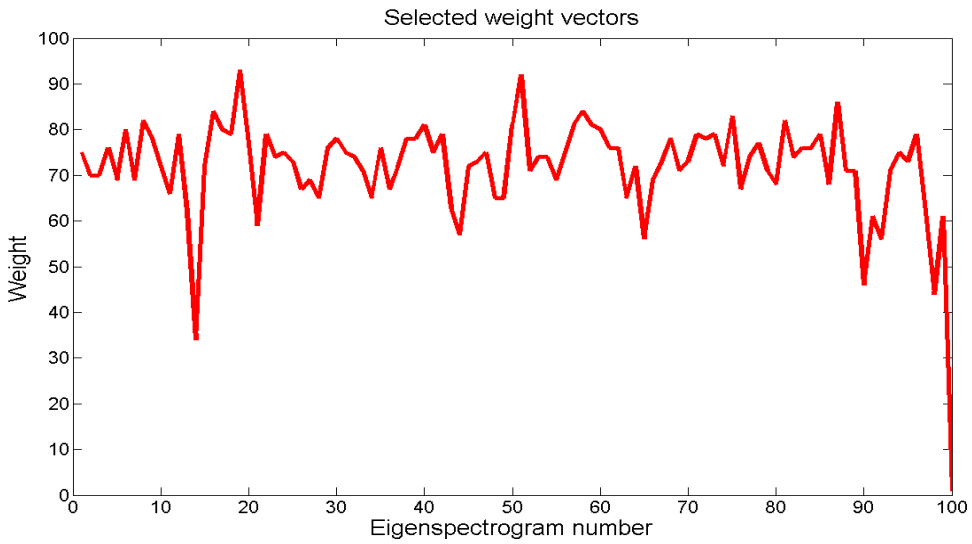


Figure 4.20: The number of eigenspectrograms required to reconstruct each of the 100 unique spectrograms such that the residual $R < 0.5$. As discussed, the particular spectrogram considered is irrelevant since the order of their inclusion is random, and are numbered sequentially here for simplicity only.

Monochromatic training spectrograms, 2-frequency-component spectrograms

We first attempt to span the problem space defined by spectrograms constructed from sinusoids with two frequency components (see Fig. 4.21). There are $\binom{100}{2} + 100 = 5050$ such constructions possible, so the problem space appears to be much larger than before, and the residual threshold has been set at $R_T = 0.01$ to reflect what we expect to be the increased difficulty of completely defining this space; we believed that the space would be hard to span, but didn't want every test spectrogram to be used to expand the eigenspace (we wanted some test spectrograms to be deemed 'near' to the parameter space, so made the residual threshold higher than before in an attempt to produce this behaviour from the algorithm).

In an attempt to use our knowledge of the previous problem space to aid us, the training set eigenspectrograms initially consists of those used to span the monochromatic spectrogram problem space. One might assume that the reconstruction of a polychromatic spectrogram can be achieved by simply using the sum of the weight vectors corresponding to each frequency component, but the inter-modulation of

the two frequency components in the sinusoids results spectrograms that do not project into the eigenspace in this way. Spectrograms with widely-separated frequency components do not suffer much from this effect and can be reconstructed from the original eigenspectrogram rather well, but in those with a small separation in the frequencies the effect is more pronounced [1].

The original eigenvector set is therefore insufficient to span the problem space, and the eigenspace it spans must be extended. The residual measurements over a large number of iterations of the PCA method (Fig. 4.22) shows that the actual eigenspace required in this situation is *much* larger than that of the monochromatic spectrograms, and needs a large number of eigenvectors to span it. Indeed, after 5000 iterations, 400 eigenvectors had already been defined. At this stage the process was halted, since the purpose was principally to understand the effect of multiple frequency components of the PCA method rather to provide a complete solution to each solution.

With an eigenspace that does not completely span the problem space, the ability of the eigenvectors to reconstruct new spectrograms is extremely varied. We are not throwing away any information (simply rearranging it), so any new spectrogram similar to one already incorporated can be reconstructed well, but those that are not are reconstructed poorly. The main disadvantage is that the behaviour is inconsistent, with the residual values for different test spectrograms varying by orders of magnitude. Fig. 4.24 shows an extremely good quality reconstruction ($R \approx 1.7 \times 10^{-14}$) of a spectrogram with two randomly chosen frequency components, in stark contrast to a second reconstruction from the same set of eigenspectrograms displayed by Fig. 4.25 (here, $R \approx 0.02$). It is this inconsistency that make it impossible to be confident about the usefulness of the output produced if the test spectrograms are going to be selected randomly.

Ignorant training spectrograms

The previous attempt is repeated, but starting with ten training spectrograms generated in the same manner as the spectrograms that will be assessed (and incorporated into the eigenspectrogram set if necessary). This time, the expectation is not that

[1] To check that the inter-modulations were responsible, test spectrograms were created by simply summing two monochromatic spectrograms together. This pseudo-spectrogram could be reconstructed perfectly as expected, with a weight vector equal to the sum of the weight vectors of the two monochromatic spectrograms.

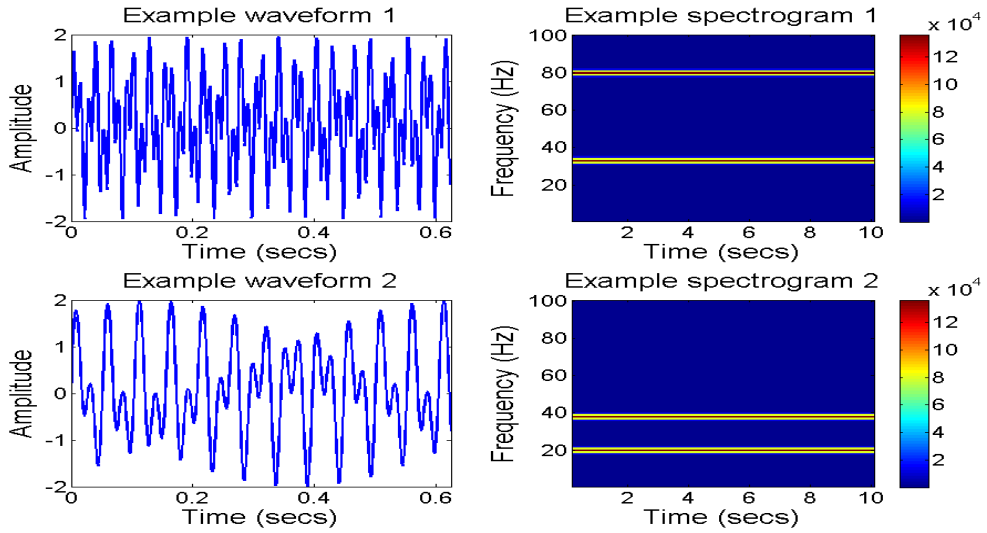


Figure 4.21: The timeseries of two polychromatic sinusoids and the corresponding spectrogram. Only the first 500 data points of the waveforms are displayed.

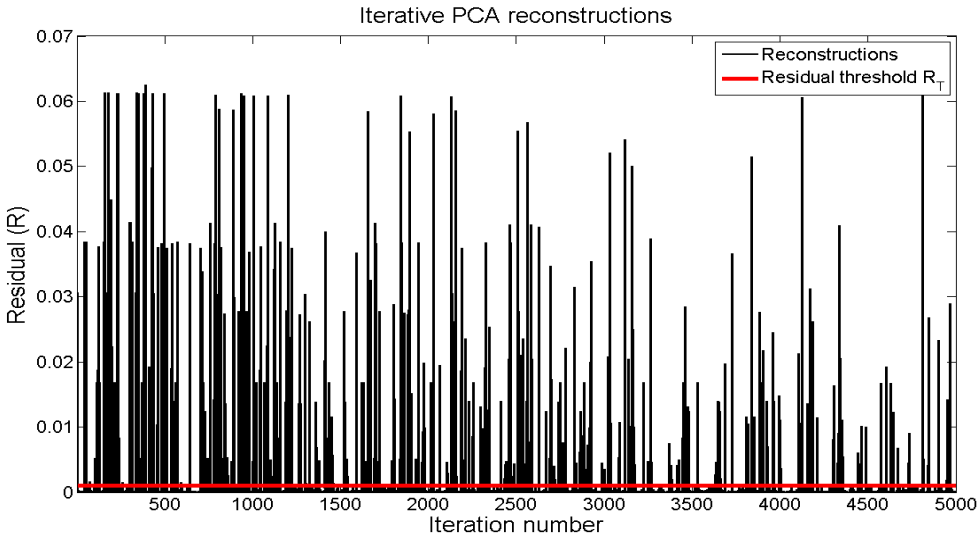


Figure 4.22: The residual measurements of reconstructions during 5000 iterations of the PCA method. The new spectrograms being assessed are polychromatic, but the training set of eigenspectrograms are those that completely defined the eigenspace of the monochromatic spectrograms.

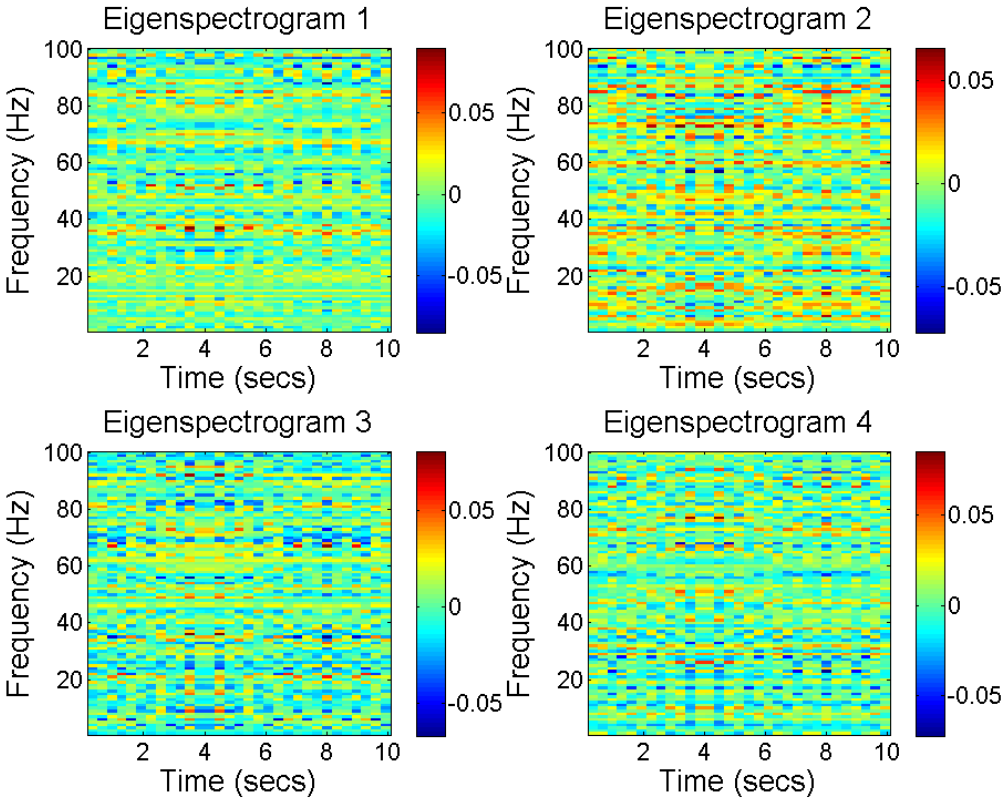


Figure 4.23: Four randomly chosen eigenspectrograms from a set constructed using monochromatic training spectrograms and further polychromatic spectrograms. It can be seen that now the structure is not time-independent.

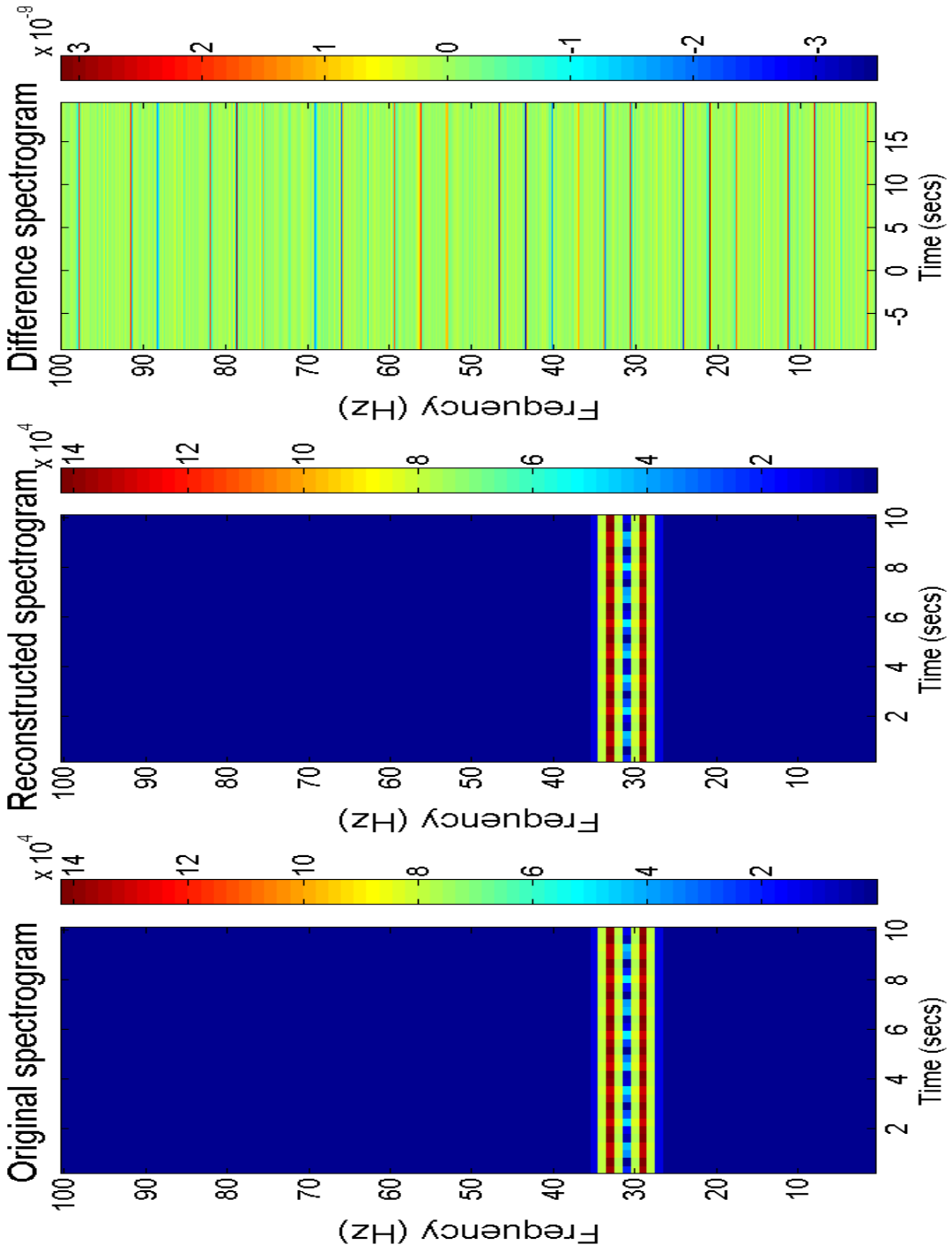


Figure 4.24: Polychromatic spectrogram and its reconstruction using 400 eigenspectrograms. The third spectrogram, showing the difference between the two, demonstrates that the residual measurement is very small $R \approx 1.7 \times 10^{-14}$.

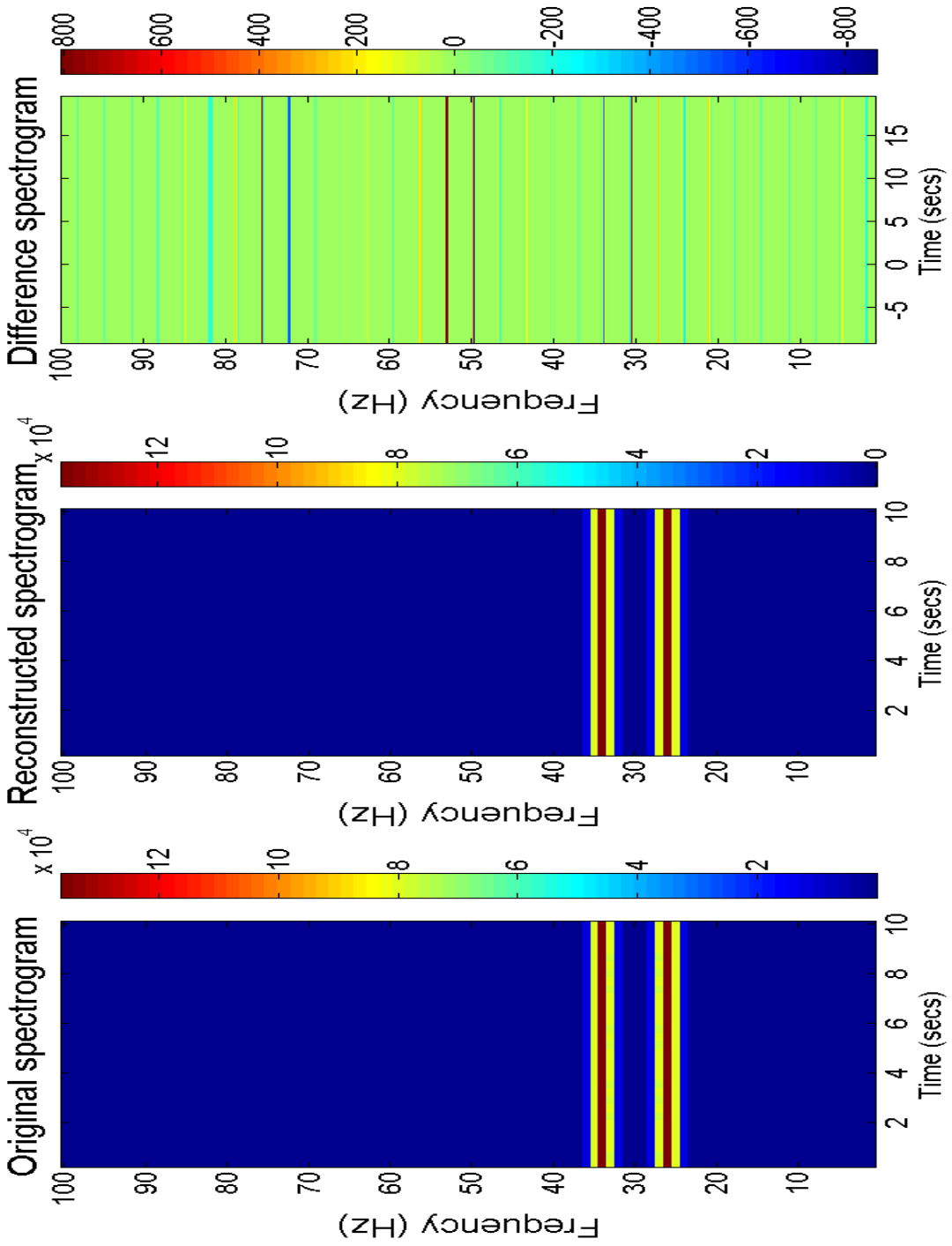


Figure 4.25: Polychromatic spectrogram and its reconstruction using 400 eigenspectrograms. The third spectrogram, showing the difference between the two, reveals that the reconstruction is of fairly poor quality, with a residual measurement of $R \approx 0.02$, although the main features are retained.

the eigenspace can be spanned (the problem space is identical), but simply that the random generation method for both the training spectrograms and the tested spectrograms will produce a similar output from the iterative PCA method as before. This will show that prior knowledge about the structure of the problem space has limited impact after a large number of iterations.

After 5000 iterations, it initially appears that this is indeed the case. 420 eigenspectrograms have been created at this point (see Fig. 4.27) with little sign that the problem space has been well spanned (see Fig. 4.26). However, in general new test spectrograms cannot be reconstructed as well as from the previous eigenspectrogram set that began with the eigenspectrograms that spanned the monochromatic spectrogram problem space.

This in not surprising, since the almost complete orthogonality of two widely separated frequency components meant that by starting with a set of eigenspectrograms that could reconstruct each frequency component individually, spectrograms of this type were no obstacle. The only unreconstructable spectrograms in this situation were those that featured two frequency components that were close in value. Starting with an eigenspectrogram set that was created randomly, on the other hand, has really only helped to remove these from the set of unreconstructable spectrograms, leaving a far larger set that still contains information that must be included.

Despite this, the residual measurements of two test spectrograms may still differ by orders of magnitude (see Fig. 4.28 and Fig. 4.29). Therefore we still face the prospect of inconsistency.

Twenty frequency components

This time, each sinusoid has 20 frequency components with integer-value frequencies in the range $1-100$ Hz and constant amplitude, and the corresponding spectrogram is created. The training set consists of ten such spectrograms, and further spectrograms are generated in an identical manner. This time, the drop off is particularly rapid, and within approximately 600 iterations, the reconstructions produce residuals that are consistently below the residual threshold used to exclude them from the set of spectrograms that to create the eigenspace ($R_T = 0.001$ once again).

The difference between this and previous circumstances is not entirely expected, since the previous polychromatic spectrograms suggested that the problem space is very large. However, the nature of the principal component analysis is to split the

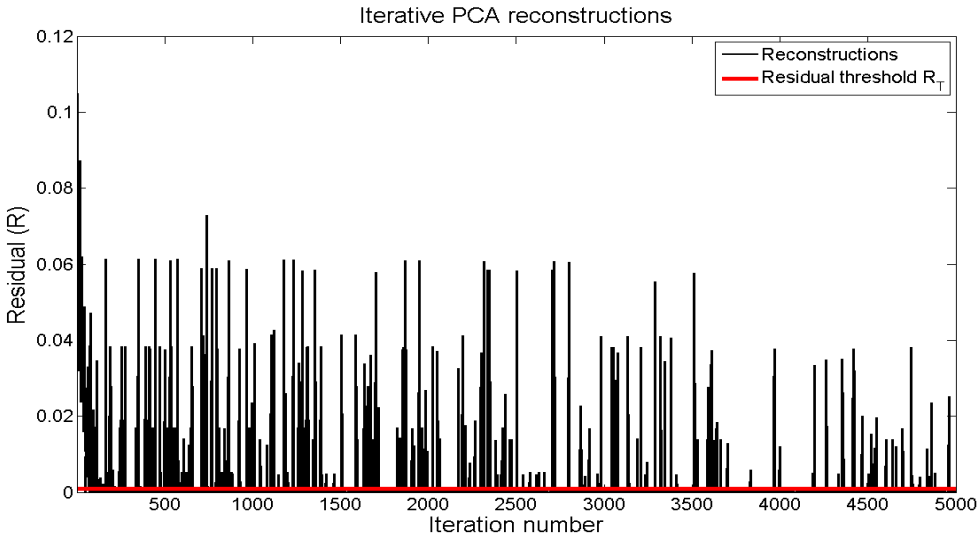


Figure 4.26: The residual measurements of reconstructions during 5000 iterations of the PCA method. New spectrograms being assessed are polychromatic waves with two frequency components, as are the training spectrograms used to create an eigenspace initially.

data provided by the spectrograms into orthogonal vectors; because each spectrogram contains a large number of components, it seems that they quickly provide all of the necessary combinations for the PCA to resolve into eigenspectrograms (after 2000 iterations, approximately 600 eigenvectors have been created). Even the numerous inter-modulations between frequency components do not provide too much of a challenge for the process, although in general the reconstruction quality is not particularly good. The main advantage over the previous eigenspectrogram output is that the residual measurements produced are much more consistent.

The polychromatic spectrograms represent a significantly larger problem space than that of the monochromatic spectrograms. Unsurprisingly, we saw that approaching the problem in the same manner to the smaller space was not successful, but that the PCA method was manipulating the data in a consistent way. It also highlighted that there was some benefit in trying to expand our defined eigenspace to span a dissimilar problem space.

The rapid drop in the residual measurement values of the reconstructions of the spectrograms with twenty frequency components suggests that the random ap-

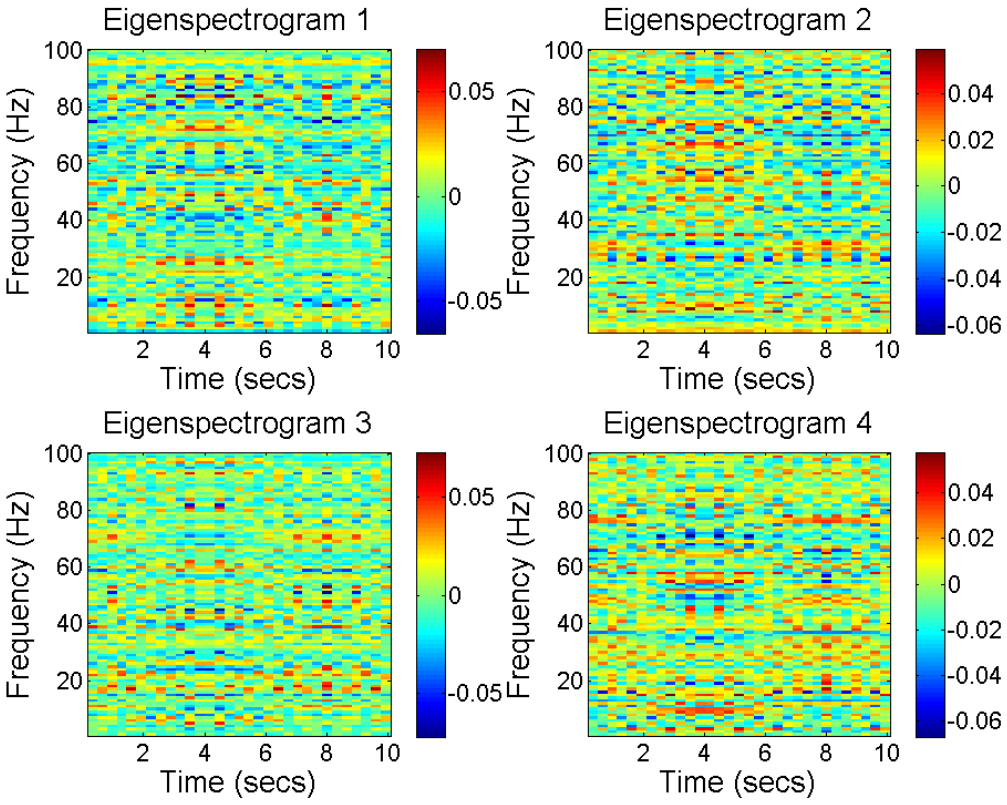


Figure 4.27: Four randomly chosen eigenspectrograms from a set constructed using polychromatic training spectrograms and further polychromatic spectrograms.

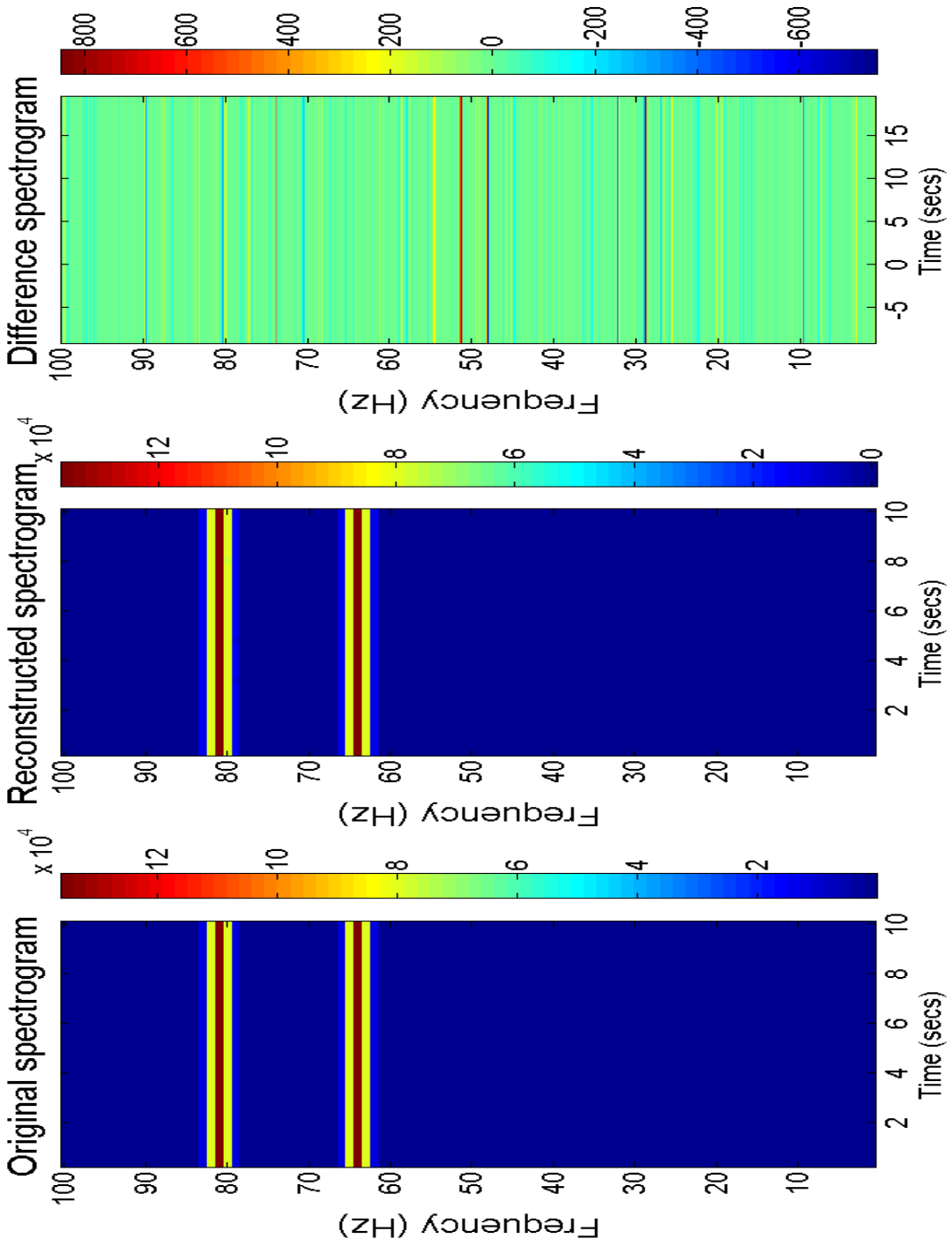


Figure 4.28: Polychromatic spectrogram and its reconstruction using 420 eigen-spectrograms. The third spectrogram, showing the difference between the two, demonstrates that the reconstruction is of poor quality (the residual measurement $R \approx 0.02$).

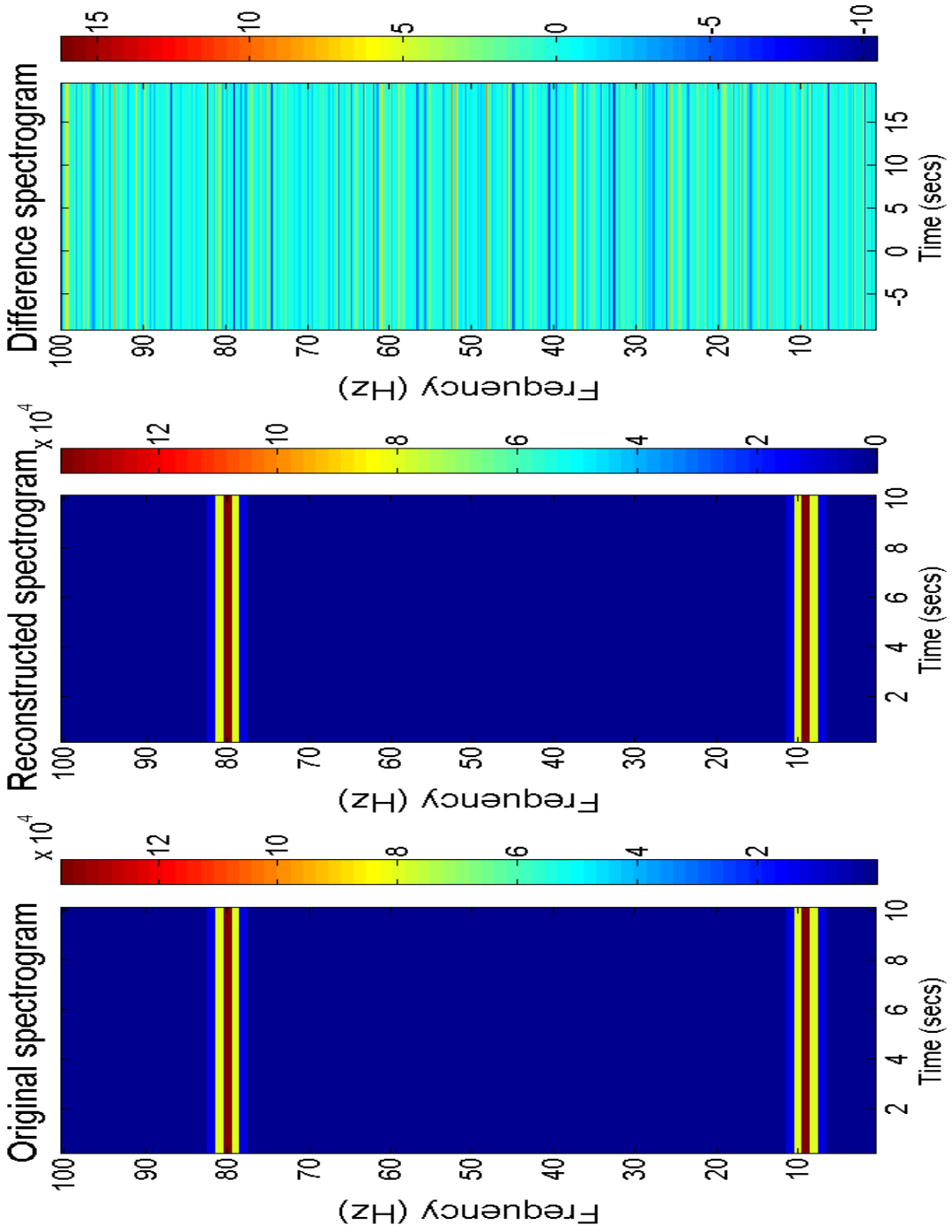


Figure 4.29: Polychromatic spectrogram and its reconstruction using 420 eigenspectrograms. The third spectrogram, showing the difference between the two, reveals that the reconstruction is of fairly high quality ($R \approx 8 \times 10^{-5}$), but still far from perfect.

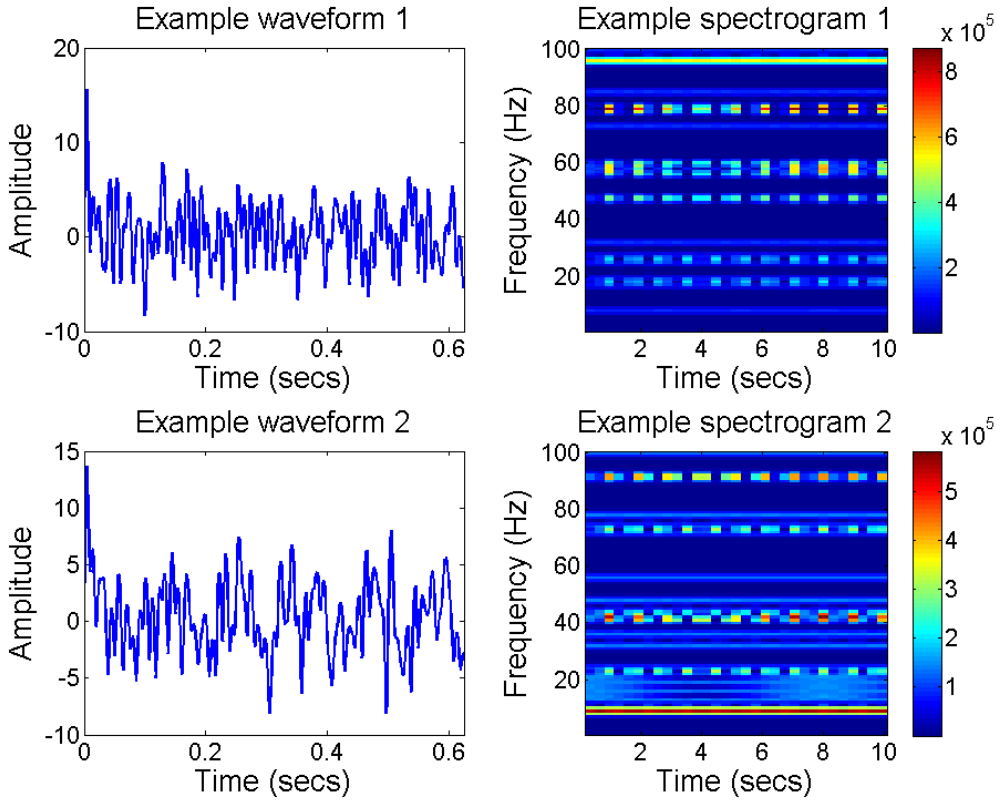


Figure 4.30: The timeseries of a polychromatic sinusoid with a 20 frequency components and the corresponding spectrogram. Only the first 500 data points of the waveforms are displayed.

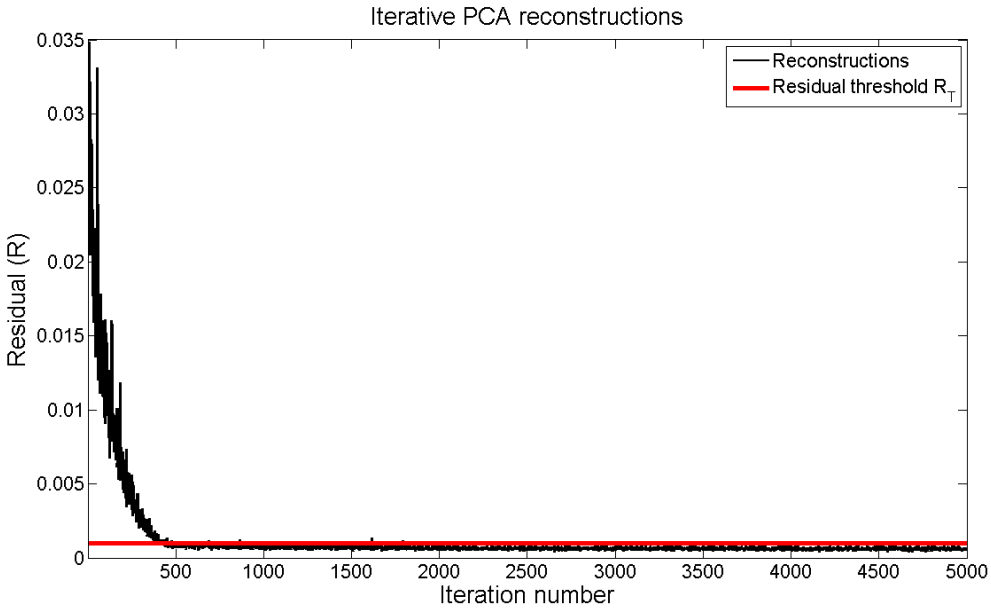


Figure 4.31: The residual measurements of reconstructions during 5000 iterations of the PCA method using training and test spectrograms with 20 frequency components.

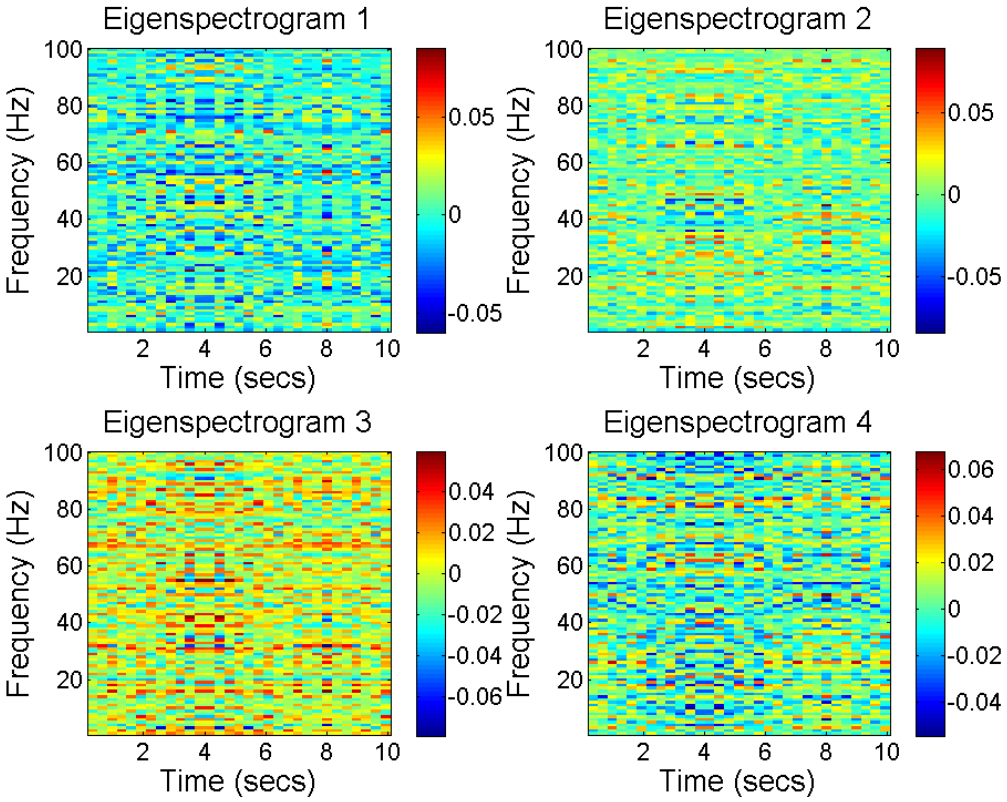


Figure 4.32: Four randomly chosen eigenspectrograms from a set constructed using polychromatic training spectrograms and further polychromatic spectrograms, each with 20 frequency components.

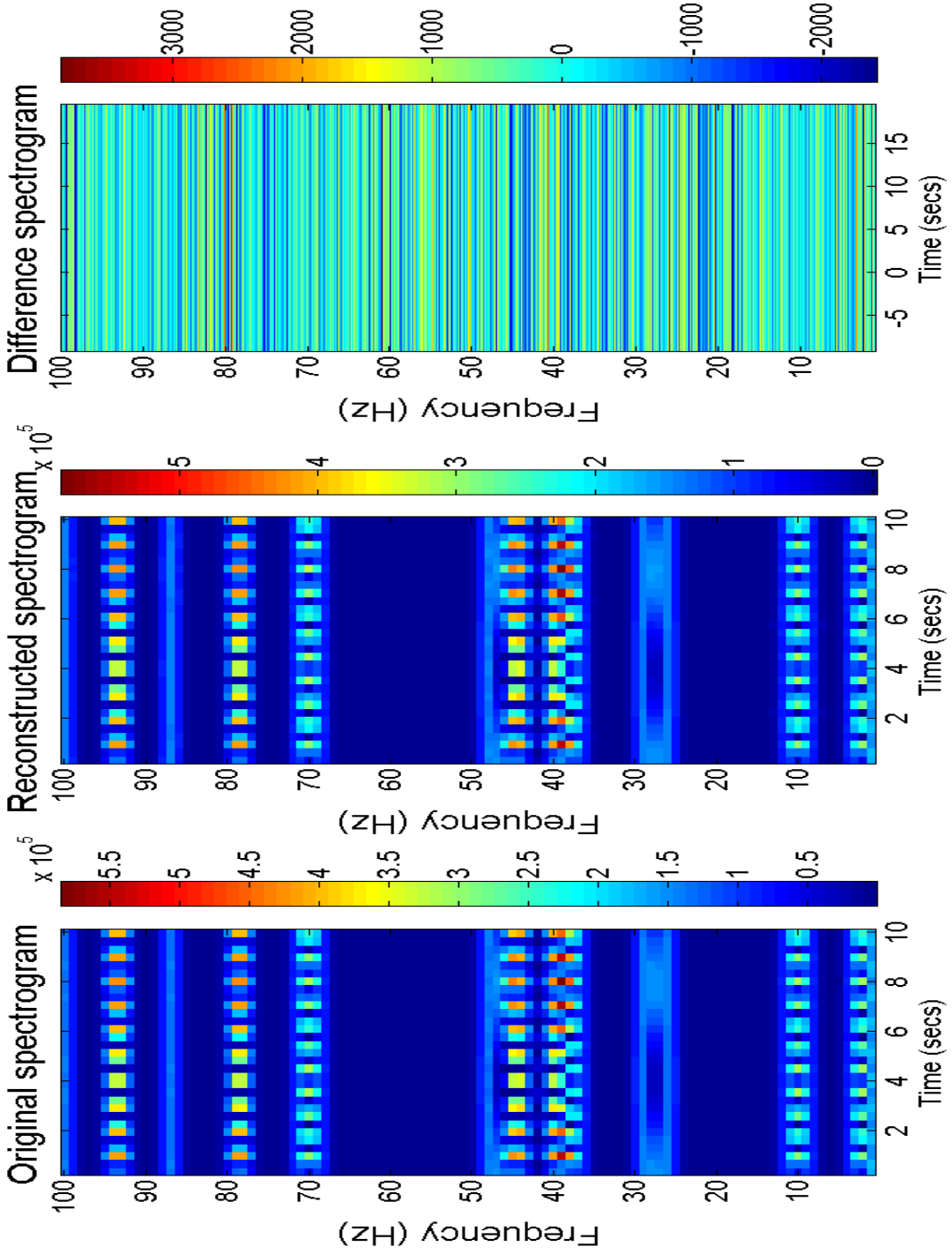


Figure 4.33: Polychromatic spectrogram with 20 frequency components and its reconstruction using 617 eigenspectrograms. The third spectrogram, showing the difference between the two, reveals that the reconstruction is not particularly good quality ($R \approx 0.008$), although the main features have been retained.

proach to adding spectrograms can work well. This is particularly important given our ignorance about the distribution of data in the EMRI spectrogram space, although the disadvantage is that a large number of iterations are required to create a far smaller number of eigenspectrograms. Potentially, the PCA method will be computationally expensive in terms of processing time rather than memory usage.

4.2.1.3 Quasi-monochromatic spectrograms

EMRI waveform spectrograms have multiple frequency components, but the polychromatic spectrograms already examined do not mimic them particularly closely; clearly, they lack the frequency evolution observed in the EMRI spectrograms, and it is necessary to examine the PCA method when faced with this situation.

A sinusoid of increasing frequency provides a rough approximation to an EMRI waveform, but to keep the situation simple we shall only add a positive first-order time derivative $\dot{f}$ to the sinusoids' starting frequency f, and construct the sinusoid according to the basic form

$$y(t) = A \cdot \sin(2\pi f t + \dot{f}). \tag{4.17}$$

Each waveform will only involve a single starting frequency and time-derivative, and are therefore *quasi-monochromatic*. The $\dot{f}$ component will determine the change in the sinusoid frequency between the beginning and end of the waveform, corresponding to the slope of the lines containing power in the resulting spectrogram (see Fig. 4.34). As in previous situations, the amplitude A and the times under consideration, t, are identical for all sinusoids created.

A uniform rate of change of frequency may seem appropriate (every sinusoid created with $\dot{f} = 1\text{Hzsec}^{-1}$, for example), but in fact this will result in an eigenspace spanned by 100 eigenspectrograms, much like the situation using monochromatic spectrograms. This is because once again there are only 100 unique spectrograms possible under these circumstances (a resulting eigenspectrogram set has been included for interest, see C.2). Despite any differences in appearance, the PCA treats each spectrogram simply as a column vector, and the elements of these column can be arranged in any order without losing any information, requiring only a reverse arrangement to become meaningful once again. Naturally, the same arrangement would have to be applied to every spectrogram whenever the PCA is attempted, but the PCA method does not *know* what it being applied to, and will manipulate an

EMRI spectrogram, the image of a face, or even a column vector of random numbers in the same manner.

Therefore, the sinusoids used to create the spectrograms have randomly selected starting frequencies f as well as a randomly selected $\dot{f}$ component, ensuring that the frequency change between the beginning and end of the sinusoid is an integer in the range $1-5$ Hz. The residual threshold is set at $R_T = 0.01$, and 5000 iterations of the PCA method are applied, beginning with a training set consisting of 10 spectrograms constructed in the described manner.

Although not as easy to span as the monochromatic spectrogram problem space, the eigenspace is quickly created (see Fig. 4.35), requiring approximately 500 eigenspectrograms to ensure that the residual measurements R is particularly low (typically, new spectrograms are recreated such that $R \leq 10^{-14}$) . Like those of the polychromatic spectrograms, the resulting eigenspectrograms do not provide much meaningful visual information themselves, and testing the PCA method multiple times under the same condition produces extremely similar output, but it can be seen that the eigenspectrograms are not time-independent. As a final complication, the process is repeated once more with frequency change range of $1 - 10$ Hz. This time, almost 800 eigenspectrograms are required to define the problem space well (see Fig. 4.38 and Fig. 4.39).

It is not surprising that allowing the frequency of the sinusoid to change requires more principal components than monochromatic spectrograms, since the quasi-monochromatic spectrograms display lines of power that can be orientated in a number of ways according to the amount of frequency change; the problem space that must be spanned is therefore bigger. How much bigger depends on the number of orientations that are available to each line.

Despite the ability of the PCA method to eventually define the eigenspace, we are still considering an approximation of EMRI spectrograms. The rate of change of frequency in the sinusoids, $\dot{f}$ will not be uniform, nor is it a first-order derivative with respect to time. Nevertheless, the quasi-monochromatic spectrograms do show that the PCA method is robust enough to generate an eigenspace that spans a problem space where the frequency components of the spectrograms evolve.

We have tested the PCA method on a number of sets of spectrograms, each imitating a particular aspect of EMRI spectrograms. Although this does not equate with tackling the EMRI spectrograms themselves, each of the simplified spectrogram

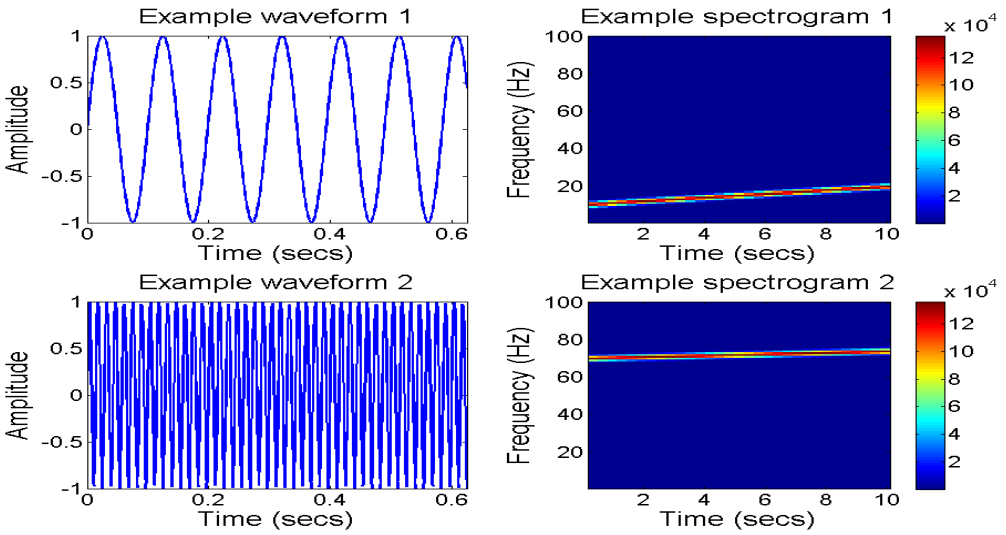


Figure 4.34: The timeseries of two quasi-monochromatic sinusoids and the corresponding spectrograms. Only the first 500 data points of the waveforms are displayed.

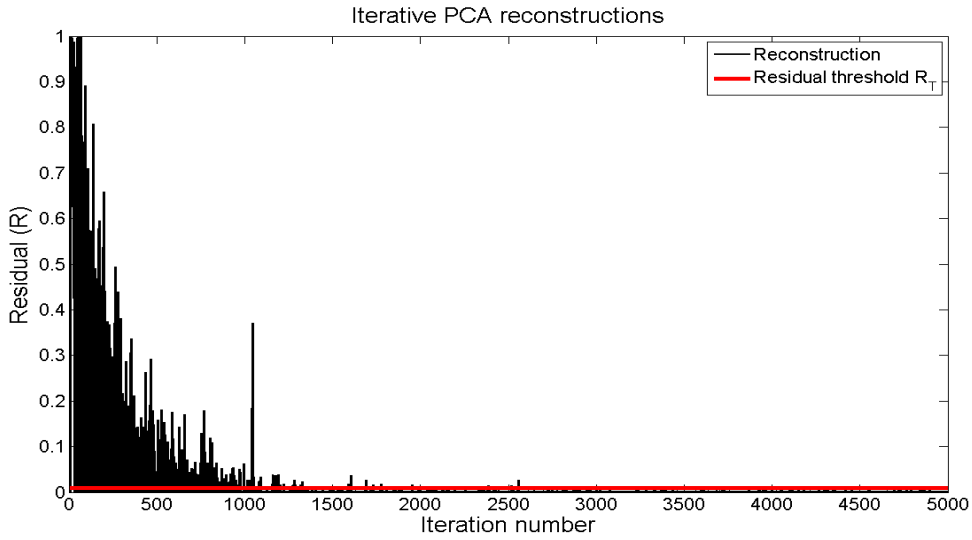


Figure 4.35: The residual measurements during 5000 iterations of the PCA method. The new spectrograms being assessed are quasi-monochromatic (of the first type described), as are the training set of eigenspectrograms.

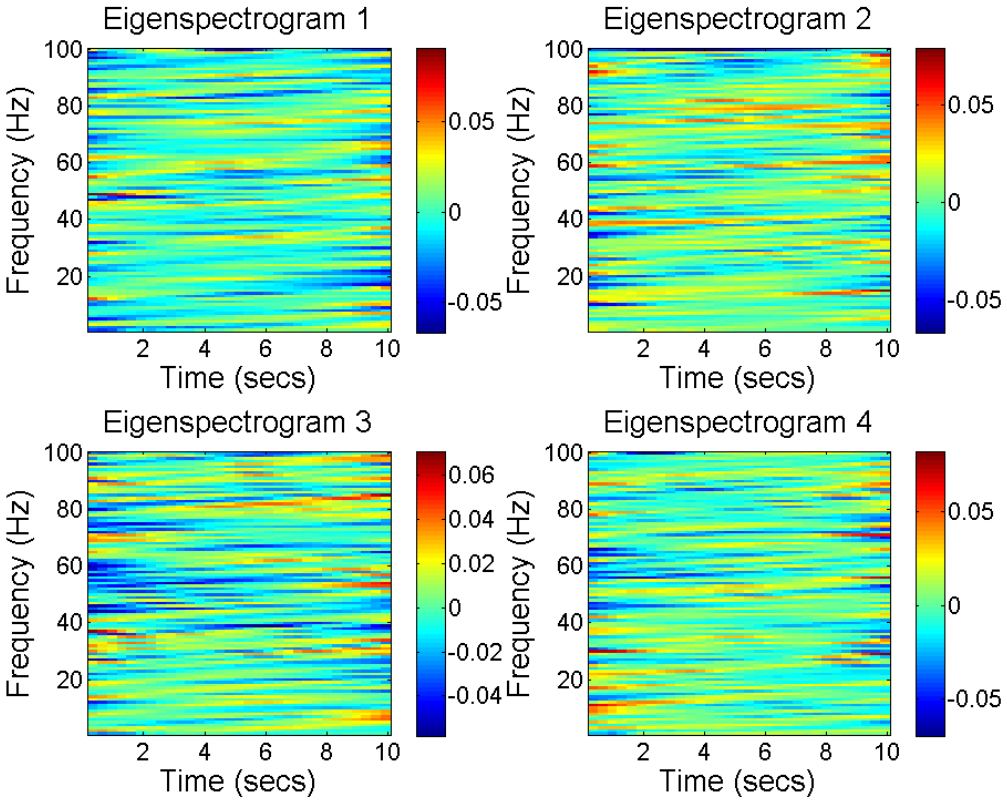


Figure 4.36: Four randomly chosen eigenspectrograms from a set constructed using the first set of quasi-monochromatic spectrograms. It can be seen that the structure is not time-independent.

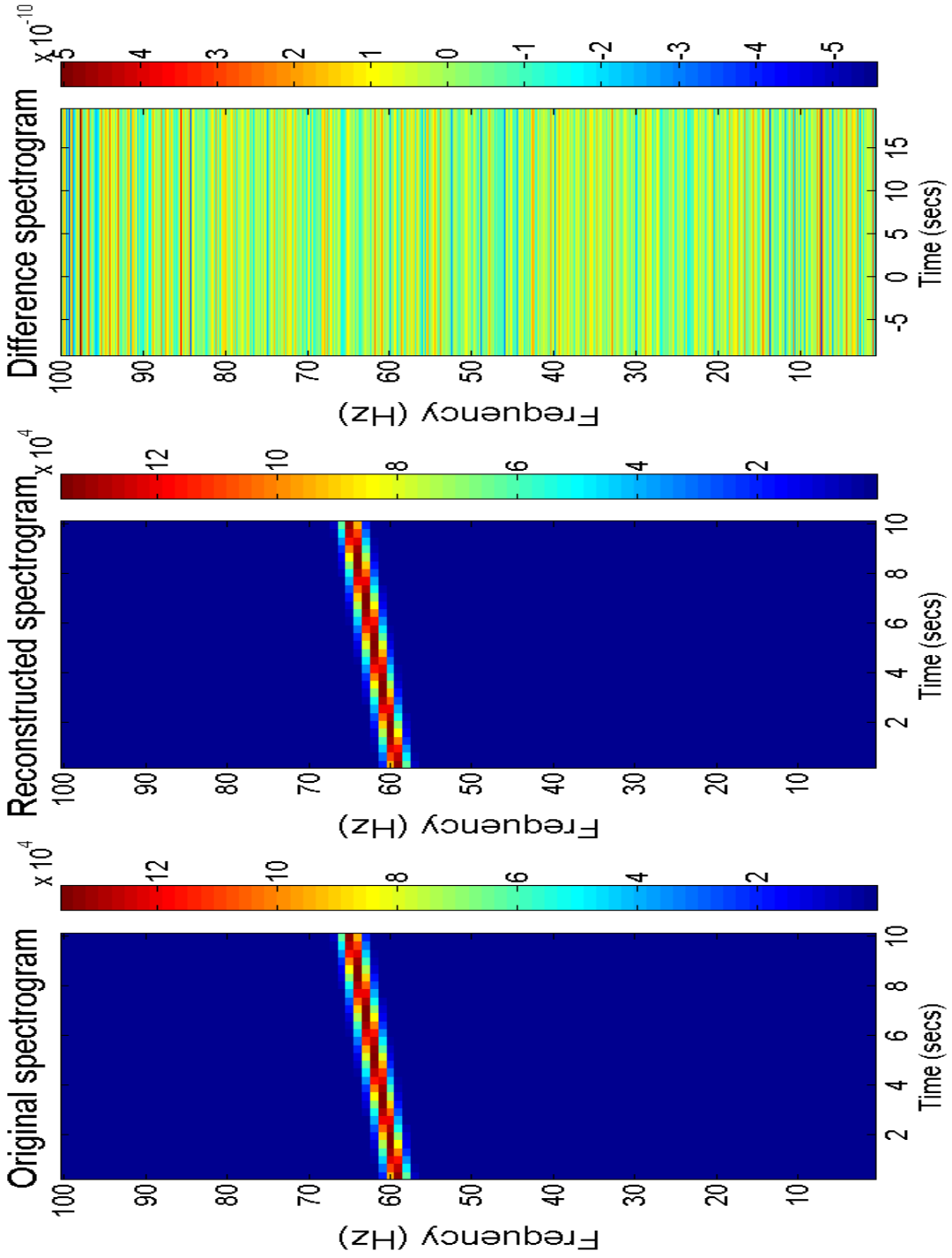


Figure 4.37: A quasi-monochromatic spectrogram and its reconstruction with a set of eigenspectrograms. Although the eigenspace has not been completely spanned, the reconstruction is very close to the original; a spectrogram showing the difference in the power between the original and the reconstruction is also included.

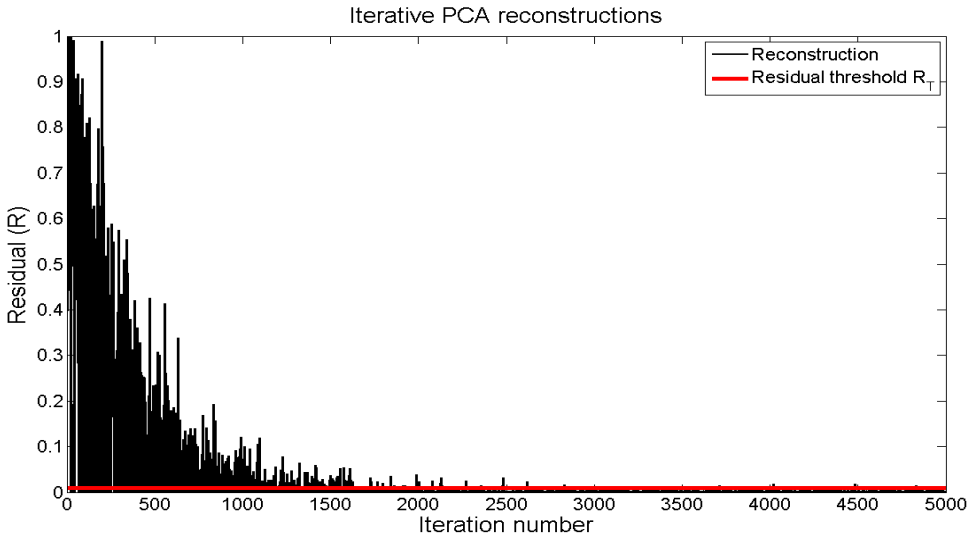


Figure 4.38: The residual measurements during 5000 iterations of the PCA method. The new spectrograms being assessed are the second set of quasi-monochromatic spectrograms, as are the training set of eigenspectrograms.

sets provide some insight about what we may expect from what is undoubtedly a more complex situation.

The monochromatic sinusoid spectrograms served primarily to prove that a simple problem space could be spanned by a set of eigenvectors, and that the PCA method could cope with the iterative incorporation of spectrograms with randomly chosen frequency components. The eigenspectrograms produced contained a couple of oddities that illuminated the inner workings of the algorithms used, but supported the notion that the PCA method was quite robust. Lastly, it provided a clear demonstration of the four categories of test spectrogram.

Polychromatic sinusoid spectrograms were no more difficult to create or manipulate, but belonged to a much larger problem space that could not be spanned easily. Although not realistic substitutes for EMRI spectrograms, they did suggest that creating an EMRI eigenspace will be much more computationally expensive, if in fact possible. They also demonstrated that because the PCA method splits the information it is provided into orthogonal vectors, having more components can actually help to span the space more quickly.

The uniform frequency evolution rate present in the first quasi-monochromatic

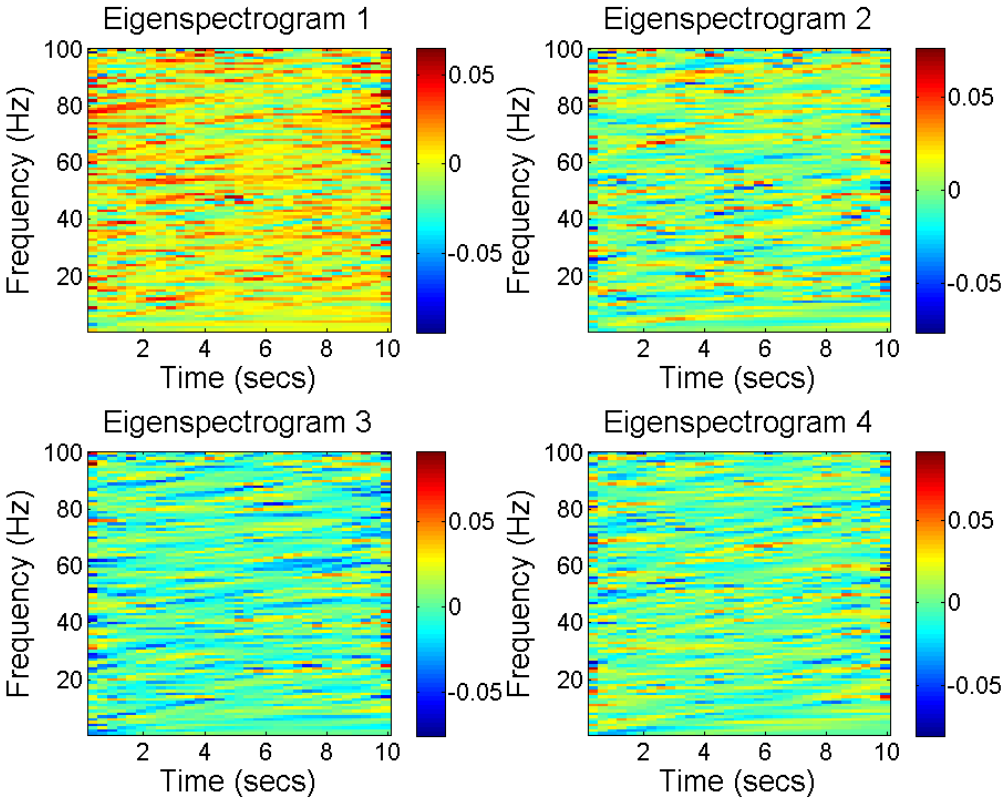


Figure 4.39: Four randomly chosen eigenspectrograms from the second set constructed using quasi-monochromatic spectrograms. Again, the structure is not time-independent.

sinusoid spectrograms proved that the transformation of spectrograms into column vectors (and its 'retransformation') needs only to be consistent, and is not meaningful itself, while the random rates of evolution broke the time-independent nature of the eigenspectrograms and created a larger problem space (though one that could be spanned rather well).

Ultimately, we return to how the sinusoid spectrograms illuminate the problem of applying principal component analysis to EMRI spectrograms. Now, we can be confident that the PCA method is robust enough to handle spectrograms, but should be wary that the problem space may be too large to span. EMRI spectrograms will combine the difficulty of multiple frequency components and the problems of their inter-modulations, as well as the fact that these components will be evolving at different rates in different spectrograms. The frequency range will have a much higher resolution than the sinusoid spectrograms and will not be confined to integer frequencies, while the observation time being considered will be several orders of magnitude larger. In addition, the waveforms and spectrograms are far more computationally expensive to generate and store than the simple sinusoid. Finally, we are very ignorant of the EMRI spectrogram problem space and have no prior information to guide when selecting parameters from which to generate waveforms (and hence spectrograms).

Together, the complexities of the EMRI spectrograms suggest they will pose quite a different challenge from the sinusoid spectrograms. Nevertheless, performing principal component analysis on the sinusoid spectrograms was a valuable exercise, allowing us to investigate more transparent situations and the issues they raised, before proceeding to the main objective.

4.3 Initial tests on EMRI spectrograms

With a preliminary investigation of principal component analysis complete, we now look to apply the PCA method to EMRI spectrograms. However, there are a number of issues that must be addressed before tackling what is a markedly more extensive problems.

Typical EMRI spectrograms

The procedure for generating EMRI spectrograms has already been described (see 3.2). Specifically, for the initial tests in this section, the waveforms are 80640 samples

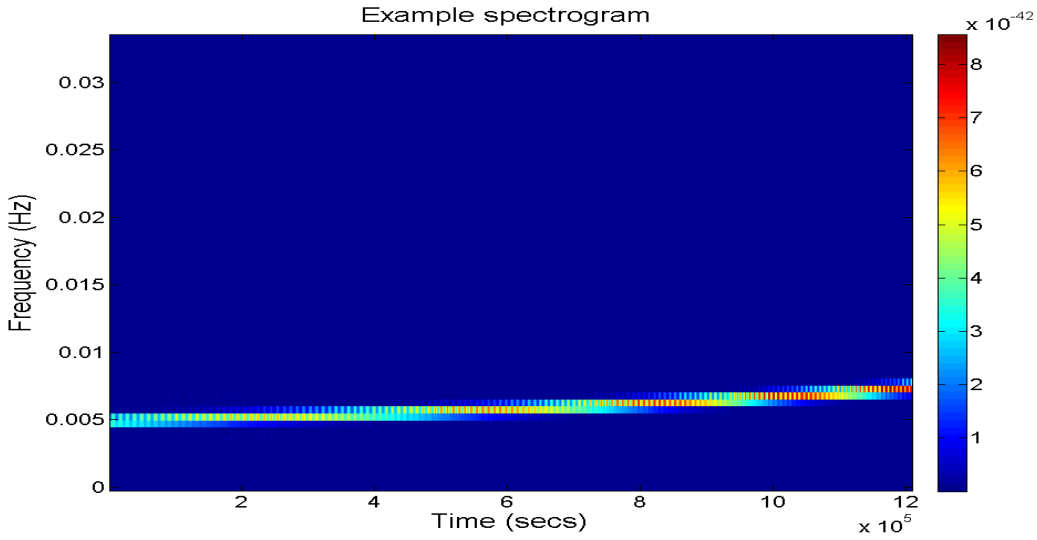


Figure 4.40: A 'typical' EMRI spectrogram. The source parameters are drawn from limited parameter space used in initial PCA attempt, but are not given explicitly.

long, with a sampling frequency of 1/15 Hz (as is the case for MLDC datasets), corresponding to approximately two weeks of simulated data. The spectrograms produced span the frequency range $0 - 33.33$ mHz with a resolution of 0.5208 mHz. As with previous EMRI spectrograms, the time displayed in the included images indicates the time remaining until the coalesence time t_c. An example spectrogram of this type is shown in Fig. 4.40.

4.3.1 Adapting the PCA method to EMRI spectrograms

Although individual EMRI spectrograms can be manipulated in the same way as any other spectrogram, the PCA method was altered slightly from its original form. The majority of these alterations were minor, generally splitting large arrays or organising output in such a way that the increased computational cost of generating and storing EMRI spectrograms and eigenspectrograms could be accommodated (and minimised where possible). Nevertheless, there are a few more significant changes that require a brief explanation.

Fig. 4.41 describes the adapted PCA method in diagrammatic form. To keep matters simple, a number of terms are introduced to denote certain sets of vectors rather than repeat their description each time.

Multiple test spectrograms per iteration

Rather than each iteration generating and assessing a single test spectrogram, it is possible to perform the same operations on a batch $\mathbf{B}$ of p spectrograms, identifying a subset $\mathbf{B}'$ of q candidate spectrograms (where $q \leq p$) that can be used to expand the eigenspace. In practice, the eigenspectrograms of this subset are calculated first (denoted $\mathbf{E_2}$) and joined to the existing eigenspectrogram set (denoted $\mathbf{E_1}$), creating the expanded set $\mathbf{E_3}$. The eigenspectrograms of $\mathbf{E_3}$ are calculated (and denoted $\mathbf{E_4}$), before the next iteration is attempted.

This may seem like needless complication, in particular the calculation of $\mathbf{E_2}$, but there is a possibility that more than one member of the candidate spectrogram set $\mathbf{B}'$ are identical or very similar. If this is the case $\mathbf{E_2}$ will contain a zero-eigenvalue eigenspectrogram (or near-zero) which can be removed before being joined to $\mathbf{E_1}$. This arrangement is less memory-intensive and faster than handling test spectrograms one at a time.

Therefore, we can construct a simple method for creating and expanding the eigenspace.

1. First, generate some training spectrograms, and construct a set of eigenspectrograms $\mathbf{E_1}$ from these.

2. Load a set of test spectrograms $\mathbf{B}$ and evaluate them according to the set $\mathbf{E_1}$.

3. Determine which of the test spectrograms $\mathbf{B}$ are sufficiently far from the eigenspace to warrant their use in expanding this space. This subset is $\mathbf{B}'$.

4. Construct a set of eigenspectrograms from $\mathbf{B}'$. These are denoted $\mathbf{E_2}$.

5. Merge the sets $\mathbf{E_1}$ and $\mathbf{E_2}$ to create $\mathbf{E_3}$. This is done by simply joining the two arrays storing the eigenspectrograms, no further mathmatical manipulation is required.

6. Determine the eigenvectors of the set $\mathbf{E_2}$, denoting them $\mathbf{E_4}$. These eigenvectors span the space occupied by the training spectrograms and the set $\mathbf{B}'$.

7. Repeat stages 2 to 6, replacing the set $\mathbf{E_1}$ with the set $\mathbf{E_4}$ and using a new set of test spectrograms $\mathbf{B}$. Multiple iterations of this method will continue to expand the eigenspace.

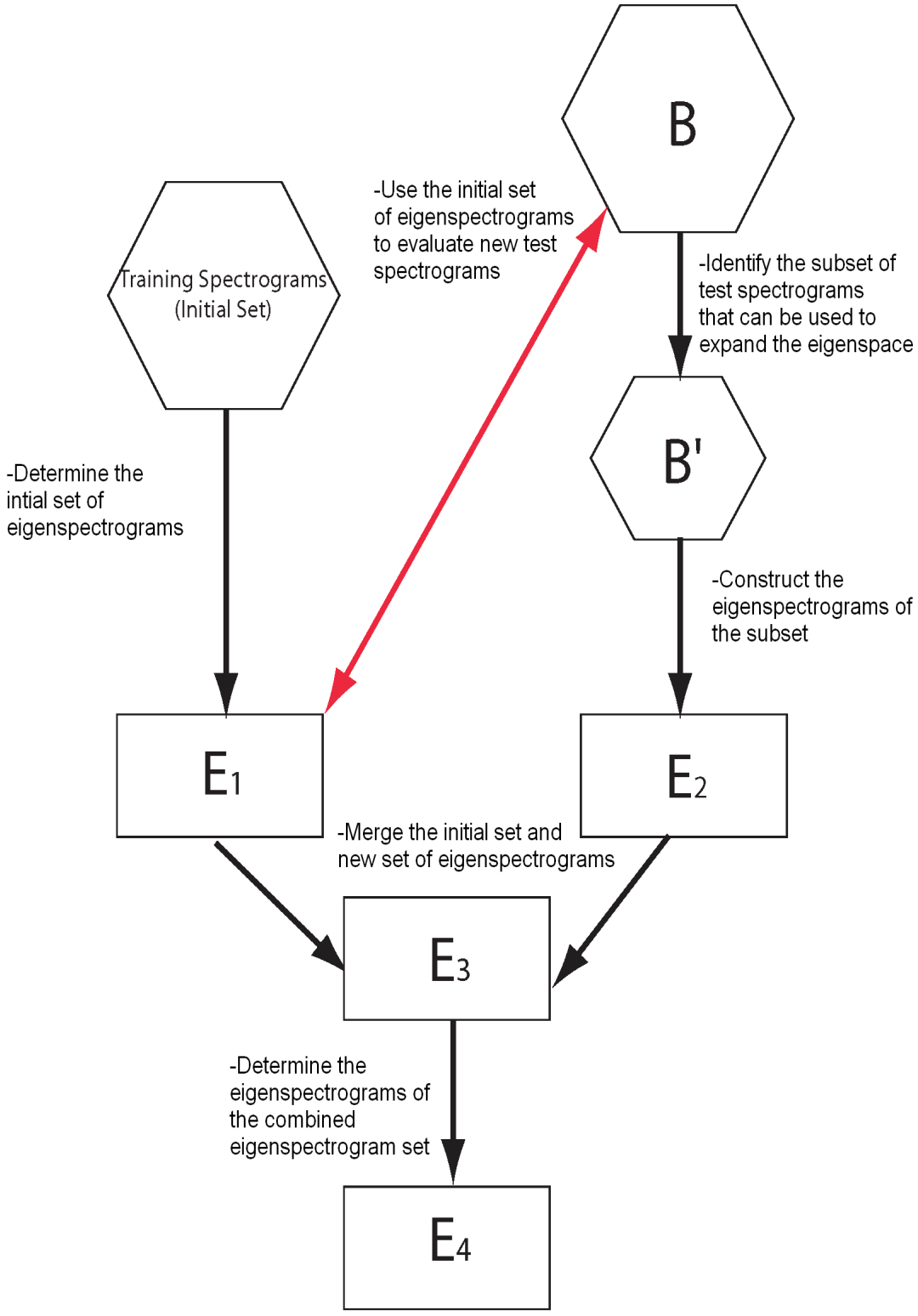


Figure 4.41: The PCA method explained diagrammatically, incorporating alterations made to accommodate the nature of EMRI spectrograms. The red arrow between **B** and $\mathbf{E_1}$ indicates that $\mathbf{E_1}$ is used to evaluate the spectrograms that make up the set **B**.

Weight vectors and class distances

Without detailed prior information about the EMRI spectrogram problem space, we cannot sensibly define spectrogram classes before the iterative part of the PCA method is applied (recall that these were defined retrospectively for the sinusoid spectrograms). Therefore, distance calculations cease to be meaningful too, and the weight vectors of specific spectrograms are only useful in order to determine the reconstruction quality as measured by the residual. As a result the weight vectors are calculated, but not stored at each iteration. The disadvantage is that they can not be recalled for analysis, but it speeds up each iteration of the PCA method and reduces the memory usage. In the adapted method, the residual R is the only value by which the test spectrograms are judged during each iteration.

h_+ and $h_\times$

The EMRI waveform is split into two components, h_+ and $h_\times$. Although theses are not completely independent, at this stage of the analysis we treat them as utterly distinct. Therefore, the eigenspectrogram sets produced only attempt to span the problem space inhabited by spectrograms from one of the components. **As before, the remainder of this chapter only concerns the $h_\times$ component and the resulting output**.

4.3.2 PCA performed over a small region of parameter space

It is not obvious if the available computing resources are able to define the entire EMRI spectrogram problem space, and certainly doubtful that such a thing could be done quickly. A sensible approach then is to apply the principal component analysis to spectrograms drawn from a very small region of the problem space before attempting the same on larger regions.

4.3.2.1 Initial test: a redundant parameter

A useful test at this stage is to identify some redundancy in the EMRI parameter space quickly. In its present format, the power in the spectrograms should vary according to the source distance, but will not be affected by things like antenna pattern or source location. Therefore, two sources that are identical in every respect other than their distance from the detector (D) should produce waveforms that

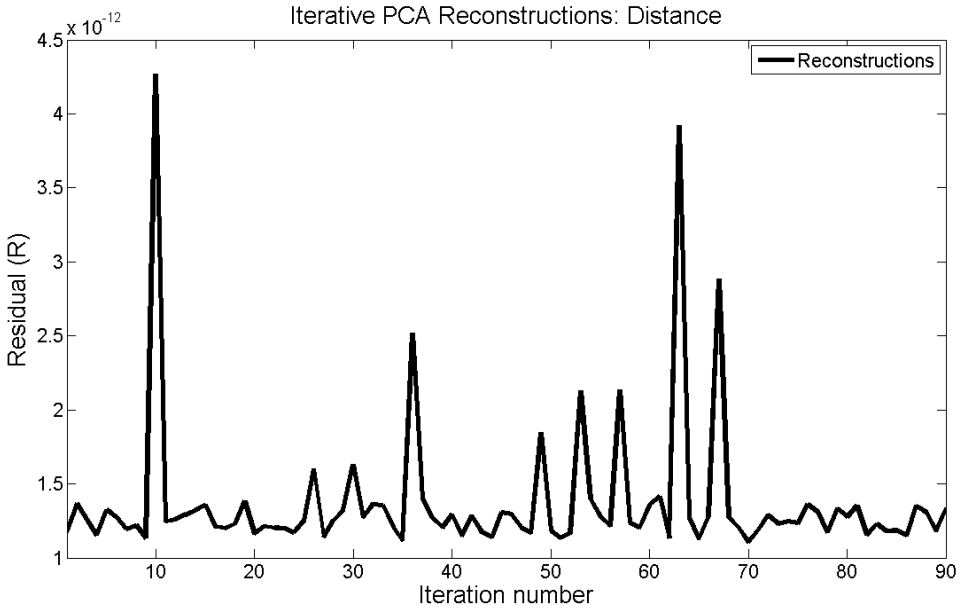


Figure 4.42: The residual measurement R of EMRI signals from sources of varying distances. There are small differences between each measurement, but these are very small. Therefore, we conclude that the source distance is a redundant parameter.

differ only in their amplitude, and the corresponding spectrograms will differ only in the amount of power stored in each time-frequency bin.

In terms of the column vectors the PCA method is manipulating, each test spectrogram will simply be a rescaled version of the training vectors. The eigenspectrogram set created from the training set should contain all the necessary information to reconstruct any test spectrogram simply by rescaling the weight vector of one of the original training set spectrograms. In fact, only one eigenspectrogram should be required; an indication that the PCA method is working is that all other eigenspectrograms should be zero-vectors, and have an associated eigenvalue of zero.

Ten spectrograms were used to create the training set and the corresponding eigenspectrogram set. The distance to the source was selected randomly from the range 1×10^9 - 2×10^9 parsecs, for an integer number of parsecs. As expected only one eigenvector with a non-zero eigenvalue was produced from this set, identical in appearance to the first of the training spectrograms apart from the scaling (see Fig. 4.43). As the iterative part of the PCA method begins, the test spectrograms can be reconstructed extremly well with this eigenspectrogram and the eigenspace is not

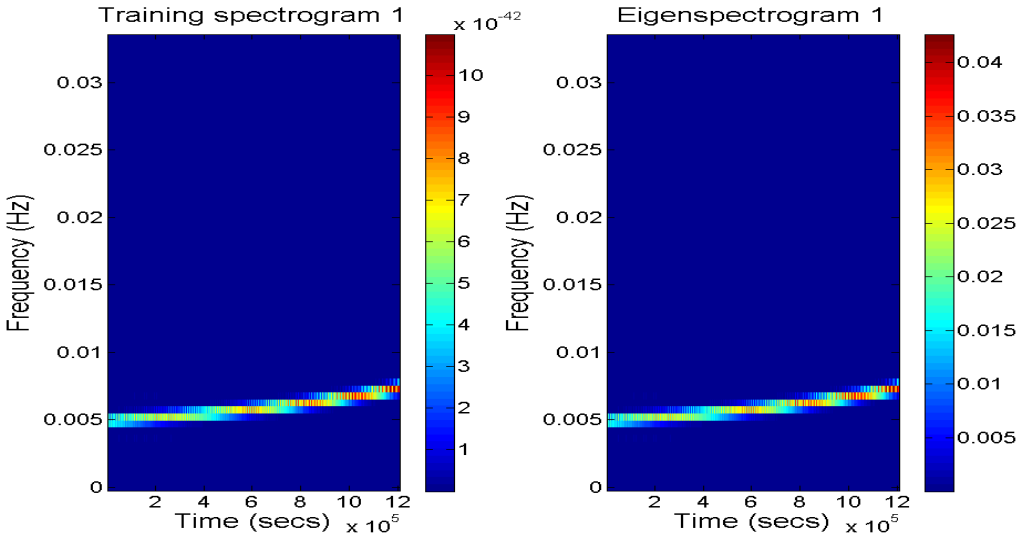


Figure 4.43: The first training spectrogram used and the first eigenvector. There is no additional information incorporated into the spectrogram set, and so the eigenspectrogram produced is identical to the original spectrogram (apart from normalisation).

expanded (we set a residual threshold of $R_T = 0.01$.

Fig. 4.42 shows the reconstruction quality of 200 test spectrograms (assessed in batches of 10) where the mean value of $R = 1.375 \times 10^{-12}$ with a standard deviation of 4.95×10^{-13}. Although this is not a particularly extensive test, it quickly becomes clear that the behaviour is not going to change and the process is halted.

We may conclude then that the single-eigenspectrogram set spans this limited problem space, as we predicted. Repeating this test using different randomly-selected values for the source distance produces the same output. Under these circumstances, the source distance is a redundant parameter because it does not affect the residual measurements to any significant degree.

4.3.2.2 Initial test: varying two parameters

Having established a source of redundany in the EMRI problem space as it has been presented, we extend the tests to examine altering two parameters. The set of parameters governing the EMRI evolution allows a large number of possible combinations, but to go through them exhaustively would be extremely time-consuming

and all may not be particularly informative. Accordingly, three combinations were chosen, each with what was believed to represent a different 'category'.

A) CO mass and SMBH mass A brief look at some example spectrograms (before PCA was being considered) suggested that the masses of the two black holes in the EMRI system seem to affect the gravitational waveform fairly significantly. By varying these two parameters, we expect the resulting spectrograms to be fairly distinct from each other; the lack of redundancy here will mean that the residual will probably be fairly high even after a large number of iterations.

B) CO mass and source distance Now, we have a combination of what we believe to a parameter that is not redundant (CO mass) with another that is (source distance). Together, we expect them to present a parameter space that has some redundancy, but as yet we are unsure how this will affect the creation of the eigenspace.

C) Φ and S These two parameters have not previously been considered, and are best classified as a combination that we are entirely ignorant of. The lack of prior information about the type of spectrograms that will be produced is another useful test of the robustness of the PCA method, since there is little chance of (even unconsciously) choosing values for either parameter that we suspect will be 'well behaved' (by which we mean they will produce residual measurements according to some fashion we know that the PCA method can cope with).

Ignorant of the redundancy that we might expect in the parameter space of each combination, the PCA was attempted with two different residual thresholds below which spectrograms would be used to expand the eigenspace. The first, $R_T = 0.1$, was deliberately high, while the second, $R_T = 0.01$ would help to define the eigenspace better. Using two thresholds allowed us to make a rough estimate of how large the eigenspectrogram set would have to be to define the eigenspace any further.

As a final condition, it was decided that allowing the parameter values to vary completely randomly throughout the selected parameter ranges may be too computationally expensive, so they were restricted to moving on a grid with uniform

Test	**A**	**B**	**C**
Range (parameter 1)	$100-200M_\odot$	$100-200M_\odot$	$0-2\pi$
Resolution (parameter 1)	$1M_\odot$	$1M_\odot$	$\pi/180$
Range (parameter 2)	$1\times10^6-1.1\times10^6 M_\odot$	$1\times10^9-2\times10^9$ pc	0.1 0.6
Resolution (parameter 2)	$100M_\odot$	1×10^6 pc	0.01
Number of possible points	1×10^5	1×10^5	18×10^3

Table 4.2: Test parameter ranges and resolutions for some two-parameter initial tests.

resolution. Naturally, this means we are not able to sample completely freely from the designated parameter ranges, and this restriction is discuessed in section 7.3. A list of the parameter ranges, their resolution, and the number of possible points in parameter space for each combination are given in Table 4.2.

What do the tests show?

A) The residual measurements follow a curve that dips sharply before leveling out, aping the type of behaviour from the tests of the sinusoid spectrograms with 20 frequency components (see Fig. 4.31). This shows that two EMRI spectrograms taken from this range are comprised of similar components to a large extent, while the differences can be incorporated into the eigenspectrogram set and expand the eigenspace.

However, there is a large difference between the number of eigenvectors required to define the eigenspace better than the two residual thresholds. When $R_T = 0.1$, only 67 are needed, while the $R_T = 0.01$ threshold generates 366 eigenvectors, and repeated tests of this parameter space produce similar results. It seems that restricted to this region of parameter space, it is fairly easy to define the problem space to a limited degree, but to define it well is significantly more difficult.

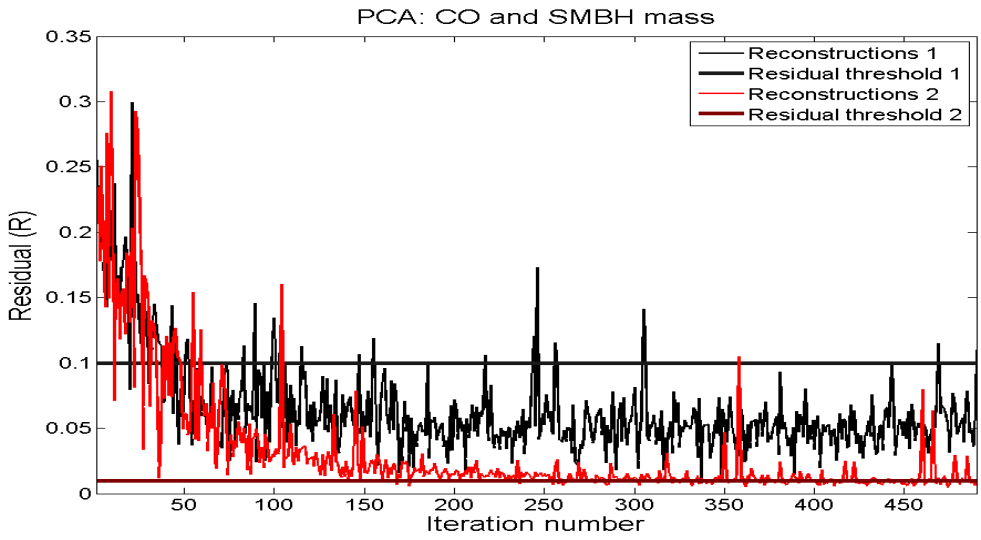


Figure 4.44: Residual measurements of test spectrograms in a two-parameter space (CO and SMBH mass). Two different residual threshold are applied (0.1 and 0.01) producing slightly different long-term behaviour.

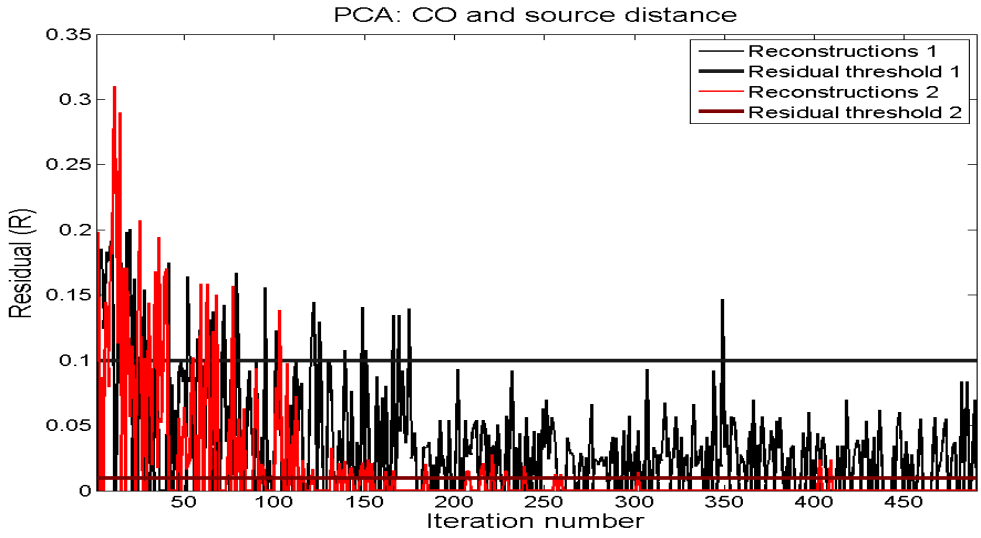


Figure 4.45: Residual measurements of test spectrograms in a two-parameter space (CO mass and source distance D). Two different residual threshold are applied (0.1 and 0.01), changing the long-term behaviour.

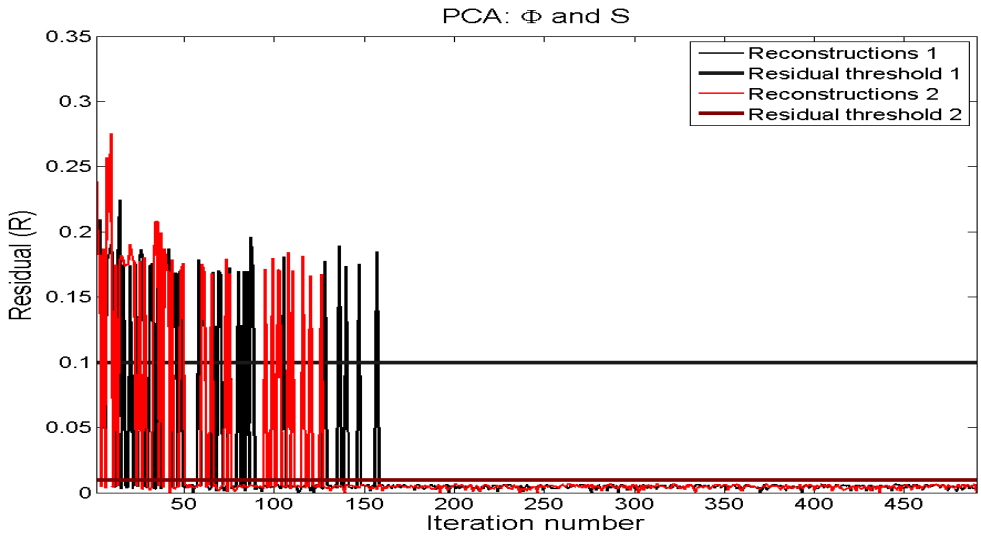


Figure 4.46: Residual measurements of test spectrograms in a two-parameter space (mean anomaly angle and spin magnitude). Two different residual threshold are applied (0.1 and 0.01), but the long-term behaviour is similar.

B) The second test behaves differently. Again, spectrograms are similar, so the residual measurements become quite small rather quickly, but further test spectrograms do not seem to expand the eigenspace incrementally; incorporating them into the eigenspectrogram set only appears to help reconstruct that particular spectrogram but not others. That is, after a rapidly dropping slope for the first few iterations, the slope disappears and we are faced with a number of fairly independent 'events'. In this way, the parameter is similar to the polychromatic spectrograms in 4.2.1.2.

The higher threshold requires 49 eigenspectrograms and the second requires 100. In particular, the set of 100 eigenspectrogram defines the eigenspace very well, as subsequent reconstruction attempts have $R \approx 1 \times 10^{-13}$. It seems that in this case, the CO mass is indeed the only significant parameter and the source distance is irrelevant. This is encouraging since we are identifying redundancy in the EMRI parameter space as intended, but is slightly troublesome since every possible CO mass spectrogram might be different enough to require their own eigenvector to define it well.

C) This time, the spectrograms are very similar, and the parameter space can be

spanned to a high accuracy with a far fewer number of eigenspectrograms than the number of possible combinations (55 and 53 for $R = 0.1$ and $R = 0.01$) respectively.

It appears that the spin magnitude of the SMBH is probably the most significant parameter here. In retrospect this might seems clear, that changing the mean anomaly angle was only really adding a small phase shift to the gravitational waveform, information that would be lost during the construction of the spectrograms. In reality, we can't make such bold claims; there are only 50 possible spin magnitude possibilities, so the mean anomaly angle must impart at least a bit of unique information to the spectrograms. Therefore, despite their combined parameter space being highly redundant (at this resolution), the 'shape' of the eigenspace is not entirely obvious.

Regardless, this is an encouraging result to some extent. Similar behaviour from other combinations of EMRI parameters would suggest significant redundancy in the EMRI problem space and the possibility that an easily manageable number of eigenspectrograms can be used to span it.

There are a number of lessons to learn from the limited parameter space tests, but the different output produced by the PCA method from each test makes it difficult to apply these to the EMRI problem space as a whole. The three tests represent a small fraction of the number of possible two-parameter combinations, and there is insufficient time to exhaustively explore each one. Furthermore, the different output produced offers few clues about the behaviour that might be expected from a three-parameter test (although source distance does appear to be consistently redundant) or larger parameter spaces. Nevertheless, the very fact that the output could be so varied for different parameter spaces forces us to appreciate the complexity of the EMRI problem space and the wisdom of investigating the simple sinusoid spectrograms first.

The tests also show that there will probably not be a linear relationship between the number of eigenspectrograms required to span the space to a certain accuracy, and that accuracy. It may be the case that computational resources are the deciding factor in the residuals that may be expected. We must consider the R-value threshold levels more carefully in future; blanket thresholds are not likely to be sufficient. It would make more sense to have a variable threshold that first starts high to

incorporate only significantly different spectrograms before lowering and gradually defining the eigenspace to better match the problem space.

4.3.3 PCA performed over a large parameter space

We may conclude from the two-parameter tests that although they demonstrate the ability of the PCA method to process the information presented by the EMRI spectrograms, we cannot yet extrapolate the performance of the program on the entire EMRI problem space. In an attempt to do so to a limited degree, we perform a quick test of the program's performance on a much wider parameter space. We are not yet ready to deal with the entire EMRI problem space, but this should give an indication of the output we might expect.

The parameter ranges and resolutions used in the large parameter space test are given in Table 4.3[1], as well as the number of possible combinations of the parameters. As can be seen, this is a truly huge space (potentially), so only a residual threshold of $R = 0.1$ is used, but otherwise the PCA method is used in exactly the same manner as the two-parameter tests.

As Fig. 4.47 shows, the large parameter space does not yield to the PCA method quickly, with the residual measurements generally remaining above $R = 0.1$. In fact, almost no spectrogram can be reconstructed to within the quality demanded, and after 489 are incorporated into the eigenspectrogram set the process is halted due to memory constraints. However, it may still be possible to glean valuable information about the eigenspace from the data we can collect.

In one sense the output is not encouraging, the number of eigenspectrograms required to span the EMRI problem space is clearly beyond the computational capabilities of a current desktop computer. Regardless, the residuals do appear to be improving, albeit slowly, and assuming that this trend continues it is possible to make a *very* rough estimate of how many eigenspectrograms are needed before the spectrograms might be expected to reconstructed with $R \leq 0.1$ on a regular basis; a linear fit of the large parameter space suggests that ≈ 1700 eigenspectrograms would be required. In its current incarnation, storing an eigenspace with this

[1]Noticeably absent is an alteration to the eccentricity of the orbits. This was due to a difficulty in maintaining the accuracy of the ODE solving process at the time. However, the extent of the parameter space was already large enough to prove impossible to span with the available computational resources.

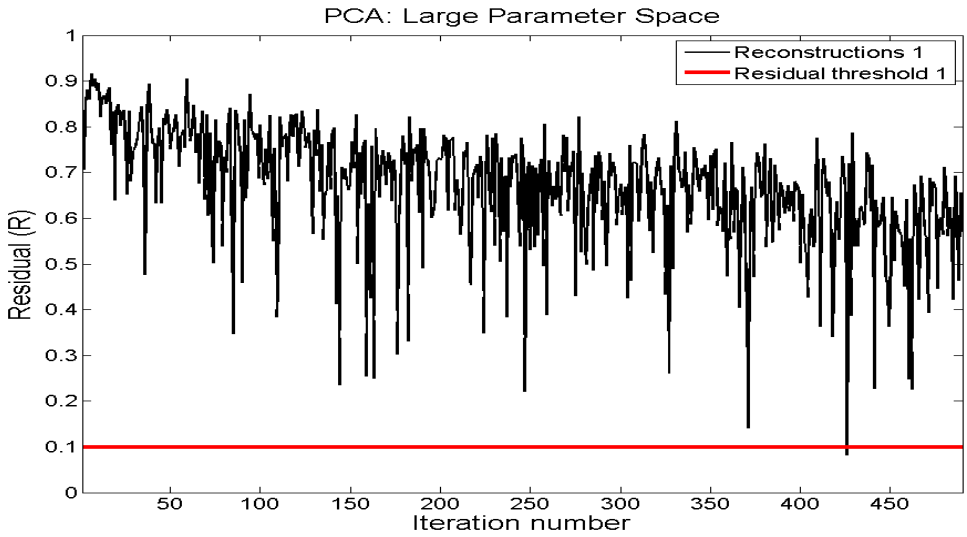


Figure 4.47: The residual measurements for multiple iterations of a large parameter space. Test spectrograms are examined and incorporated into the eigenspectrogram set; the residual decreases as the problem space begins to be spanned, but the reduction is gradual and is fuelled by the inclusion of so many new spectrograms that there is insufficient memory to follow this process for a large number of iterations.

Parameter	Parameter Range	Parameter Resolution	Number of possible points
μ	$100-200M_\odot$	$1M_\odot$	100
M	$100-200M_\odot$	$1\times10^6-1.5\times10^6M_\odot$	5000
S/M^2	0.1–0.6	0.01	50
$\tilde{\gamma}_0$	$0-2\pi$	$\pi/180$	360
Φ_0	$0-2\pi$	$\pi/180$	360
θ_s	$0-2\pi$	$\pi/180$	360
ϕ_s	$0-2\pi$	$\pi/180$	360
λ	$\pi/24-11\pi/24$	$\pi/180$	75
α	$0-2\pi$	$\pi/180$	360
θ_k	$0-2\pi$	$\pi/180$	360
ϕ_k	$0-2\pi$	$\pi/180$	360
θ_k	$1\times10^9-2\times10^9$ pc	1×10^6 pc	1×10^5
	Total		1.469×10^{32}

Table 4.3: Parameter ranges and sampling resolutions for a large parameter space. The total number of possible selections is vast, but the principal components of the spectrograms within the parameter space may be significantly smaller.

number of eigenspectrograms would require an order of magnitude more memory than currently available to a standard desktop computer. Of course, this doesn't take into account that in the long-term, were it feasible, the PCA output may be significantly different; we are assuming that we are looking at the initial stages of something like Fig. 4.44, although this is not necessarily true. Furthermore, it is difficult to be sure the length of the flat tail in the residual measurements that would be required before we could be confident that the space had been well spanned, but if the estimate was roughly accurate (to within even a few orders of magnitude) we are still using a great deal less computational power than required by matched filtering.

We know for sure now that the EMRI problem is too difficult to 'solve' completely, even within the boundaries that we had originally established. Consequently, we shall stay away from further larger parameter space tests so that we can examine the issues raised by the tests in Section 4.3 in greater detail without falling foul of finite computing resources. Subsequent examinations will concentrate on individual aspects of a PCA-based approach to detecting EMRI gravitational wave signals.

Chapter 5

A PCA-based search for EMRI signals

We have shown that it is possible to perform a principal component analysis on EMRI spectrograms, but must address the applications for the information produced; that is, we have to show how the method we have employed can be used to search for gravitational wave signals from EMRIs. The following chapter discusses a possible search method and the results it produces under a variety of conditions.

5.1 Search method

Previously, two intuitive methods of assessment have been discussed (see 4.1.2), but since we have not successfully spanned the entire EMRI spectrogram problem space we have concentrated on the residual measurements of test spectrograms while the distance from spectrogram classes has been abandoned. Although this is not an ideal situation, it does suggest a very simple search method building on concepts already introduced.

Rather than consider a single test spectrogram, we now consider a timeseries $H(t)$, of length L, containing a $h_\times$ polarisation EMRI gravitational waveform. This timeseries can then be split into i segments of a length equal to that used to generate the EMRI spectrograms that are used to construct the eigenspace (L_s, say). Each segment can be turned into a test spectrogram T_i and reconstructed according to its projection into the eigenspace; segments that can be reconstructed well inhabit the

problem space and are candidate detections, while those that cannot are far from the eigenspace and are deemed unlikely to be EMRI signals.

In short, the search method moves through the timeseries and selects a segment of it, checking to see if this segment 'looks like' an EMRI spectrogram. A detection may be registered, or not, depending on the similarity demanded, but either way this slice is abandoned and the search moves on until the entire timeseries has been segmented and examined.

Naturally there are a number of issues that require clarification, given below. However, the complexity of the EMRI spectrograms immediately suggests a number of circumstances that should be looked at individually before a search of more realistic timeseries' are conducted, and each of these demonstrates the workings of the search as well as any description.

The timeseries $H(t)$ This data needs only to be in the same format as the usual EMRI analytic kludge waveforms. The sampling frequency is a crucial aspect, but the timeseries itself can cover a much longer period of time than the usual EMRI waveforms. The only constraint is that the timeseries must clearly be at least as long as each segment (that is, $L \geq L_s$).

'Waveform time' The length of the timeseries (and its sample frequency) defines a period of time, but the waveforms do not correspond to a particular date, as they would if the searches were performed during the time the LISA mission was active; this is natural enough, since we are looking at simulated data. Rather than attempt to create some reference point (such as the date we might expect LISA to begin broadcasting strain measurements), we will instead simply consider the time only with respect to the start of a timeseries, which will begin at time $t = 0$ seconds. We will always assume a sample frequency of $f_s = 1/15$ Hz, and will therefore give Waveform time in seconds.

The spectrograms Each segment must be turned into a spectrogram in the same manner as those used to define the eigenspace, otherwise a valid EMRI signal might project into the eigenspace in an unfamiliar manner and escape detection. Similarly, there is the potential for false detection if the spectrogram generation method is altered, projecting non-EMRI signals into the eigenspace such that they appear as EMRIs.

In practice it is really only necessary to ensure that a segment of the correct length L_s is selected, and that any windowing or other manipulation of the waveforms that generated the original spectrograms is also performed on the segments.

The segments We determine the segments by creating an index of starting positions in the timeseries that indicate the starting point of the data that we wish to test, and create a related index that marks the end point. If we define a start position x, and calculate the end position $y = x + L_s$, then every timeseries data point between these two points define the segment that is turned into a spectrogram and tested.

The search method allows us to define a search 'resolution' in terms of the values of consecutive entries in the starting/end position indexes, or equivalently in terms of the number of timeseries data points that consecutive segments will differ by. For example, with a search resolution of k, a segment T_a will contain the timeseries data between x and $x + L_s$, and the next segment T_{a+1} to be turned into a spectrogram will contain the timeseries data between $x + k$ and $x + k + L_s$.

The total number of segments that will be turned into spectrograms and analysed, i, is therefore

$$i = \frac{L - L_s}{k} + 1. \tag{5.1}$$

Depending on the length of the timeseries and the search resolution desired, i will not be always be an integer value, in which case it is rounded down to the nearest integer.

There are also two particularly significant segments. The first is the segment that produces the minimum reconstruction residual in the entire search T_m. The second, T_c, is the segment which is closest to the EMRI signal in terms of the start and end data points in the timeseries; depending on the position of the waveform within the timeseries and the resolution of the search this may or may not be an exact match. Although these designations are often assigned to the same segment, this is not always the case.

T_m can be determined easily, but we are only able to determine T_c in tests where we can position the EMRI signal where we wish. Due to this it will be a

useful indicator of how well the search method performs when we control the construction of the timeseries $H(t)$, but cannot be used when we do not have this control.

When plotting the residual measurements of spectrograms, the final data point of the segment used will define the corresponding Waveform time representing that spectrogram. This is the most intuitive designation given the backwards-evolution of the EMRI signals; if we know the coalescence time of a signal then the spectrogram assigned this time will contain the segment-length of evolution leading up to this point.

The possibility of $0/0$ Though we do not have any prior knowledge of how the search will work under some circumstances, we can prepare ourselves for the situation where a segment will contain no gravitational strain at all. In this situation the resulting spectrogram will have zero weight associated with each possible eigenspectrogram. This is entirely expected, but the reconstruction will also be contain no data and be considered 'perfect'. However, the lack of power in either spectrogram will create a situation where the reconstruction residual will be $R = \frac{0}{0}$, which is determined to be 'NaN' according to Matlab ('Not-a-Number').

Because this will *only* occur where there is a complete lack of gravitational wave strain, these entries are automatically replaced with $R = 1$ to make it clear that there is no signal in the segment.

Fig. 5.1 presents the search method in diagrammatical form, and the annotations **A** - **F** are explained below.

A The timeseries in loaded into memory.

B A 'index list' of start and end points within the timeseries is constructed, defining the segments of the waveform data to evaluate. The specific contents of the list will depend on the search resolution, the length of the timeseries data and the length of the eigenspectrograms used to evaluate each segment (equal to that of each segment).

C Select a segment of the timeseries data, using the index list points. Each segment is evaluated once, and it is most straightforward to work methodically through the index list in order of increasing Waveform time.

D Create a spectrogram from the selected timeseries segment.

E Evaluate the segment spectrogram using an eigenspectrogram set as a test spectrogram according to the usual PCA method described previously, producing a residual measurement for this segment.

F Store the residual measurement from the segment. At this point, processes **B** - **E** are repeated, selecting a different entry from the index list and using it to select and evaluate a different segment of the timeseries.

5.2 Searching for noiseless EMRI spectrograms

As noted, the search method can be applied to any timeseries in the correct format. To keep matters straightforward, we consider only isolated EMRI signals free from the effects of an antenna pattern or noise to begin with, yet we must simplify the problem space even further in order to produce useful output.

In order to be clear about the detections, we test the timeseries $H(t)$ on a set of eigenspectrograms that span a limited parameter space to a high degree. The most obvious choice is to identify a single, redundant parameter in the EMRI problem space such as distance D, but in fact it will be beneficial to engineer a situation where we can modify this single parameter in a way that will change the resulting spectrogram so that it cannot be reconstructed perfectly. This seems confusing initially, but will prove to be informative as we examine different circumstances under which the search will operate.

Thus, eigenspectrograms that define the eigenspace to be searched are drawn from the limited parameter range in Table 5.1, and are generated in the same manner as described in Section 4.3. Due to the finite resolution of the search, the parameter space is only 'spanned' as long as the test parameters are drawn from the same finite grid. This allows us to generate EMRI spectrograms with identical parameters, those that are slightly different, and finally those that are significantly different.

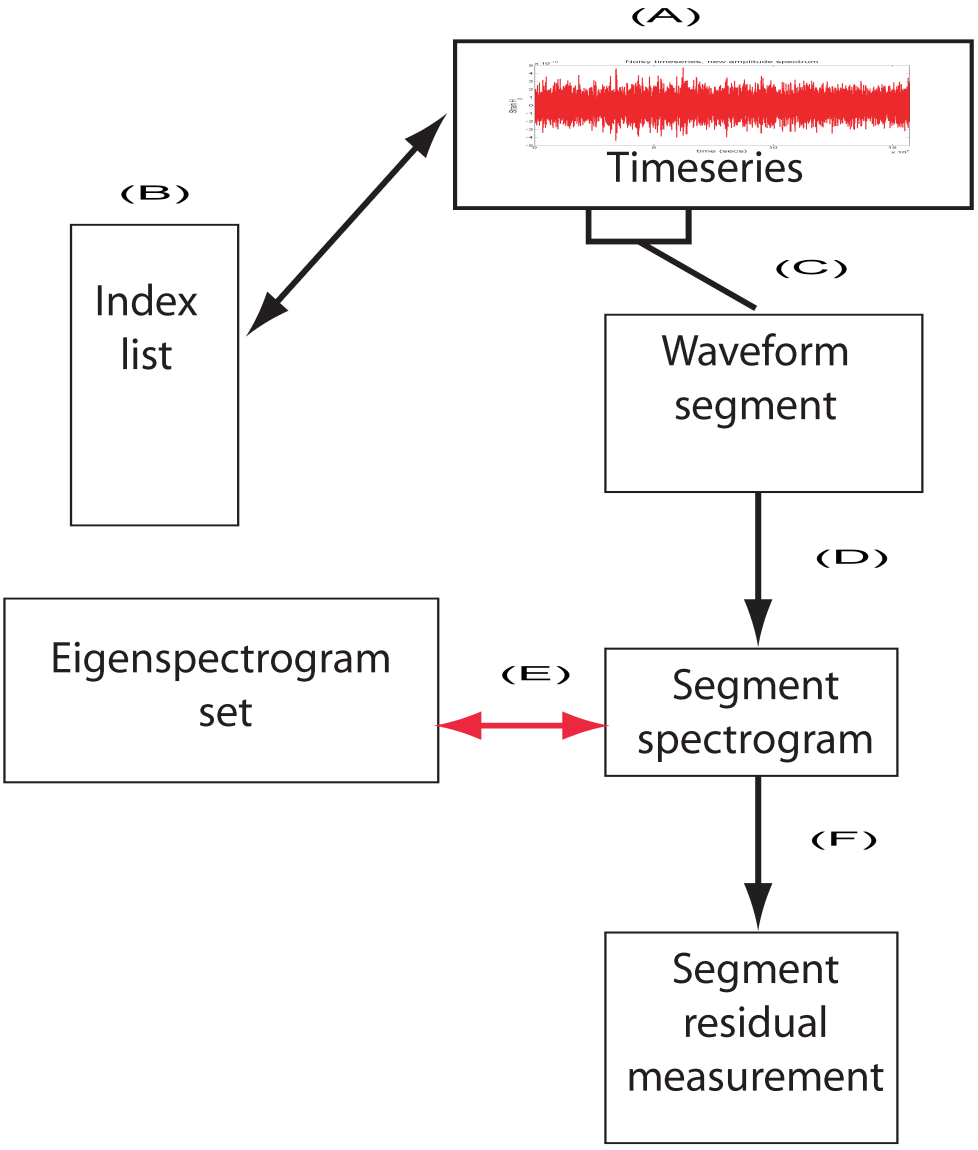
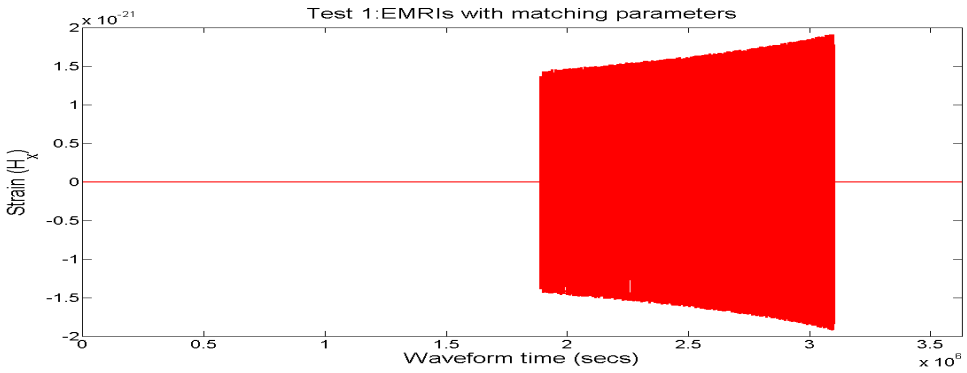


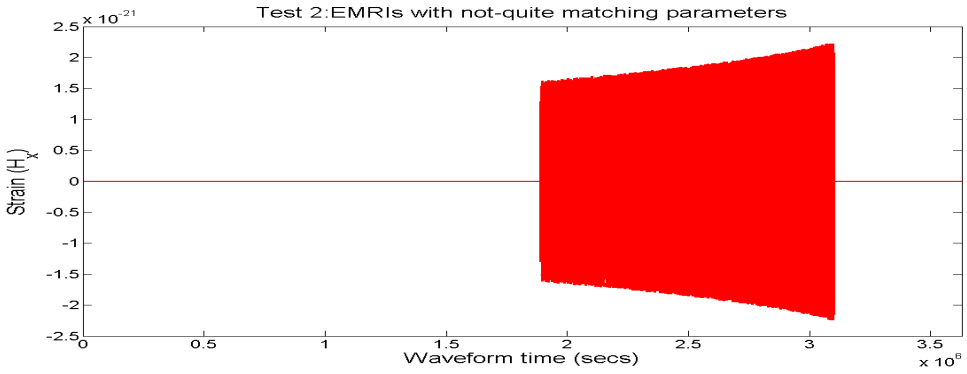
Figure 5.1: The PCA search method explained diagrammatically. Each process is explained in the main text (see Section 5.1).

Parameter	Value for defining eigenspace	Test 1	Test 2	Test 3
μ	$100 - 200M_\odot$ (with resolution of $1M_\odot$)	152	177.7	144.4
M	$1 \times 10^6 M_\odot$	$1 \times 10^6 M_\odot$	$1 \times 10^6 M_\odot$	$1.2 \times 10^6 M_\odot$
S/M^2	0.6	0.6	0.6	0.4
e_0	0.1	0.1	0.1	0.1
$\tilde{\gamma}_0$	$\pi/2$	$\pi/2$	$\pi/2$	$\pi/2$
Φ_0	$\pi/2$	$\pi/2$	$\pi/2$	$\pi/2$
θ_s	0	0	0	0
ϕ_s	π	π	π	π
λ	$\pi/6$	$\pi/6$	$\pi/6$	$\pi/5$
α	$\pi/2$	$\pi/2$	$\pi/2$	$\pi/2$
θ_k	0.1	0.1	0.1	0.1
ϕ_k	0	0	0	0
$D(pc)$	2×10^9	2×10^9	2×10^9	2×10^9

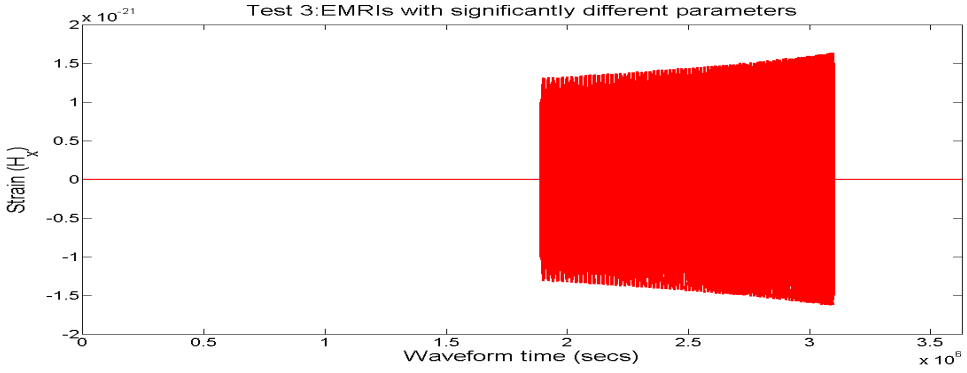
Table 5.1: The physical parameters of the EMRI signals used in Test $1-3$ in Section 5.2, and the parameters of the EMRI signal spectrograms used to generated the eigenspace used to conduct the search.



(a) Searchable waveform containing an EMRI signal with parameters matching those used to create the eigenspectrograms and hence the eigenspace.



(b) Searchable waveform containing an EMRI signal with parameters that are slightly different from those used to create the eigenspectrograms and eigenspace.



(c) Searchable waveform containing an EMRI signal with parameters that are significantly different from those used to create the eigenspectrograms and eigenspace.

Figure 5.2: Searchable waveforms containing EMRI signals.

5.2.1 Test 1:EMRIs with matching parameters

We first examine a timeseries $H(t)$ containing a single noiseless EMRI waveform, with parameters defined by Table 5.1. Specifically, the timeseries waveform is three times longer than the EMRI waveform it includes (length $L = 241920$ data points ≈ 42 days with a sample frequency of $f_s = 1/15$ Hz), and the EMRI signal begins at the 125970th data point. All other entries are zero. The waveform is shown in Fig. 5.2(a), and the result of the search in Fig. 5.3.

To search with a resolution of one data point would require 161281 search spectrograms to be constructed, which is not practical. However, because the timeseries clearly contains a signal we shall search with a resolution of ten data points, requiring 16129 search spectrograms. We are abusing our prior knowledge of the situation to some extent; a person who is ignorant of EMRI waveforms would be able to see some sort of signal in Fig. 5.2(a), but would not necessarily know that it was an EMRI signal. However, we must be clear that we are presenting the search method the simplest conditions possible in order to judge its performance under circumstances that we can control.

The search shows that the residual power from the reconstructions drops to $R = 0.0032$ before climbing once more, and by highlighting the spectrograms T_c and T_m, we see that the two are one and the same. This segement of the timeseries used to create this spectrogram actually begins with the 100001th data point, so is not an exact match for an EMRI spectrogram used to create the eigenspace, but it is very close and subsequently can be reconstructed very well. As expected, the 'NaN'spectrograms preceeding and following the EMRI signal give this search it shape, but the drop in residual measurements is unmistakable and clearly highlights the presence of the EMRI signal. The residual measurements of segments surrounding T_c/T_m do contain some structure, shown in 5.4, but the minimum point is unambiguous.

5.2.2 Test 2:EMRIs with not-quite matching parameters

Now, we search a waveform that contains a gravitational wave signal from a single noiseless EMRI waveform generated with parameters that are similar to, but not identical to, the parameters used to make spectrograms that constructed the eigenspace (the parameter are given explicitly in Table 5.1). The timeseries is the

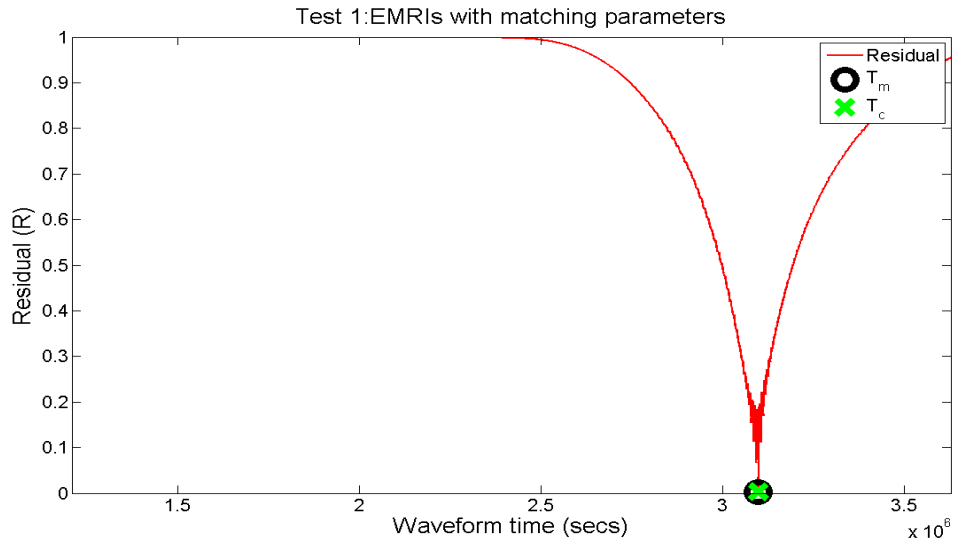


Figure 5.3: Residual measurements of the segment spectrograms of the waveform containing an EMRI signal described in Section 5.2.1. The segments T_m and T_c are the same; there is a strong suggestion that an EMRI signal has been found.

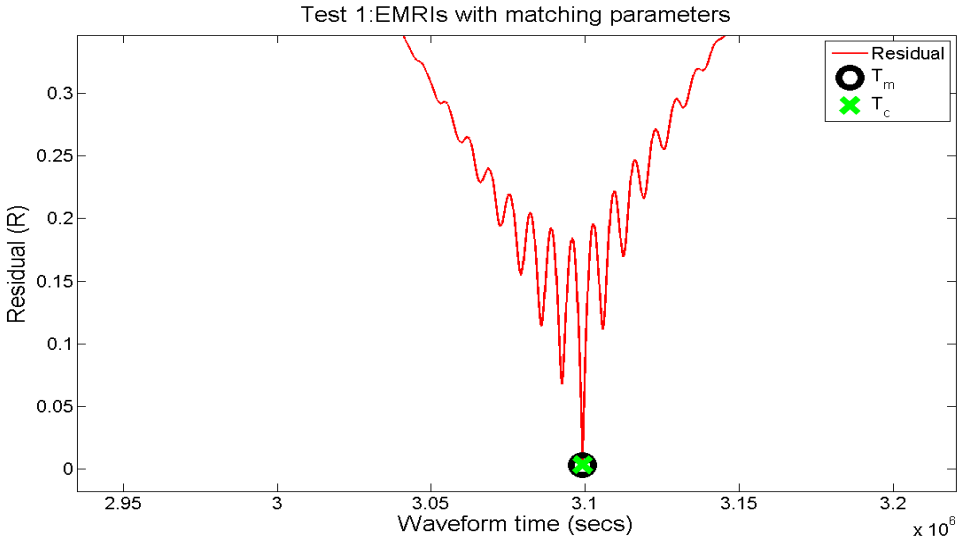


Figure 5.4: A closer look at the residual measurements around the T_c/T_m segment in Fig. 5.3. There is some structure in surrounding segments, but the most likely location of the EMRI signal is unambiguous.

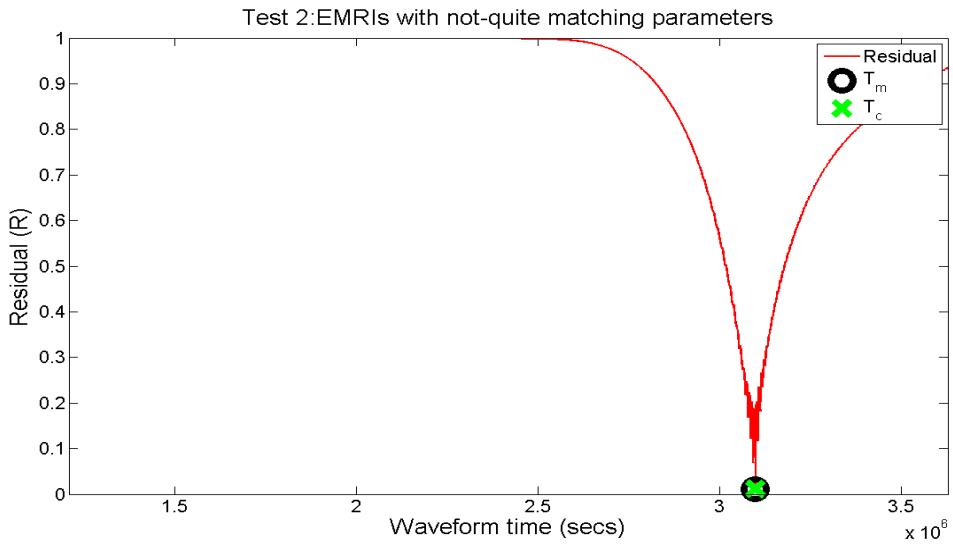


Figure 5.5: Residual measurements of the segment spectrograms of the waveform containing an EMRI signal described in Section 5.2.2. As in Fig. 5.3 segments T_m and T_c are the same.

same length as is in the previous test, again the start point for the EMRI signal is the 125970th data point, and the search resolution is ten data points.

The search produces very similar output to the first test waveform, shown in Fig. 5.5, with a clearly defined dip centred on the segments T_c and T_m which again refer to the same segment. The residual in question is $R = 0.0177$, higher than the first test but still low. Once more it is clear that the search reveals a likely EMRI signal very close to the actual signal, and again we see a similar structure to the measurements around this point (see Fig. 5.6).

5.2.3 Test 3:EMRIs with significantly different parameters

This time, we generate an EMRI from parameters that are significantly different from those used to generate the eigenspace (given in Table 5.1), but keep the start position of the signal and the resolution of the search from the previous search. Under these conditions, the search struggles, and concludes that T_c and T_m are quite widely separated; its ability to reconstruct the search spectrograms with the eigenspectrogram set available is actually best when it is not actually looking at the EMRI signal spectrogram.

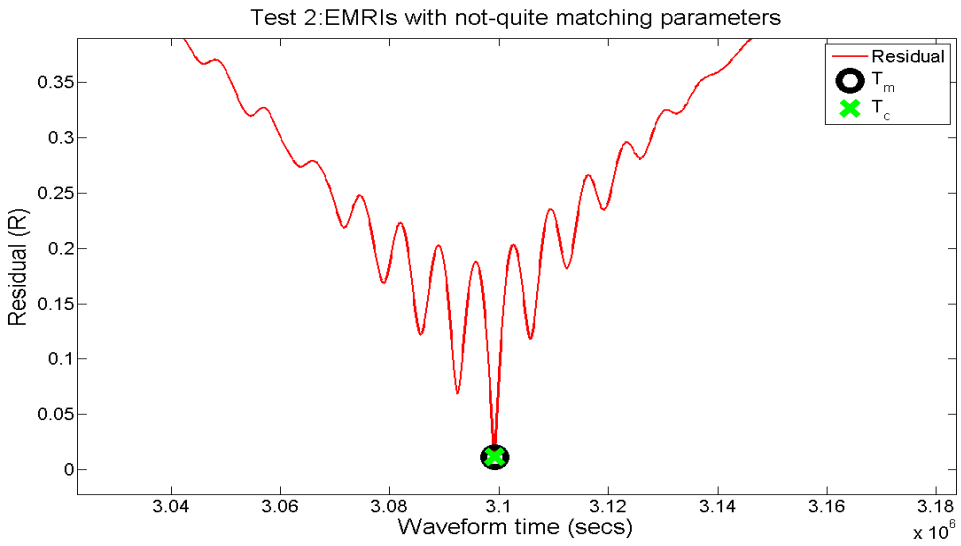


Figure 5.6: A closer look at the residual measurements around the T_c/T_m segment in Fig. 5.5. Similar to Fig. 5.4, there is some structure in surrounding segments, but the most likely location of the EMRI signal can still be seen clearly.

Furthermore, the shape of the residual measurements, given in Fig. 5.7, is rather different than in previous tests. The dip is superficially similar, but the minimum has been shifted to later spectrograms, is wider. Most telling, the residual measurements are much higher, with the minimum $R = 0.52$. Under normal conditions it is unlikely that this segment would be a viable candidate for an EMRI detection.

5.2.4 Test 4:A non-EMRI signal

Finally, two non-EMRI signals were searched. This was mainly to assert that a search could be performed on pretty much any timeseries as long as it was in the correct format, and partly out of curiousity to see how the search would perform on unusual timeseries.

The first test (test **4A**) replaced an EMRI signal with some random strain measurements, with amplitudes of the same order of magnitude to that of previous test EMRI signals ($\approx 1 \times 10^{-21}$), and is shown in Fig. 5.9(a). This was not intended to be a representation of realistic noise, nor even a signal with any physical basis, but simply a quick method of generating a waveform that looked nothing like an EMRI. As the search results show (Fig. 5.9(b) and Fig. 5.9(c)), the search does not provide

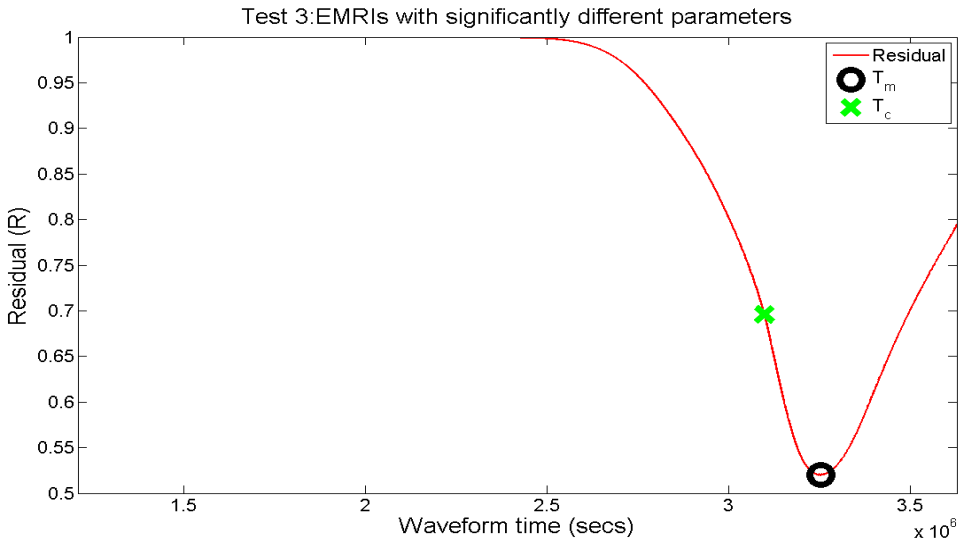


Figure 5.7: Residual measurements of the segment spectrograms of the waveform containing an EMRI signal described in Section 5.2.3. Unlike previous searches, T_c and T_m are not the same segment, and the shape of the slope is different. The residual measurements for this search are much higher as well.

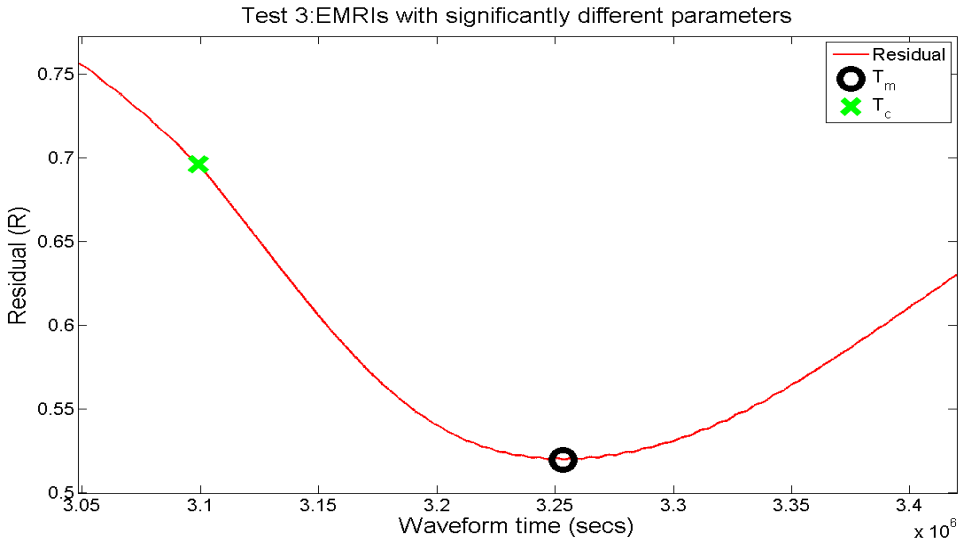


Figure 5.8: A closer look at the residual measurements of some segments in Fig. 5.7. The fine structure found in other searches is missing here.

any reasonable candidates for detection. Even the apparent structure in the search is the result of a very small differences in the residual measurements; the segment that is designated T_m here has a residual measurement of $R = 0.99995$.

Next (test **4B**), an EMRI signal was produced but then 'flipped', so that it appears as an EMRI evolving backwards in time (see Fig. 5.10(a)). Again, there was no real physical basis for this signal, but it proved a useful check that structure that shows evolution of the EMRI signal is built in to the eigenspectrograms. The search produced high residuals as expected, with a minimum of $R = 0.919$ indicating that the reconstructions bore little resemblance to the segment spectrograms.

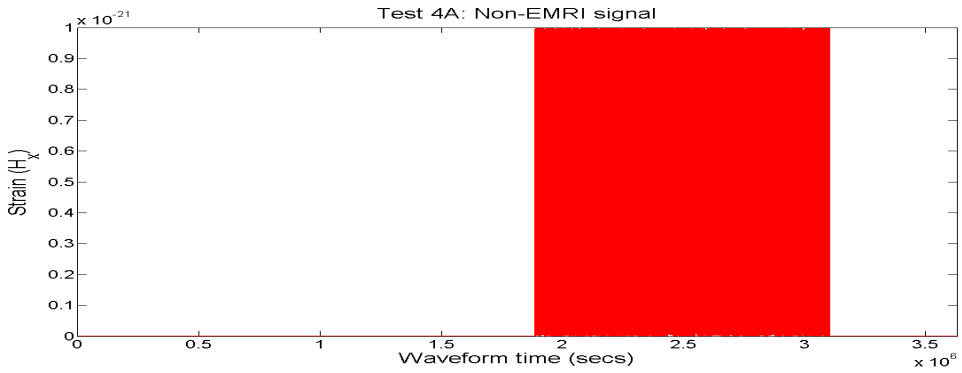
5.3 What do the simple tests tell us?

It is important to be frank with ourselves when looking at the results of the searches in Section 5.2; none of the timeseries analysed are 'realistic'. They do not take into account effects such as noise sources or LISA antenna patterns, presenting themselves only as a single isolated signal. Thus the searches cannot yet be said to indicate how adept the search method itself is at searching for EMRI signals, but can still provide a wealth of useful information which we can build on. Naturally, future work would include a realistic estimate of the noise in a timeseries of LISA data.

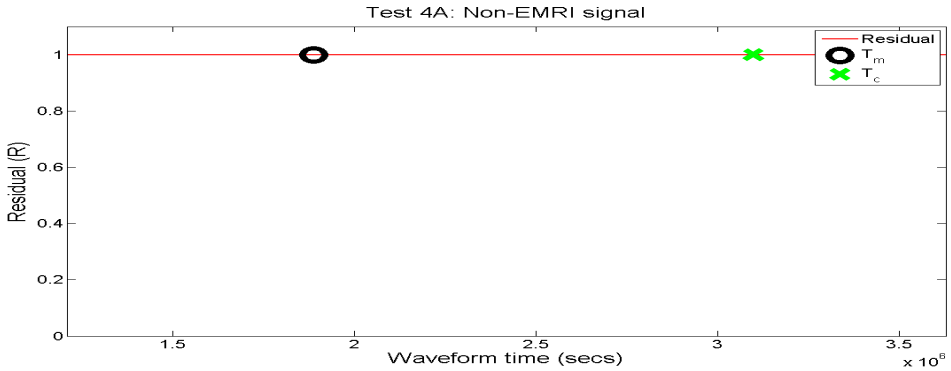
5.3.1 A searchable waveform

Creating a searchable timeseries itself is straightforward, but actually conducting the searching requires more thought. Rather quickly, the number of spectrograms that must be searched becomes very large if the search resolution is small (and hence computationally expensive). Thankfully, it is not necessary to process each one in turn; once the starting/end point indexes are determined, different segments can be examined individually, and could be sent to separate processors for analysis before the results were collected.

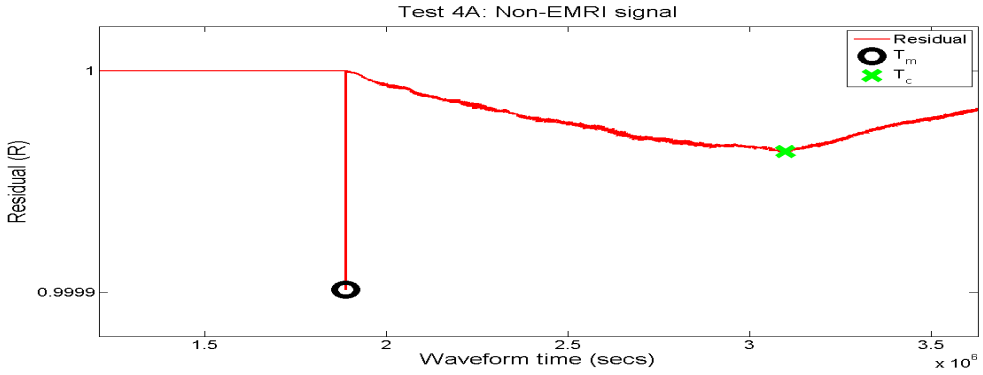
Still, in a situation where we had to analyse a years worth of data the computational cost is not insignificant. Despite this, in its current form the search of each spectrogram is fast enough to generate residual measurements in real-time (assum-



(a) A searchable waveform containing some random strain measurements. The signal is not intended as realistic noise, but simply to create a waveform that is unlike any EMRI waveform.

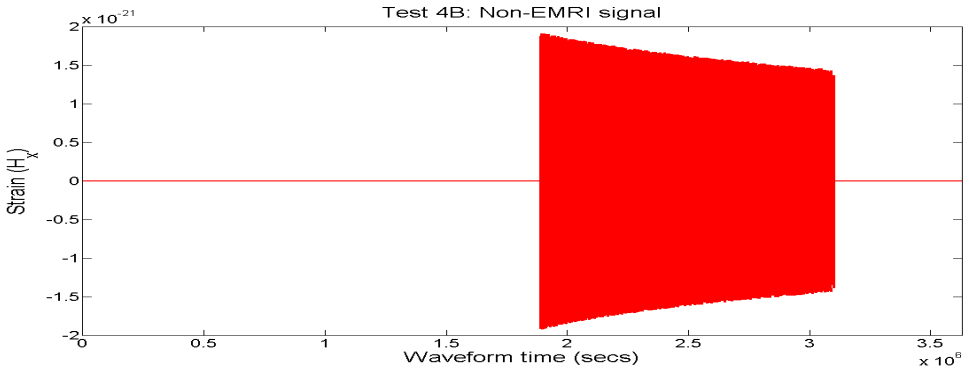


(b) Residual measurements of the waveform above. There is some structure, but it is not apparent here. Note that the segments T_c and T_m are widely separated.

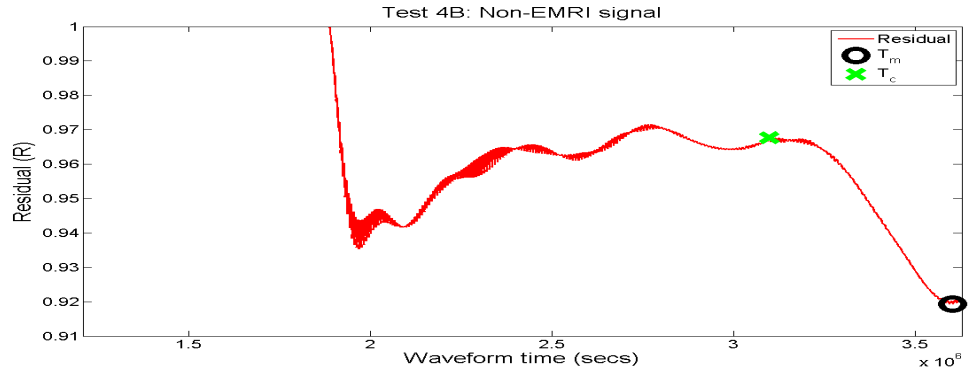


(c) A much closer look at the residual measurements. There are tiny variations present, but they do not reveal any likely candidates for EMRI detection.

Figure 5.9: Non-EMRI signal **4A** search results.



(a) A searchable waveform containing a 'backwards-evolving' EMRI signal.



(b) Residual measurements of the waveform above. They display some interesting structure, but none are low enough to be considered possible detections. Note that the segments T_c and T_m are widely separated.

Figure 5.10: Non-EMRI signal **4B** search results.

ing $f_s = 1/15$ Hz) with a search resolution of 1 data point[1] although this would require a timeseries that updated in real-time too. Nevertheless, the test timeseries typify those that we must examine for the moment; apart from our own constructions, the MLDC datasets will produce year-long timeseries and data from LISA will cover similar periods of time, but the approach remains unchanged.

The search resolution can change things fairly dramatically, however. Fig. 5.12 shows the effect of changing the search resolution on the residual measurements for the waveform created in Section 5.2.1. Only the search with a resolution of one data point really finds the tiny residual power we expect from an EMRI signal with the parameters that were chosen ($R \approx 1 \times 10^{-14}$), even though larger resolutions do produce a clear indication that the residual alters significantly around a particular time. Taken to an extreme, where the entire timeseries is only divided into a few segments, the shape does become harder to see and the residual power measured is higher even for the best matching segment. Table 5.2 converts a number of search resolutions into approximate times based on a sample frequency of $f_s = 1/15$ Hz as an aid.

Nevertheless, it can be seen in Fig. 5.12 that the shape of the search results from the waveform Section 5.2.1 is largely unchanged when considering the ten data point and one data point resolutions even when looked at closely. Indeed, because of the computational cost of searching the waveform with a resolution of one data point, the ten-data-point search was used to find a minimum residual point, and the timeseries surrounding this position were then re-searched at the smaller resolution. While there was a risk that a different structure would be revealed by this closer search, it does not appear to be the case in this situation. It must be kept in mind that there is no guarantee such a situation will occur in other waveforms, especially those where the signal is not so obvious.

Thus we must balance the computational expense of the searches and the quality of the results they produce. The test timeseries of Section 5.2 all revealed their shape and character with a resolution of ten data points (limited tests with a resolution of one data point revealed very little change in the results), but it is clear that to extract the most information each data point must be searched.

[1] This is of course dependent on the number of eigenspectrograms used in the reconstructions, but even an eigenspace spanned by many thousands of eigenvectors would still be processable in near real-time.

Search resolution (data points)	Search resolution conversion
1	15 seconds
10	21/2 minutes
100	25 minutes
1000	≈ 4 hours
10000	≈ 42 hours

Table 5.2: Conversion of various search resolutions into approximate equivalent timescales for quick reference, based on a sample frequency of 1/15 Hz.

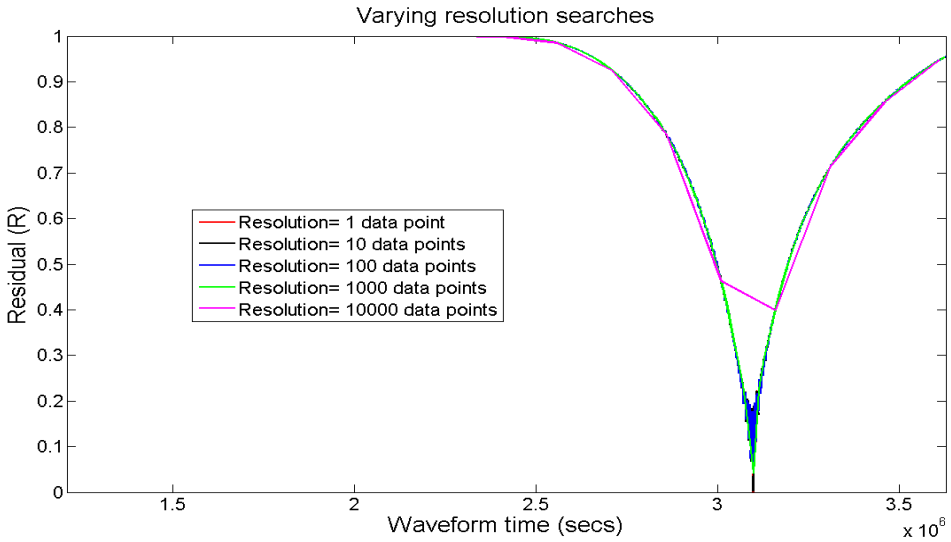


Figure 5.11: The effect of varying the search resolution on residual measurements for the waveform described in Section 5.2.1. Only a resolution of 1 data point generates the residual measurements where the segment will match a spectrogram used to generate the eigenspace.

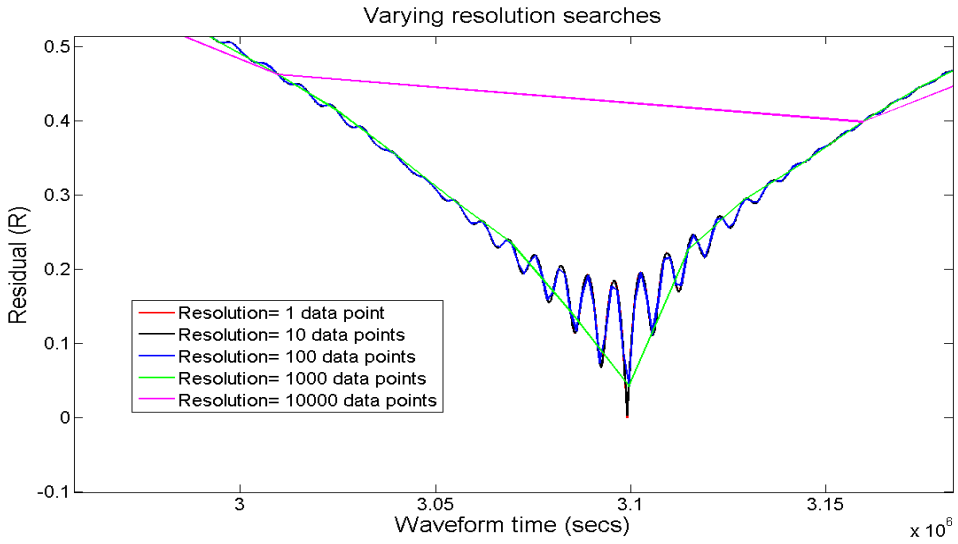


Figure 5.12: A close look at the residual measurements from searches of varying resolutions on the waveforms described in Section 5.2.1. The general shape is revealed with a search of 100 data points, with only slight differences generated by the 10-data-point and 1-data-point searches.

5.3.2 Signal placement

The placement of the EMRI signal within the timeseries can be important too. Only a search resolution of one data point is guaranteed to overlap with the EMRI signal completely but if this is not practical then we risk not having sufficient overlap to produce a good quality reconstruction revealed by a small residual measurement. All the same, the search is fairly robust in that it tolerates the placement of the EMRI signal waveform pretty much anywhere within the timeseries; as long as it is entirely within the timeseries, the output produced from the segments covering that period will be similar.

This principle is demonstrated in Fig. 5.13 where the EMRI signal from Section 5.2.1 is moved around the timeseries and searched with a search resolution of ten data points. There are a number of placements considered:

A The original positioning, producing exactly the same results as detailed in Section 5.2.1.

B The EMRI signal begins right at the start of the timeseries. The drop in residual

power is absent, since there are no segments leading up to the signal but the recovery is present as the segment contain progressively less of the EMRI signal. The minimum residual is the first segment of the signal, matching the spectrogram exactly, with $R \approx 8 \times 10^{-14}$.

C The signal begins at the 125975th entry, so does not exactly match the spectrogram from any segement. Despite this, the residual measurements are not changed significantly, with a minimum of $R = 0.0075$. Fig. 5.13 shows that plotting these values over those of placement **A** obscures the latter. Fig. 5.15 shows the differences in residual measurements between placement **C** and **A**.

D The EMRI signal is only partially present, it does not lie entirely within the timeseries being searched. Here we see the drop in residual power, but not the point at which the segment and the EMRI signal overlap exactly, nor the recovery. The minimum residual here is $R = 0.861$.

It seems that we must return to the search resolution problem, since the proportion of the waveform captured by any segment depends on the resolution and placement together. Again, the ideal solution is a one-data-point search resolution of a timeseries that is continually updated. This eliminates that concern that residual measurements are high simply because there isn't the correct overlap between the segment being examined and the signal, and mitigates the computational cost of processing long timeseries each time. Naturally, placements **B** and **D** are difficult to deal with, but are unavoidable depending on what period the timeseries covers; the operational time of the LISA mission and what EMRI signals occur during this time are inevitable constraints on the signals that this search method can potentially detect.

5.3.3 Waveform duration

As stated the spectrograms being examined must be generated in the same manner as those used to create the eigenspectrograms that define the eigenspace, but this generation can be changed. Thus far we have not repeated the eigenspace generation or searches using EMRI waveforms of different durations, but it is feasible and may have a significant effect.

This topic is explored further in Section 6.6, but the most immediate concern is that we are currently considering signals that begin and end abruptly. The duration

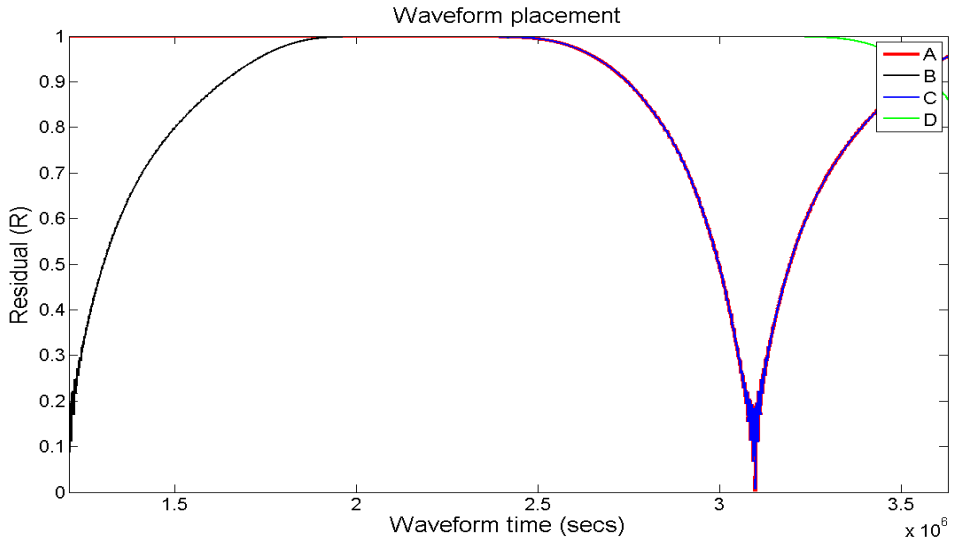


Figure 5.13: The effect of signal placement within a searchable waveform on the residual measurements generated, as described in Section 5.3.2. In particular, placements **A** and **C** are indistinguishable in this plot.

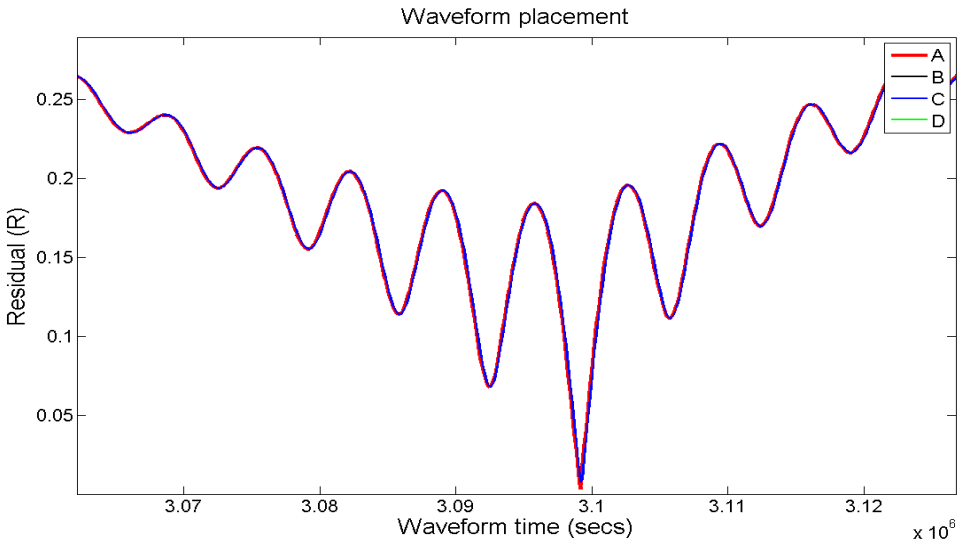


Figure 5.14: A closer look at the search around the minimum residual measurements generated by signal placements **A** and **C** from Fig. 5.13. Tiny differences are visible, but the shape of the residual measurements for both placements is very similar. The small offset between the two signal placements does not affect things to any great degree.

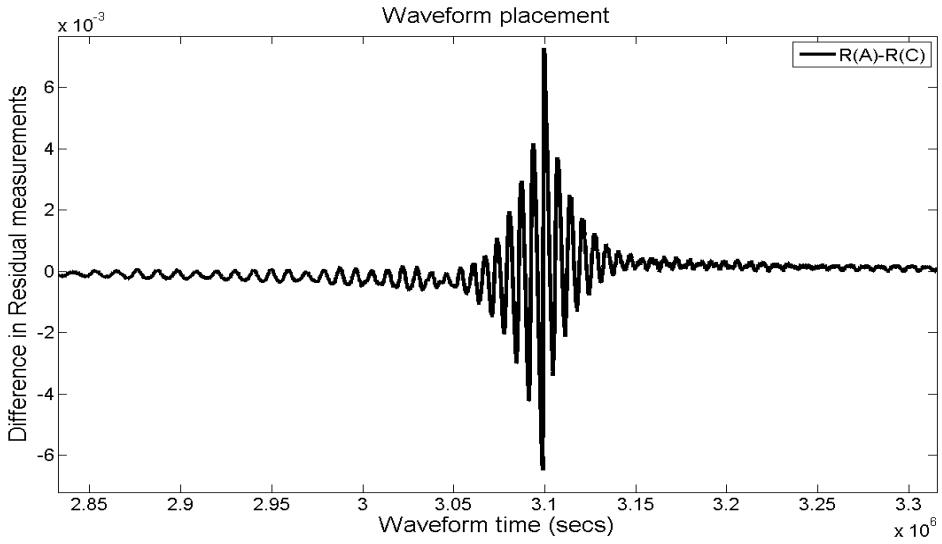


Figure 5.15: The difference between the residual measurements generated by signal placements **A** and **C** from Fig. 5.13, around the time of the minimum residual measurements. The slight differences suggest that the difference between signal placement does not affect things greatly.

of an actual EMRI waveform will be much longer than the searchable waveforms we are constructing, and in reality EMRI signals early on in their evolution may be present (and will not coalesce during the waveform). This will doubtlessly affect the shape of the search residual measurements, making a sharp near-symmetric shape such as that produced by the tests in Section 5.2.1 and 5.2.2 less likely.

5.3.4 Final remarks

We must address the question of what the search method is actually looking for. As mentioned, none of the waveforms produced represents realistic challenges for the search, so it is hard to infer its behaviour under different circumstances from the initial tests. These issues are addressed in Chapter 6, but it is important to be clear on a few points before proceeding.

Although we have stated that tackling the EMRI waveforms themselves in a traditional manner might not be computationally feasible, we have not yet proven that the spectrogram/eigenspectrogram approach is the best one. Certainly, when looking at waveforms such as those in Fig. 5.2, anyone would be able to identify the

EMRI signal with little prompting. The searches of these waveforms do produce different results though, in particular the one featured in Section 5.2.3. Furthermore, the inability of the search method to reconstruct the non-EMRI signals is encouraging. Therefore, we can *cautiously* proceed with the knowledge that in simple situations the search should be able to distinguish between EMRI and non-EMRI signals, and the makeup of the particular EMRI signal will affect the search results.

So far so good, but what should we count as a detection? Clearly selecting the segment T_m will not always work even when an EMRI signal is present, and is not a helpful guide when the signal is non-EMRI. To complicate matters, our residual measurements to not produce 'spikes' with a sharp drop off that can be easily isolated, but are more likely to produce slopes. For these reasons, we will will delay defining 'a detection' until we have a better grasp on how the search behaves under more realistic conditions. Although it may seem easier to make such a definition at this stage, we do not want to prematurely assume anything based on the very simple searches performed so far.

Chapter 6

Further searches for EMRI signals

Preliminary PCA-based searches for EMRI signals showed promise under constrained, largely idealised circumstances, but their usefulness can only really be assessed under more realistic conditions. This chapter identifies some major complications to the types of signals and searchable waveforms examined in the previous chapter, considering the effects of each on the search for the EMRI signals separately and in combination. In doing so, the difficulty of identifying the signals will increase, but so to will the relevance of the searches in the detection of gravitational waves.

6.1 Waveform lengths and a searchable waveforms

Previous searchable waveforms have only contained EMRI signal waveforms of the same duration as those that are being searched for (and equivalently, the same length as each segment being evaluated). Now that we are attempting to construct more realistic tests, we must acknowledge that the EMRI signals will not 'switch on' in this manner, and that we should include their evolution leading up to this $\approx$ 2-week-long section that we have previously concentrated on.

The increased complexity of the waveforms that will be searched increases the possibility of confusion. In order to ensure that things remain comprehensible, we need unambiguous definitions for some of the concepts or parameters required. Simplified versions of these definitions are also included in the Nomenclature section preceding the main body of this thesis for reference.

'Searchable waveform' A gravitational strain waveform that is the target of the search for EMRI signals. The waveform is represented by a timeseries of grav-

itational $H_\times$ strain measurements, sections of which are manipulated further (being turned into spectrograms and projected into the eigenspace). We replace the old convention of simply referring to the timeseries $H(t)$ because the searchable waveform may be the combination of several timeseries, typically multiple EMRI signals and noise.

'Time from coalescence' In the event that the EMRI signal is present but difficult or impossible to distinguish by eye within the searchable waveform, a measure of the time before (negative values) and after the coalescence time (positive values) is useful. The previously-defined T_c compliments this measurement, but we can avoid having to define the signal in terms of which spectrogram incorporates the signal (this can be useful when showing multiple residual measurements on the same plot).

A new set of eigenspectrograms has been constructed and will be used for the remainder of this chapter. Table 6.1 details the EMRI spectrograms set used to create the eigenspace being examined (we will name this set Υ), while each searchable waveform will have 2^{20} data points with a sample frequency of 1/15 Hz, corresponding to approximately six months of data. EMRI waveforms that have been added to the searchable waveform were evolved backwards from the coalescence time for 2^{20} data points as well; as long as the coalesence falls within the extent of the searchable waveform's *waveform time*, we can avoid the signal 'switching on'. Naturally, gravitational strain from any EMRI signals that would fall before the beginning of waveform time are ignored (for example the searchable waveform shown in Fig. 6.1 only reveals approximately three months of strain measurements from an EMRI signal, since the evolution further back in time would correspond to times $t < 0$).

The searches being conducted are still analysing the approximately-two-week-long gravitational wave signal from an EMRI, regardless of the duration of the searchable waveform. Fig. 6.1 shows part of an EMRI signal subsequently referred to as ξ, generated with parameters given in Table 6.1 and inserted with a coalescence time (in waveform time) of 7869600 seconds (approximately three months from the beginning of the waveform) to create a searchable waveform. ξ will be used extensively for demonstration purposes throughout this chapter, and unless otherwise stated, will be inserted into a searchable waveform with a coalescence time $t_c = 7869600$ seconds. The particular parameters of ξ were chosen at random, constrained only by the fact that they must match those of one of the spectrograms

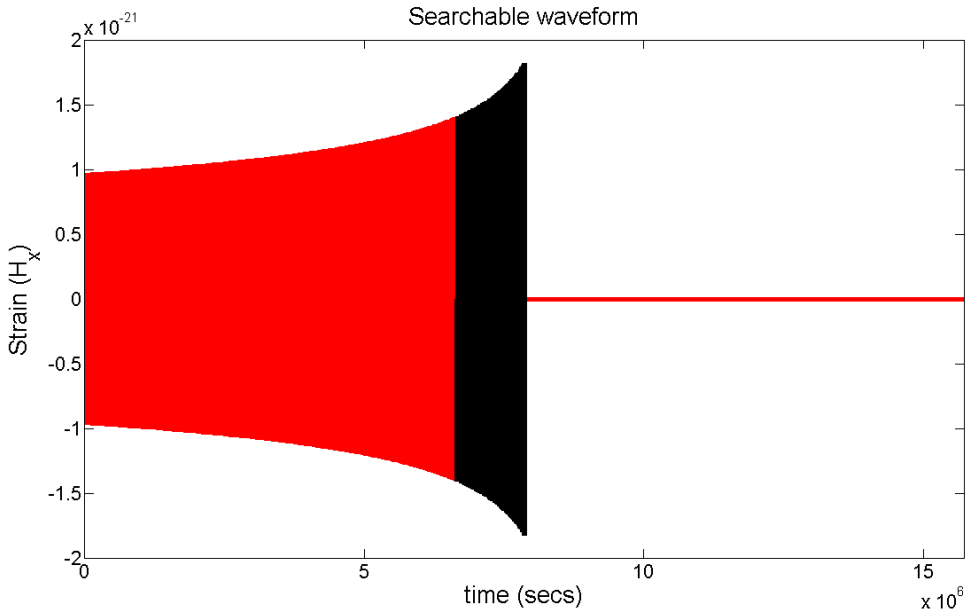


Figure 6.1: A searchable waveform containing an EMRI signal ξ. The coalescence occurs at waveform time 7869600 seconds (approximately three months from the beginning of the waveform). The segment that will produce a spectrogram belonging to the set Υ is marked in black.

in the set Υ, albeit evolved backwards in time for six months. The extent of the signal matching one of the Υ spectrograms within ξ is also highlighted.

The increased length of the waveforms being searched means that searches naturally take longer to process. Keeping in mind the results displayed in Fig. 5.12, the default search resolution throughout this chapter will be 100 data points. This represents a compromise between the ability to perform searches within a reasonable time (on a desktop computer a search of this type will take approximately 20 minutes) and the ability to pick out fine structure within the residual measurements of the searches. As before, the resolution can be changed at will if deemed necessary.

6.2 Noisy EMRI signals

The most obvious shortcoming of previous searches is that the EMRI signal is the only gravitational strain present in the searchable waveform. There, we tacitly supposed some flawless detector that was able to measure gravitational strain without

Parameter	Value for defining eigenspace	ξ
μ	$100 - 150M_\odot$ (with resolution of $1M_\odot$)	$140M_\odot$
M	$1 \times 10^6 M_\odot$	$1 \times 10^6 M_\odot$
S/M^2	0.6	0.6
e_0	0.01	0.01
$\tilde{\gamma}_0$	$\pi/2$	$\pi/2$
Φ_0	$\pi/2$	$\pi/2$
θ_s	0	0
ϕ_s	π	π
λ	$\pi/6$	$\pi/6$
α	$\pi/2$	$\pi/2$
θ_k	0.1	0.1
ϕ_k	0	0
D(pc)	2×10^9	2×10^9
Waveform duration (secs)	1209600	15728640

Table 6.1: The physical parameters of the EMRI signal spectrogram set Υ used throughout Chapter 6 to generated the eigenspace used to conduct searches. The specific parameters of the EMRI signal ξ is also given, since it will be used extensively.

any noise whatsoever, but this is not a useful way of approaching the problem; we must get to grips with likely sources of noise that will alter the signals in order to test the effectiveness of the PCA-based searches.

6.2.1 Adding noise to searchable waveforms

A standard LISA sensitivity curve (given by (W..q)) was used to generate noise timeseries, to which EMRI waveforms could be added. This process and a description of the noise sources present is rather involved and is given in detail in Appendix A in an attempt to keep matters in this section straightforward.

6.2.1.1 Searching the noise

Timeseries of LISA's noise have large contributions from low-frequency noise (see Fig. 6.2), producing strain measurements that look very different from an EMRI waveform (see Fig. 6.3). However, before adding EMRI signals we conduct a search of two noise-only timeseries' (denoted SW1 & SW 2), serving much the same function as the non-EMRI signal searches described in Section 5.2.4 but with a more instructive focus; the performance of the searches when there are no signals will produce a sort of baseline with which we can judge how large an effect an EMRI signal will produce and whether such an effect can be deemed 'significant'.

The searches themselves are performed in exactly the same manner as those in Chapter 5, and the residual measurements show that noisy segments are far removed from the eigenspace (see Fig. 6.4). In fact, the residual measurements are all so close to $R = 1$ that we define another parameter (see Fig. 6.4(b))

$R_2 = 1 - R$. When $R \approx 1$, this can be used to reveal structure in residual measurements. Note that as R increases, R_2 decreases and vice versa; searches highlighting a minimal value of R in a searchable waveform will highlight a maximum R_2.

These results are actually encouraging because they indicate that a segment containing only noise will project into the EMRI eigenspace in a manner that is very unlike any of the EMRI spectrograms. Further, there seems to be no overall pattern to the structure revealed by the R_2 values of either search. It is of course premature to try and apply some rule such as 'ignore all R_2 measurements below 1×10^{-9}' based on these two searches, but we have established that while variations

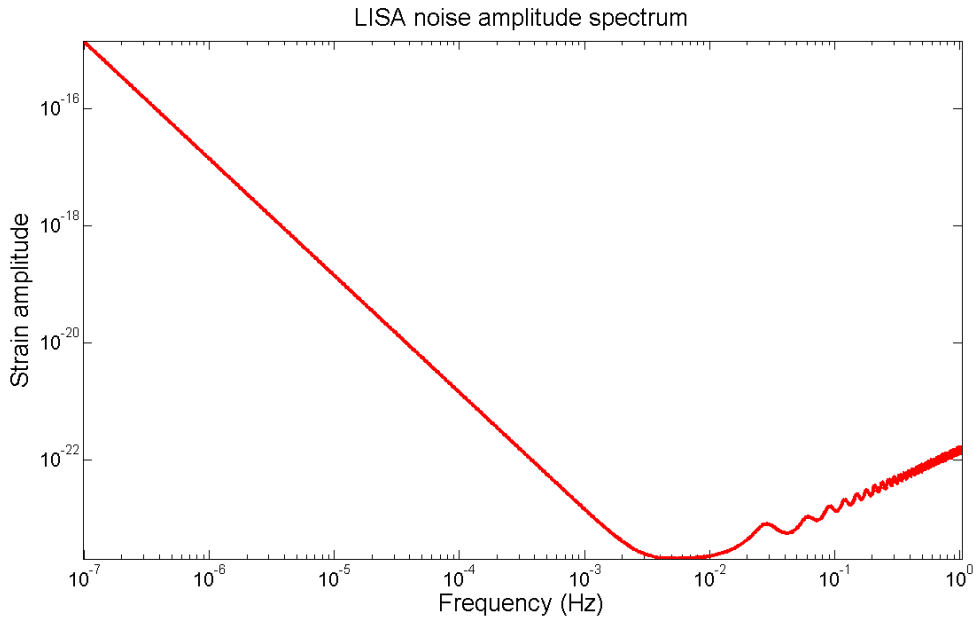


Figure 6.2: The strain amplitude spectrum of LISA noise, as given by (W..q).

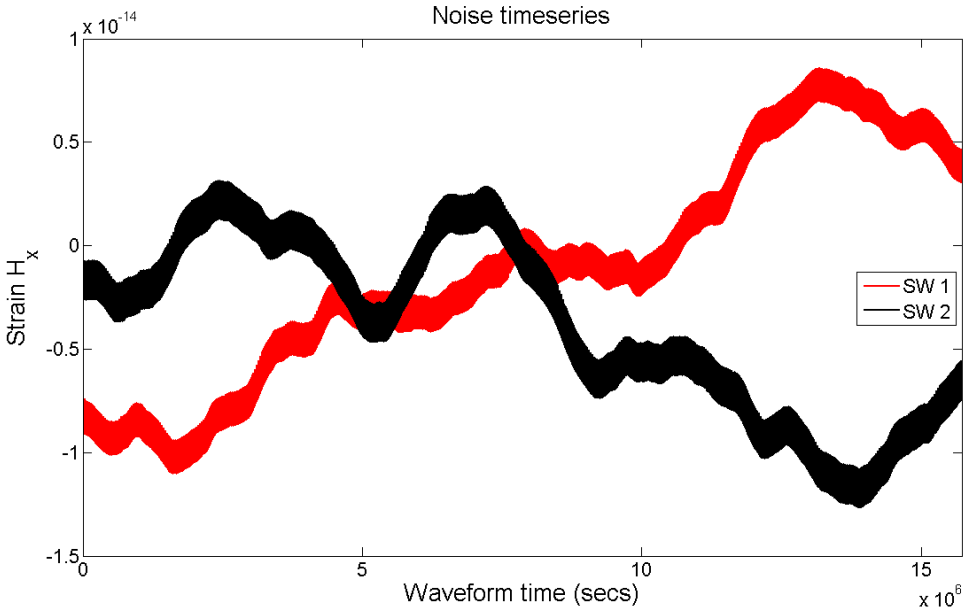
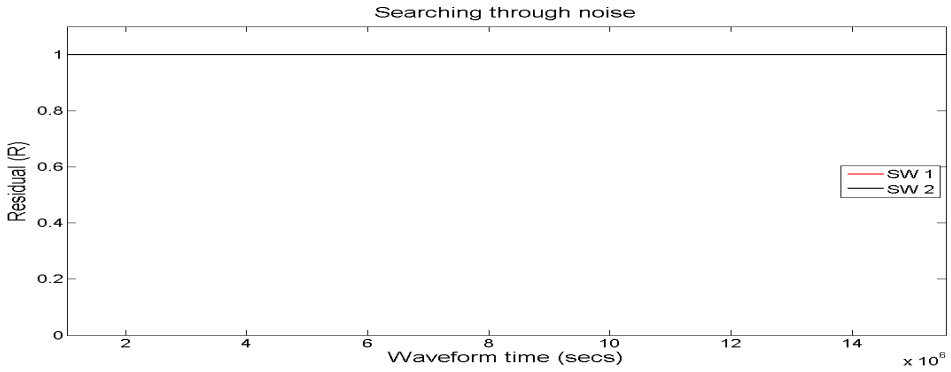
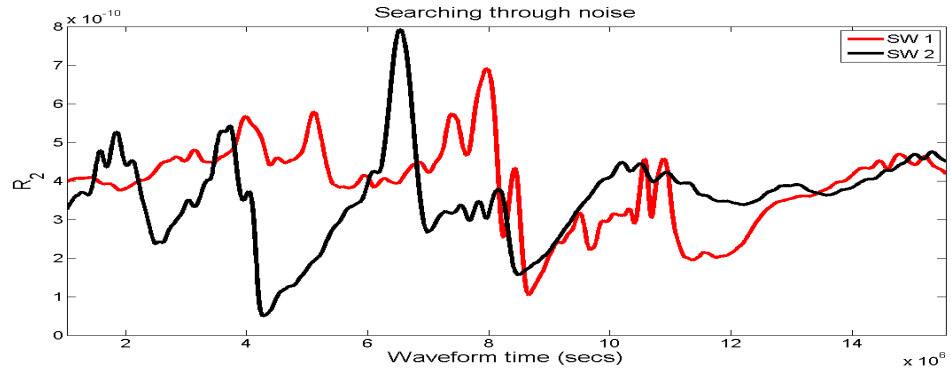


Figure 6.3: Two searchable waveforms (denoted SW1 & SW 2) constructed from the sensitivity curve given in Fig. 6.2. Neither waveform contains an EMRI signal, only noise.



(a) The residual measurements R of searchable waveforms SW 1 & SW 2 (see Fig. 6.3). On this scale they are indistinguishable and appear to contain no structure.



(b) R_2 measurements for the searchable waveforms SW 1 & SW 2. Fine structure is revealed, but the measurements do not appear to have any obviously meaningful shape.

Figure 6.4: Residual measurements of searchable waveforms SW 1 & SW 2

in the residuals occur even without the presence of a signal, they are extremely small.

6.2.1.2 The first search

Having looked at searchable waveforms containing only noise, we now add an EMRI signal to the noise and search again, hoping to find lower residual measurements around the time of the EMRI signal.

The noise waveform is exactly the same as SW 1, to which the EMRI waveform ξ has been added. The searchable waveform produced is given by Fig. 6.5, but the residual measurements do not immediately suggest the presence of a signal (see Fig. 6.6, noting that R_2 is required to highlight the structure once again), being of the same order of magnitude as those produced by the noise-only search.

The presence of an EMRI signal prompts us to use the 'Time from coalescence' to determine how well the search finds the signal, but it is clear that simply looking for the highest R_2 value does will not work, nor does there appear to be any meaningful shape preceding or following the coalescence time which might indicate the presence of a signal (at least, nothing immediately obvious). As might be expected, the R_2 measurements in the latter half of the searchable waveform are identical to those measured for SW 1, since there is no EMRI signal contribution to the strain after the coalescence.

Not unsurprisingly then, the noise has a huge impact on the effectiveness of the search, and as presented in the examples above is sufficient to completely obscure an EMRI signal. Fig. 6.7 shows a spectrogram of the segment T_c of the searchable waveform in Fig. 6.5 and it does not suggest the presence of an EMRI signal to the naked eye either.

We now attempt to mitigate some of the adverse effects of the noise in order to reveal the presence of a signal in the searchable waveforms.

6.2.2 Whitening the noise

The large contributions of low-frequency noise (which we shall classify as noise below 1×10^{-4} Hz for the moment[1]) as well as contributions from frequencies above the

[1]EMRI signals of the sort we have been investigating are typically above this frequency, but we desire a small 'buffer' in the event of different behaviour from signals selected from different parts of the EMRI problem space.

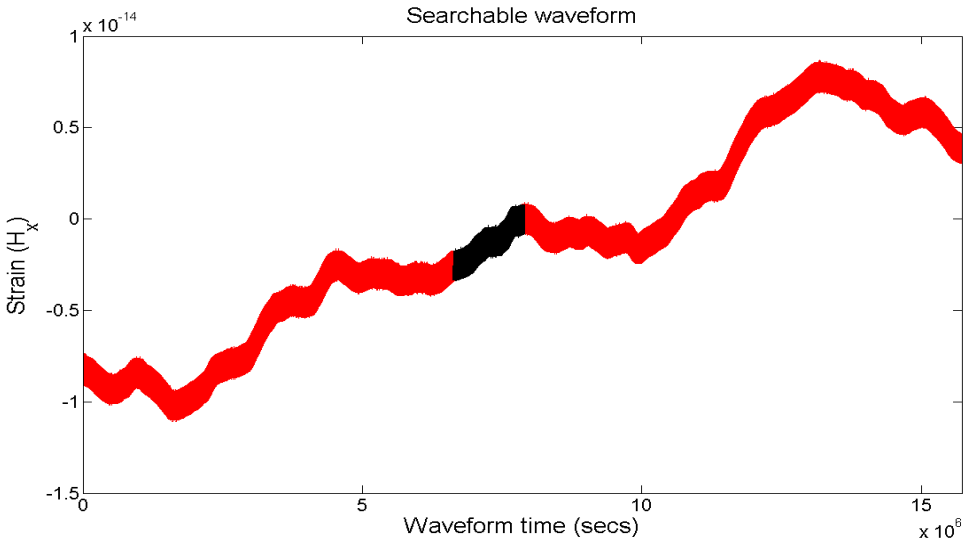


Figure 6.5: A searchable waveform containing the EMRI signal waveform ξ (the segment of the waveform containing the signal belonging to the set Υ -the 'target' of the search- is in black).

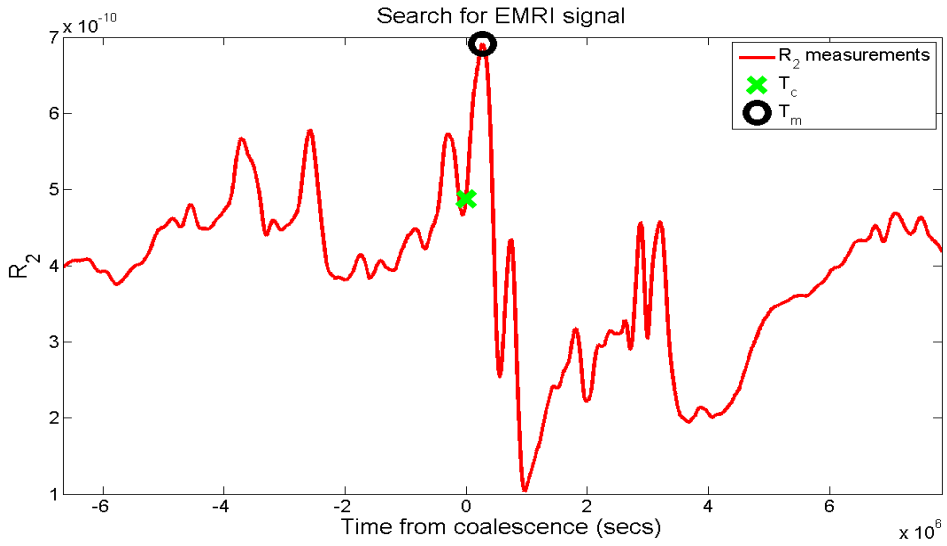


Figure 6.6: Residual measurements of the searchable waveform shown in Fig. 6.5, given as $R_2 = 1 - R$. Note that under these conditions, T_m is actually the segment containing the highest R_2 value.

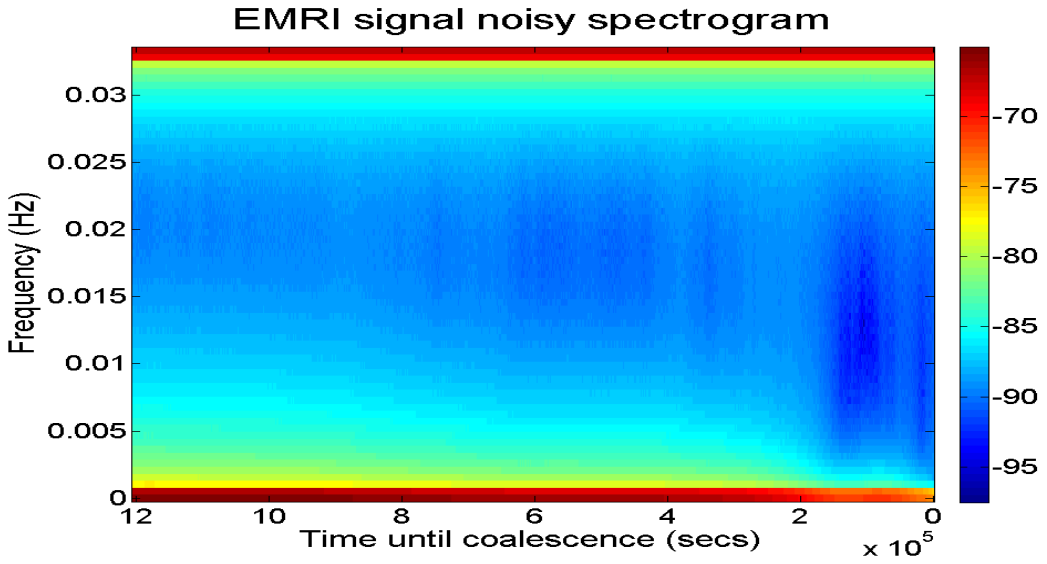


Figure 6.7: A spectrogram of the segment T_c in the the searchable waveform shown in Fig. 6.5. The EMRI signal does not stand out (the logarithm of the power is displayed in an attempt to highlight the fine structure in the signal, to no avail).

Nyquist frequency of the waveforms based on a 1/15 Hz sample frequency (that is, above 33 mHz), produces strain measurements that vary over a range that is orders of magnitude larger than the strains produced by the EMRI signal waveform.

We are not only being stymied by the magnitude of the noise but its shape as well, something that can be seen easily by looking at typical EMRI spectrograms (such as those in Fig. 4.40) and the strain amplitude spectrum of LISA noise (Fig. 6.2); thankfully the signal we are looking for is strongest at frequencies free from the low or high-frequency noise. However, in terms of principal components of the eigenspace, no eigenspectrogram will be able to account for the low-frequency noise, but the power will still overwhelm the signal itself. As a result, residual measurements will remain high (as we have seen).

Rather than try and cut out the noise, it might make more sense to approach the problem in a similar manner to the SMBH inspiral signal search presented in Section 2.2.1 and whiten the noise in each segment being searched. Armed with what we believe to be a good representation of the noise, we might be able to higlight an EMRI signal in a searchable waveform.

This approach requires only a slight modification to the existing search method.

When the spectrogram of a segment is constructed, the power in each frequency and time bin $P_{\mathrm{s}}(\nu, t)$ is recalculated according to

$$P_{\mathrm{s}}(\nu, t) = \frac{P_{\mathrm{o}}(\nu, t)}{\langle P_{\mathrm{noise}}(\nu) \rangle}$$

where $P_{\mathrm{o}}(\nu, t)$ refers to the original segment.

In order to effectively use the whitened signal, we must take care to compare like with like. Therefore, we construct a new eigenspace using the set Υ but whitening each spectrogram (with the first L measurements of waveform **A**, described below). There are still 50 unique spectrograms by virtue of the 50 unique spectrograms, hence we still produce a set of 50 eigenspectrograms, although they do look different from the usual eigenspectrograms (see Fig. 6.13). After the eigenspace is generated though, the projection of new whitened spectrograms into the eigenspace and their reconstruction proceed as normal.

Once again, we look at a couple of searchable waveform containing only no noise in order to gauge the typical behaviour the search when there is no EMRI signal present. Fig. 6.8 shows two searchable waveforms that were used to investigate this behaviour, in three different ways, and Fig. 6.9 shows the residual measurements of the searches, each of which requires some explanation.

A The searchable waveform corresponds to **A** in Fig. 6.8, and each segment spectrogram is whitened by a spectrogram constructed from the first L measurements of that same waveform. We are free to do this because there is no EMRI signal contributing to the searchable waveform, but it is an unusual situation in that we are using some of the noise to whiten the noise itself. However, because we are using the mean noise power at each frequency based only on the content of the first L measurements, there will still be variations in subsequent test spectrograms; they will not be perfectly whitened.

B The searchable waveform now corresponds to **B** in Fig. 6.8, but the test spectrograms are whitened according to the first L measurements of the searchable waveform **A**. This allows us to check how variations in the residual measurements change with different searchable waveforms but the same whitening source.

C Finally, the searchable waveform **B** in Fig. 6.8 is used again and we change the whitening source to another L measurements from the searchable waveform **A**

(specifically, the measurements between 6660015 and 7869600 seconds in waveform time). With this arrangement we can double check the effect of keeping the searchable waveform constant but whitening with a different source.

The residual measurements obtained from from the arrangements **A**,**B** and **C** are still high, but display variations that are noticeable without resorting to calculating R_2 equivalents. **A** in particular shows some strong features, but there is no obvious pattern to them, as we might expect given that the noise does not contain any other structure other than as a result of its amplitude spectrum.

On the other hand, both **B** and **C** display one main feature towards the end of the measurements, but again their similarity could have been foreseen since they are the results of testing the same waveform but with different whitening sources. The difference between the two in terms of residual measurement is not as great as the features themselves, nor does it compare to most of the features seen in **A**.

It appears then that whitening the test spectrograms does impact the residual measurements greatly, producing variations that are many orders of magnitude larger than those in similar searchable waveforms that are not whitened. Furthermore, the differences between the searchable waveforms themselves are more significant than those produced by using different sources to whiten test spectrograms (to within reason; the notion of using radically different whitening sources is discussed in Section.6.2.6).

Again, applying some blanket rule that would just ignore the features seen in Fig. 6.8 is overly hasty, but we must be prepared to accept that if the test spectrograms are whitened then there will be variations in the residual measurements of a searchable waveform that have nothing to do with an EMRI signal. Only if the addition of a signal produces rather different results will we be able to infer its presence using this search method.

The next logical step is to add an EMRI signal to the noise and search for it. Three new noisy timeseries were generated, and ξ added to each to create three new searchable waveforms (labelled in Fig. 6.10 simply as **1**, **2** and **3**), before the usual search method was applied. The same whitening source used to create **A** in Fig. 6.8 was used for each searchable waveform.

Fig. 6.10 reveals residual measurements that show seemingly random structure for each searchable waveform. There is no apparent pattern to the size of the

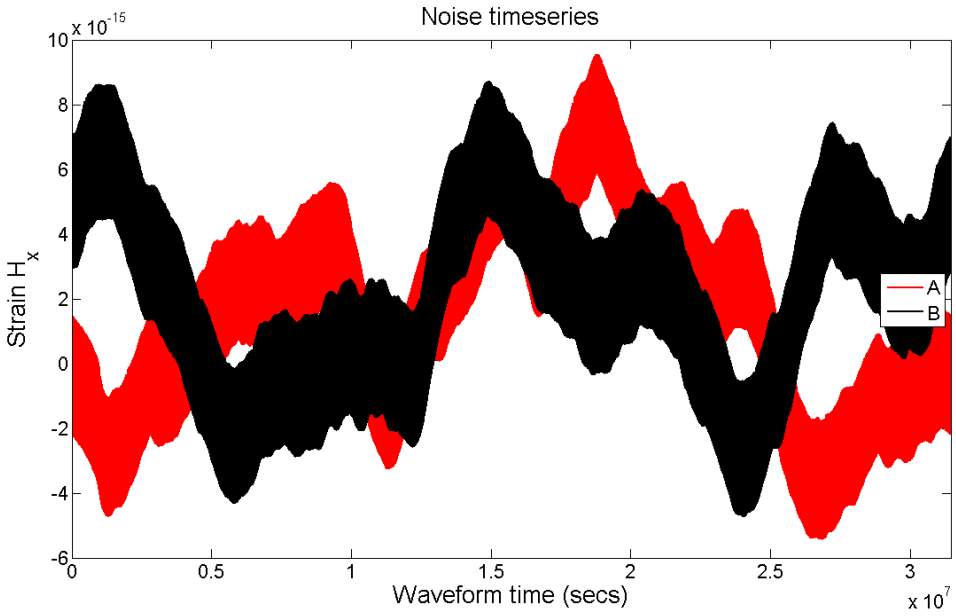


Figure 6.8: Noisy timeseries used either as a noise source or a whitening source.

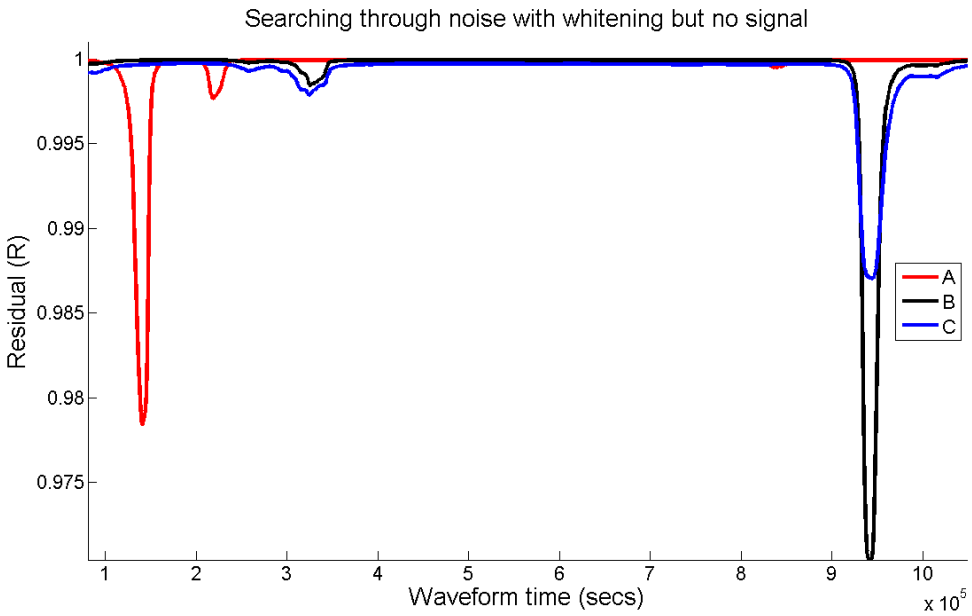


Figure 6.9: Residual measurements of searchable waveforms containing noise but no EMRI signal. The search includes the whitening stage.

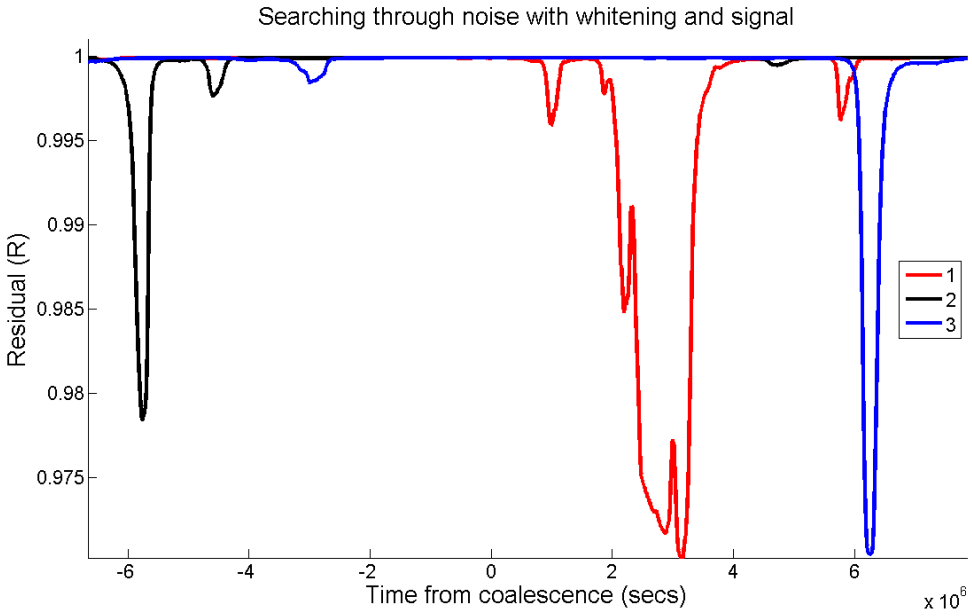


Figure 6.10: Residual measurements of searchable waveforms containing EMRI signal ξ and noisy timeseries. The search includes the whitening stage. Each searchable waveform contains a different contribution from noise, but are whitened by the same source. Variations in the residuals on the same scale as those in Fig. 6.9 are present, but not in the region close to the coalescence time for the EMRI signal.

features in the residual measurements, and no apparent correlation between any two searchable waveform results. Crucially, there does not appear to be any significant feature centered around the coalescence time for the EMRI signal or any similarity in the behaviour of the residual measurement plots around this time. Furthermore, none of the features have residual measurements that are as small as those found in the largest feature in Fig. 6.8.

Therefore, we can be confident in saying that simply whitening the test spectrograms in the manner described will not show the presence of an EMRI signal in a noisy searchable waveform in any obvious way, and any change that the signal makes to the residual measurements of test spectrograms around its coalescence time will not stand out from the changes caused by the noise.

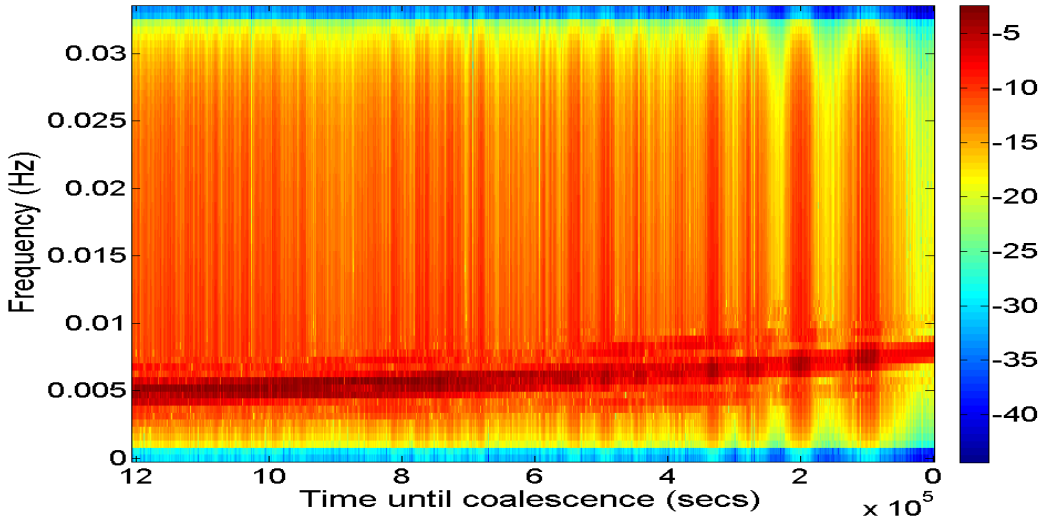


Figure 6.11: An eigenspectrogram whitened by the original noise amplitude spectrum. The presence of the EMRI signal is noticable, but the structure is slightly different from noise-free EMRI signal eigenspectrograms. The logarithm of the power is displayed to highlight the fine structure in the signal.

6.2.3 Suppressing low level noise

We are still being thwarted by the noise, but as remarked upon previously, the power from the EMRI signals is strongest at frequencies free from the strongest noise. The higher-frequency noise is not too concerning give our standard sample frequency of 1/15 Hz, but the low-frequency noise is troublesome.

To suppress the low-frequency contributions we might ordinarily apply a highpass filter to our noise timeseries but as Appendix A describes, our process is somewhat backwards, creating a timeseries from a strain amplitude spectrum. Therefore, we generate a new timeseries by zeroing the power in the amplitude spectrum for any frequency below $1e^{-4}$ Hz first and then following the procedure discussed in the appendix using this new amplitude spectrum as our starting point.

Fig. 6.14(a) shows a searchable waveform created in this manner, while Fig. 6.12 is the corresponding amplitude spectrum recalculated from this waveform. In the recalculation, a small amount of residual low-frequency noise creeps back in, but this is approximately ten orders of magnitude smaller than the noise at the 'useful' frequencies and was not deemed to significantly alter the noisy timeseries produced.

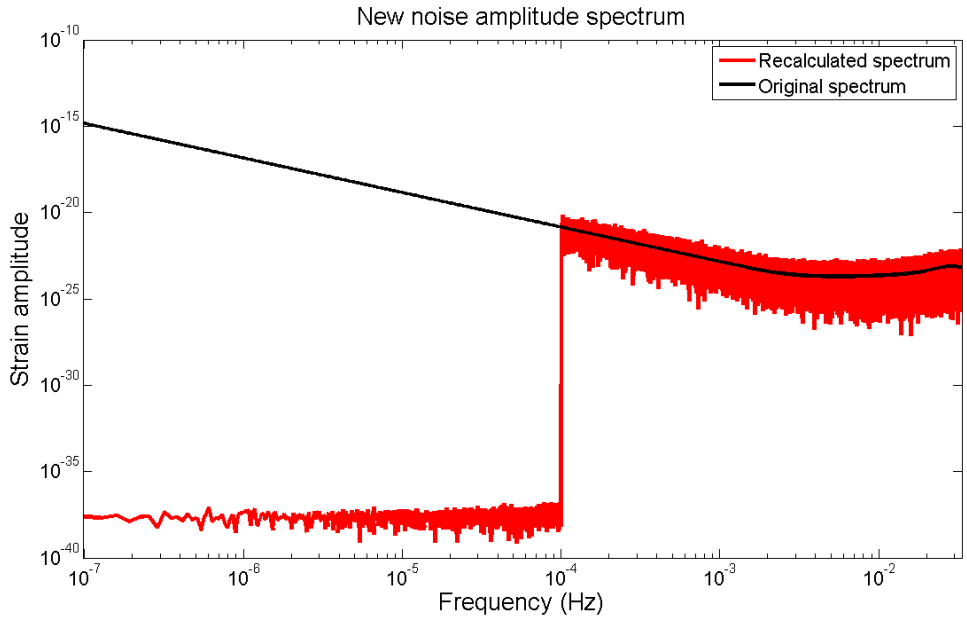


Figure 6.12: A recalculated strain amplitude spectrum of LISA noise, with the low-frequency noise suppressed. The original amplitude spectrum is also shown in black.

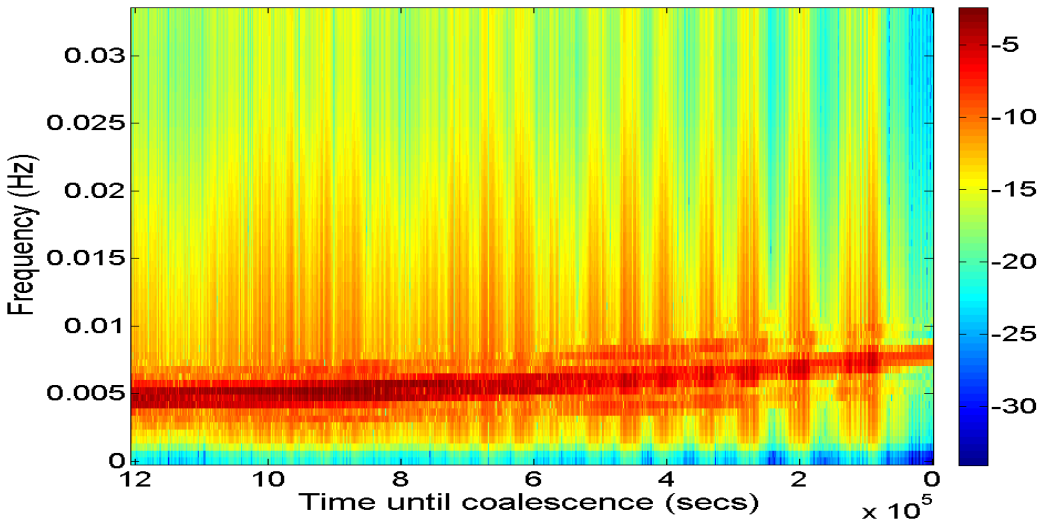
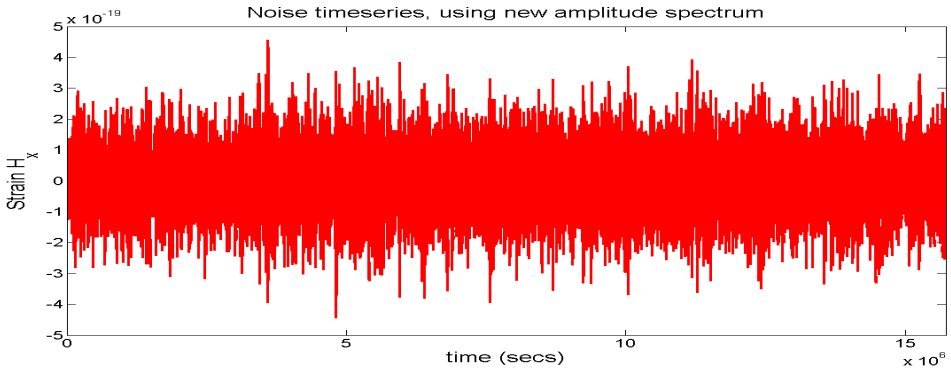
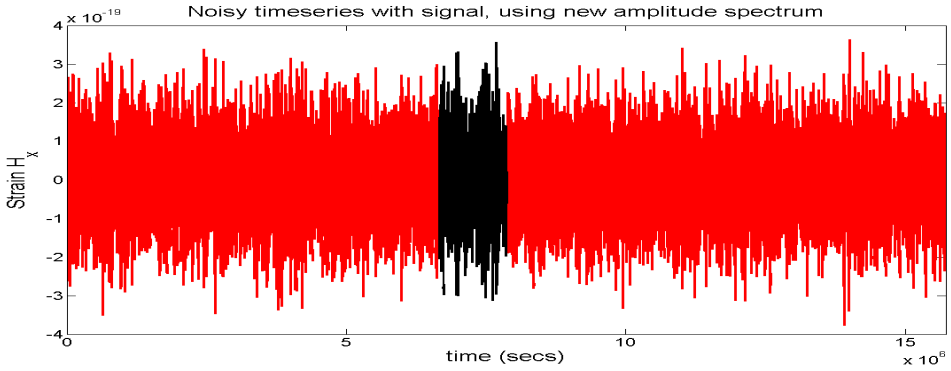


Figure 6.13: An eigenspectrogram whitened by the noise amplitude spectrum with low-frequency noise suppression. Again, the EMRI signal stands out. The logarithm of the power is displayed to highlight the fine structure in the signal.



(a) A searchable waveform with supressed low-frequency noise but containing no EMRI signal.



(b) A searchable waveform with supressed low-frequency noise and containing the EMRI signal ξ. The segment containing the signal belonging to the set Υ is shown in black. The significance of the difference between the waveform here and the waveform in Fig. 6.14(a) is not detectable by eye, but the projection into the eigenspace of the two waveforms is radically different.

Figure 6.14: Searchable waveforms with suppressed low-frequency noise

We now apply the usual search method to two searchable waveforms containing EMRI signal ξ and noisy timeseries with the low-frequency components supressed (these waveforms are defined only as 'EMRI signal 1' and 'EMRI signal 2'). Each noise timeseries is generated seperately, while a third noise timeseries is created for use in the whitening calculation, but otherwise the searches proceed as normal.

This time, there is an unmistakable dip in the residual measurements centred on the segment T_c. The drop in residual power itself is not particularly large ≈ 0.04, but it is far larger than the typical variations in segments far removed from T_c. Just to confirm that there was nothing remarkable about the positioning of ξ, another searchable waveform containing only noise (another fresh generation of a noise timeseries, defined as 'No EMRI signal') was searched as well[1].

The results of these three searches is given in Fig. 6.15, and are certainly much more pleasing than previous results at first glance. There is no confusion about the probable location of a signal due to similar drops in residual power during the waveform, and the behaviour of the residual measurements of the entire waveform is somewhat reminiscent of the results produced from noiseless EMRI spectrograms in Section 5.2.

There is a suggestion then that this dipping behaviour in the residual measurements is characteristic of the presence of an isolated EMRI signal in a searchable waveform, but we will not try to extend this into a *definition* yet. Nevertheless, this is a comforting *indication* of an isolated EMRI signal detection in a noisy timeseries.

Although the presence of the EMRI signal within the searchable waveform is not obvious, a spectrogram of the T_c segment as shown in Fig. 6.16 **will** reveal it, and a spectrogram of the same segment after whitening (Fig. 6.17) makes things even easier to see. The implications of being able to see the EMRI signal without resorting to principal component analysis are discussed in Section 6.2.6.

6.2.4 Was the whitening necessary?

The possibility that EMRI signals could be clearly indicated in a plot of residual measurements is exciting, but we must be cautious. The searches in Section 6.2.3 used test spectrograms that featured suppressed noise *and* were whitened, while we

[1] to aid in the comparison, the searchable waveform containing no signal has its time defined with respect to a non-existent coalescent time identical to that of the ξ waveform

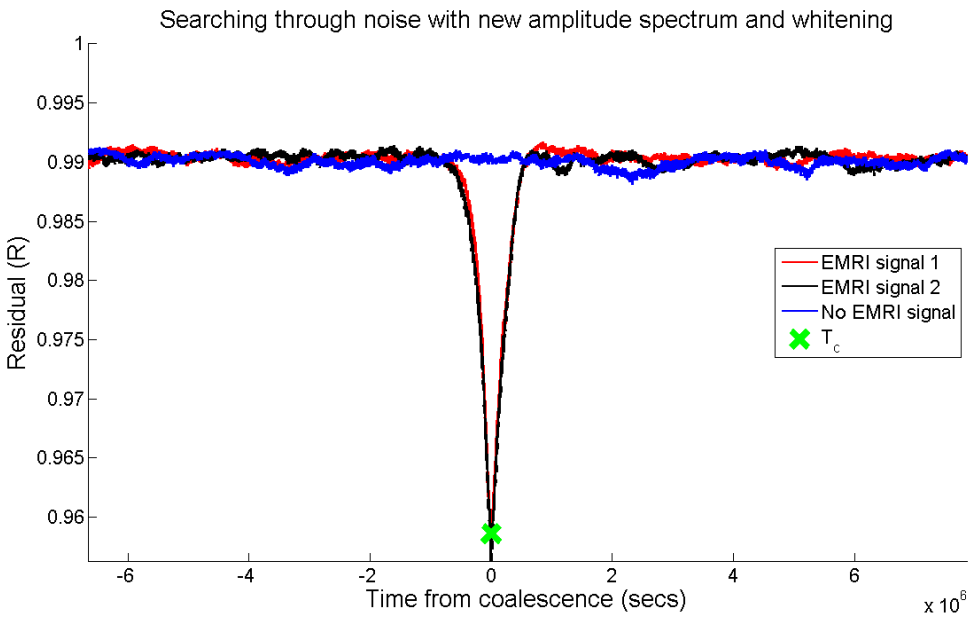


Figure 6.15: Residual measurements of searchable waveforms containing EMRI signal ξ and noisy timeseries with the low-frequency components supressed, as well as one waveform that does not contain an EMRI signal. The dip produced by the presence of an EMRI signal is obvious, suggesting that such behaviour in residual measurements may be characteristic of the presence of an isolated EMRI signal in a searchable waveform.

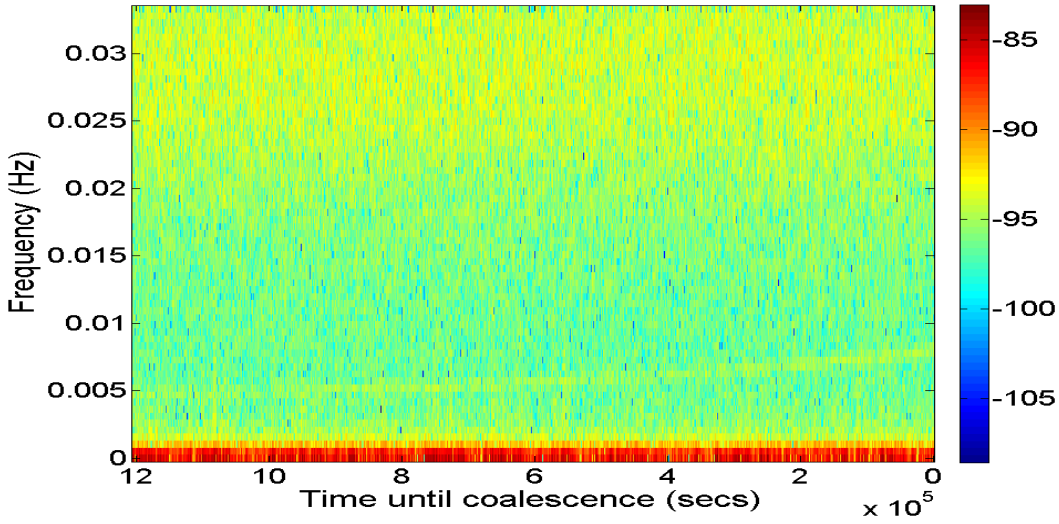


Figure 6.16: A spectrogram of the segment T_c in the searchable waveform in Fig. 6.14(b) before whitening. The EMRI signal can be seen faintly (as a line sloping between 0.005 and 0.01 mHz). The logarithm of the power is displayed to highlight the fine structure in the signal.

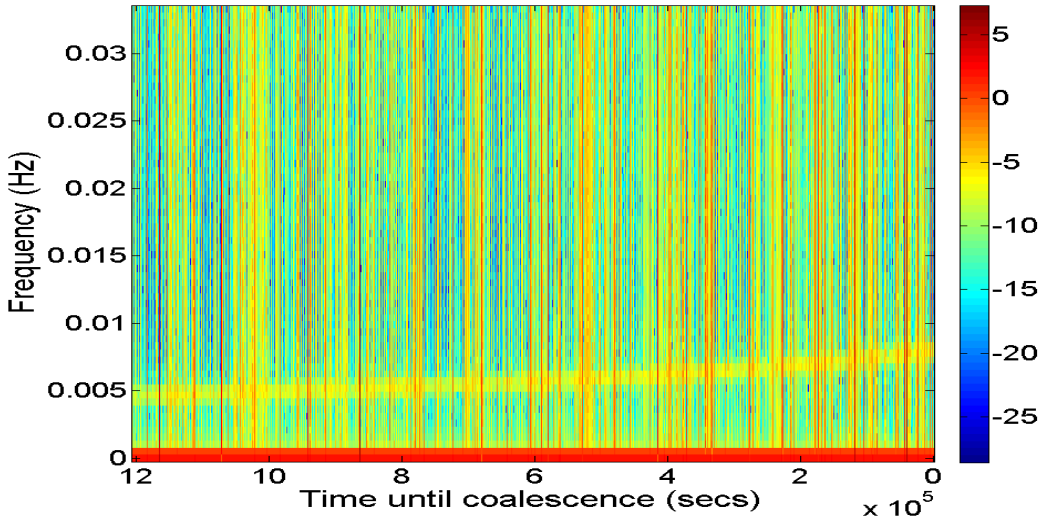


Figure 6.17: A spectrogram of the segment T_c in the searchable waveform in Fig. 6.14(b) after whitening. The EMRI signal can be seen, more clearly than in the non-whitened spectrogram. The logarithm of the power is displayed to highlight the fine structure in the signal.

saw in Section 6.2.2 that whitening without suppressing the noise produces very different results (and did not suggest the existence of a signal even when one was present).

It may be that the noise suppression was the only necessary operation, and that whitening does not produce any worthwhile effect. To settle this matter, the searches of the three searchable waveforms 6.2.3 were repeated, without whitening the test spectrograms during the search and using the original eigenspectrograms created from the spectrogram set Υ.

The residual measurements produced by these searches are shown in Fig. 6.2.4, with the two waveforms containing the EMRI signal ξ indicating the segement T_c. The presence of the EMRI signals can be seen in the overall structure of the residual measurement, but individual measurements in these 'regions of interest' are not significantly larger than those outside it, and once more we must calculate R_2 values because the residual measurements R are so large. Further, the minimum residual measurement (maximum R_2 value) in the waveform T_m does not match T_c and is not even the peak of the region of interest.

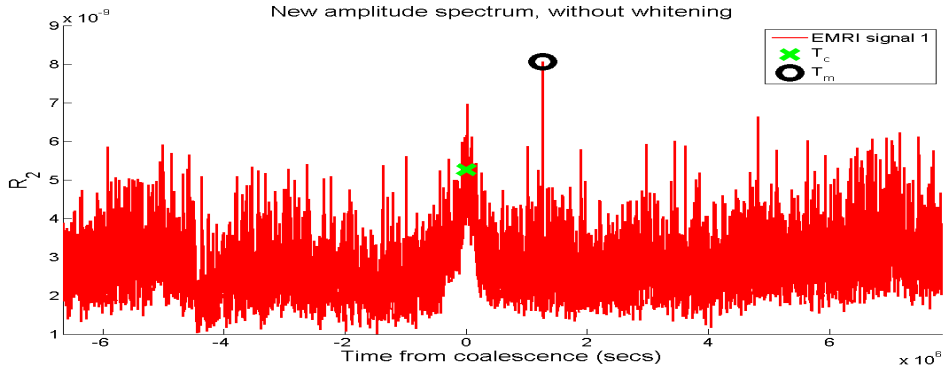
There is still the possibility that omitting the whitening could lead to a detection if the R_2 could be evaluated by eye, but automatically flagging up these regions of interest would be particularly difficult. We conclude that we need suppressed low-frequency noise *and* whitening to produce a clear indication of the presence of an EMRI signal and that neither modification will be successful on their own.

6.2.5 Noise levels

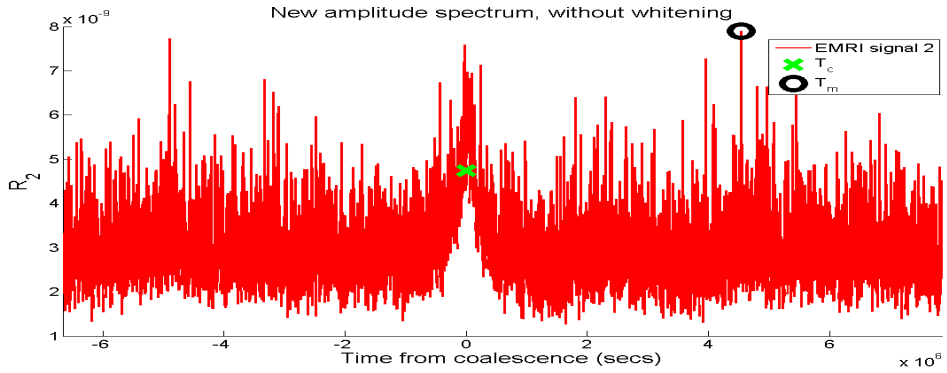
It appears that we *can* see an EMRI signal in a noisy timeseries provided we suppress the low-frequency noise and whiten the spectrograms properly. However, we must try and establish just how strong the evidence really is; after all, the drop in residual power in Fig. 6.15 is actually very small. What we want is to calculate a signal-to-noise ratio (SNR) that we can apply to a set of test spectrograms in order to define how strong potential EMRI signals are with respect to typical fluctuations.

To do so, we calculate the power in weight vectors of the test spectrograms within a searchable waveform, giving a measure of the power of their projection into the eigenspace, and will refer to this as their **Weight Power**, denoted **WP**. Thus

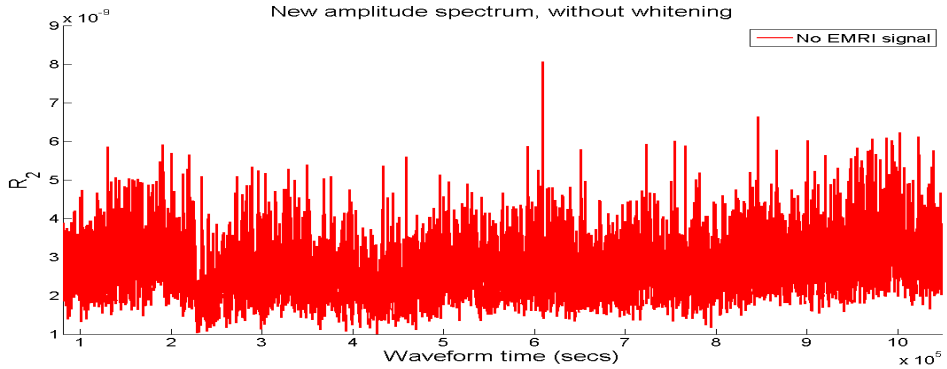
$$\mathbf{WP} = \| \Omega \| . \tag{6.1}$$



(a)



(b)



(c)

Figure 6.18: Residual measurements of searchable waveforms containing noise with suppressed low-frequency components but no whitening. In two, the presence of an EMRI signal can be seen, but the evidence is not particularly strong, and cannot be identified automatically by highlighting the T_c segment.

What we need is to determine the Weight Power found in searchable waveforms containing EMRI signal(s) and noise, as well as those containing only noise, and will denote these $\mathbf{WP_S}$ and $\mathbf{WP_N}$ respectively (note that there will be as many of these measurements as there are search spectrograms for any given searchable waveform). Using the set of noise Weight Power measurements, we calculate the *mean* noise Weight Power $< \mathbf{WP_N} >$ and the standard deviation in the set δ; with this, we have an estimate of the power in spectrograms within a noisy searchable waveform, as well as an estimate of the typical fluctuations throughout the searchable waveform.

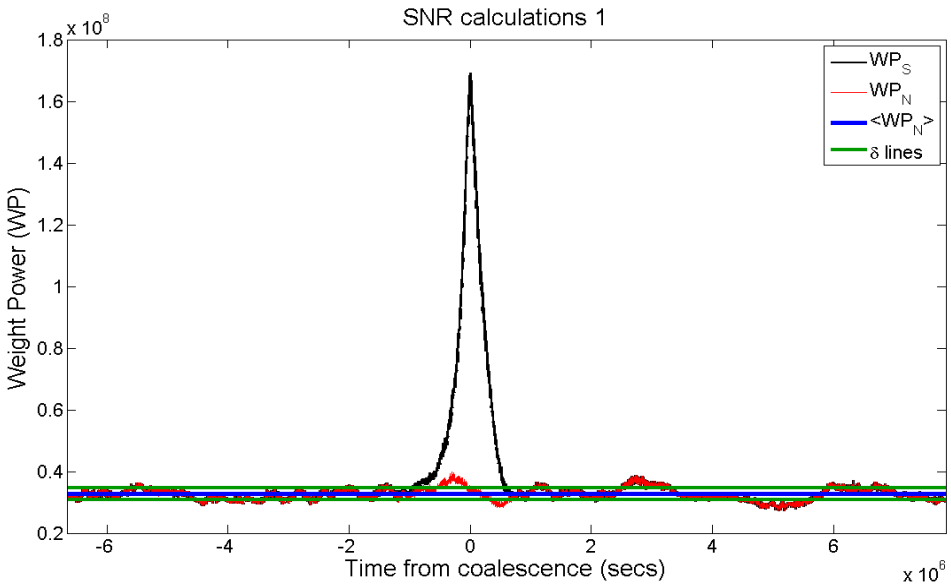
Now, when analysing a searchable waveform, we take a test spectrogram ($\mathbf{q}$ say), and calculate its Weight Power $\mathbf{WP_S}(q)$. We define the SNR for this spectrogram to be

$$SNR(q) = \frac{\mathbf{WP_S}(q) - < \mathbf{WP_N} >}{2\delta} \tag{6.2}$$

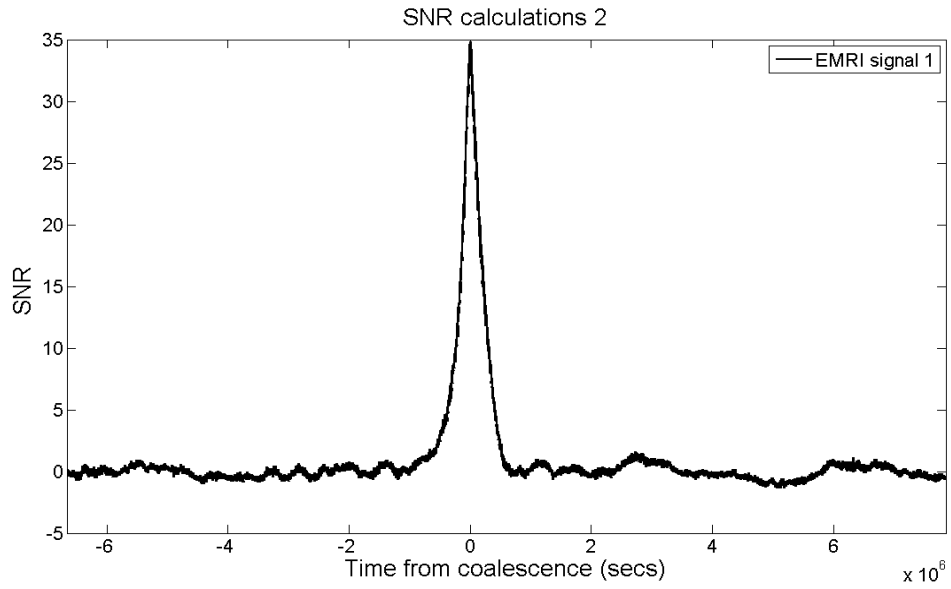
The reasoning behind this calculation is straightforward when shown pictorially. Fig. 6.19(a) shows the Weight Power of spectrograms within a searchable waveform containing an EMRI signal (specifically, the same waveform denoted 'EMRI signal 1' in Fig. 6.15) as well as the Weight Power within a searchable waveform containing only noise (here, the waveform denoted 'No EMRI signal' in Fig. 6.15). These measurements are $\mathbf{WP_S}$ (in black) and $\mathbf{WP_N}$ (in red).

From this, the mean noise Weight Power $< \mathbf{WP_N} >$ is determined (included as a blue line) and the values of noise Weight Power one standard deviation either side of this measurement are calculated too, revealing a 'channel' marking the typical fluctations in the noise Weight Power with width 2δ. The SNR then, is how far a test spectrogram's Weight Power is above the average noise Weight Power, as a measure of the typical fluctuations of noise Weight Power.

The definition we have used is straightforward in its calculation, but isn't the usual way of thinking about SNR in a noisy waveform since we are only defining the signal in terms of a its projection into a space we have constructed. This is potentially misleading, since if we do not define this space correctly then we may ignore some contribution from an EMRI signal. Nevertheless, we have always been aware that the PCA-based construction of a problem space would not always be complete and would be at best a useful approximation of the original parameter space. Given this, our definition of SNR behaves appropriately; the process is insensitive to EMRI



(a) The Weight Power of test spectrograms in the EMRI signal 1 searchable waveform shown in Fig.reffig: whitenened 1 (in black), as well as the Weight Power of test spectrograms containing only noise (in red). Also included are the mean noise Weight Power (blue line) and boundaries one standard deviation from this measure (in green).



(b) The signal-to-noise ratio determined from the plot above. The basic shape is retained, but the measure is more meaningful.

Figure 6.19: Weight Power and SNR calculations for a searchable waveform.

signals or EMRI-like signals that do not project into the eigenspace, and similarly only the noise that projects into the eigenspace is counted.

In practice, we can see easily that if a searchable waveform contains only noise and no EMRI signal then the SNR measurements will be around 0, while a stronger EMRI signal (resulting from an increase in the gravitational strain amplitude) will increase the SNR. More power in the noise will decrease the difference between the signal weight power and the mean noise weight power (hence reducing the 'signal' part of the SNR calculation), while if the fluctuations in the noise weight power increase (consequently increasing δ) then the SNR will be reduced[1]. Always, there is the underlying assumption that the noise used to calculate $< \mathbf{WP_N} >$ is an appropriate estimate of the noise present in the searchable waveform containing the EMRI signal(s).

To demonstrate how the SNR changes, we create new searchable waveforms containing EMRI signals of different strain amplitudes; easily done by changing the distance of the EMRI source. Each searchable waveform is created including the contribution from a different noise timeseries, but is whitened by the same source (we use the waveform shown by Fig. 6.14(a)). In all other respects the EMRI signals are exactly the same as ξ. Table 6.2 gives details of the SNR measurements shown in Fig. 6.20, including the maximum SNR for each set of measurements (SNR_{max}) and the distance of the EMRI source. Also included is the minimum residual measurement from each set (R_{min}, corresponding to $R(T_m)$), which in all the waveforms tested was the same spectrogram as T_c.

The effects are clear, if unsuprising; increasing the amplitude of the EMRI signal added to the searchable waveform increases the SNR.

6.2.5.1 The Principal Components of the noise?

Since we are able to produce spectrograms of the noise, it might seem that we could avoid some of the problems caused by noise by determining their principal components and using them to expand the eigenspace. We might then imagine that

[1] One must be careful about what we mean by the 'noise'. In the context of the searchable waveforms, noise refers to the fluctuations in the gravitational strain resulting from the LISA sensitivity curve, while in the context of the SNR calculations we refer to the fluctuations in the weight power measurements of a searchable waveform containing LISA sensitivity noise but no EMRI signal.

SNR measurement set	EMRI distance (pc)	SNR_{max}	R_{min}
1	2×10^9	34.47	0.958
2	1×10^9	308.01	0.784
3	0.1×10^9	2.24×10^6	0.058
4	0.01×10^9	2.23×10^{10}	0.006
5	0.001×10^9	2.23×10^{14}	5.54×10^{-4}

Table 6.2: The distance of an EMRI source, and the corresponding maximum SNR (SNR_{max}) and minimum residual measurement R_{min}.

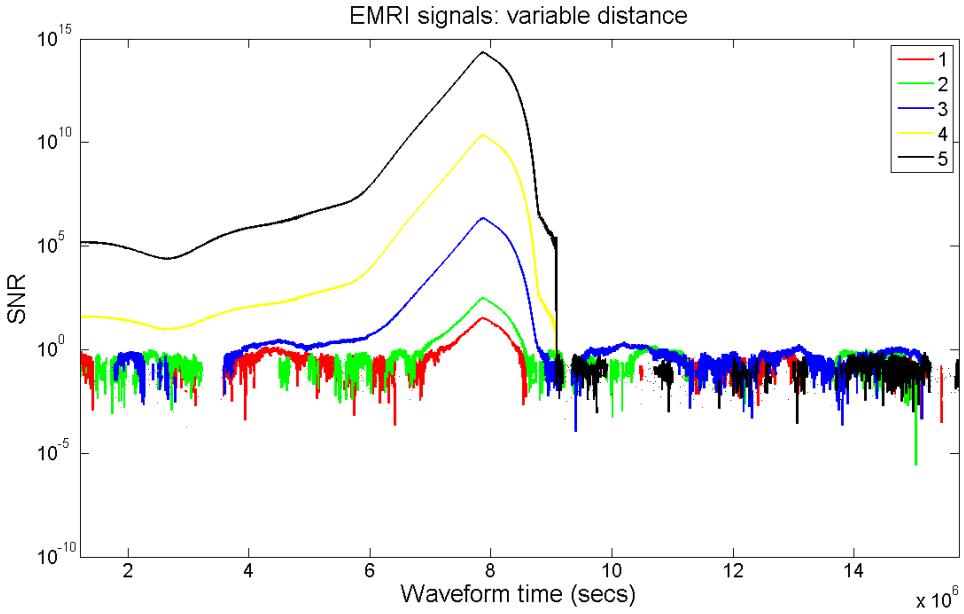


Figure 6.20: The SNR of EMRI signals from varying source distances; as the distance decreases, the SNR increases. The range in SNR necessitates a logarithmic scale, so all SNR < 0 values are omitted.

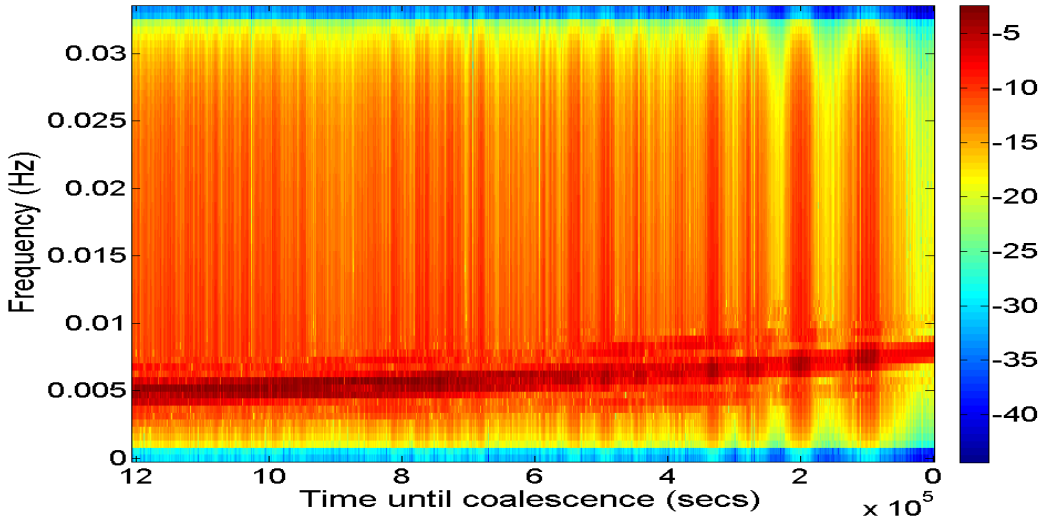


Figure 6.21: An eigenspectrogram, calculated from whitened EMRI spectrograms without the low frequency suppression. The logarithm of the power is displayed to highlight the fine structure in the signal.

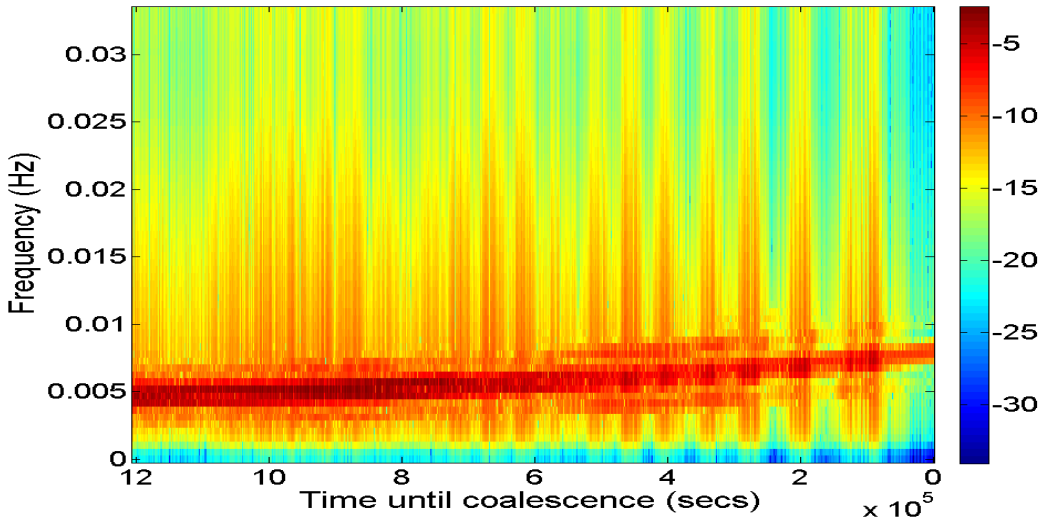


Figure 6.22: An eigenspectrogram, calculated from whitened EMRI spectrograms and low frequency suppression. The logarithm of the power is displayed to highlight the fine structure in the signal.

a noisy EMRI signal within a searchable waveform can be reconstructed perfectly by the eigenspectrograms that span this space, and that the presence of this kind of signal would be quite obvious once the results of the search.

This may indeed be the case, but we would now be attempting to span the problem space inhabited by EMRI signals *and* LISA noise. Regardless of the extra computational cost of such an approach (which may be significant, although we are currently ignorant of this), this shift in the focus of the search would not actually help; pure noise segments would still project into the eigenspace that had been created, and could also be reconstructed to some degree, presenting themselves as possible candidates for EMRI detections.

Even if the noise was orthogonal to the EMRI signals -which is largely true since the noise doesn't project well into the EMRI signal eigenspace according to Section 6.2.1.1- and created eigenspectrograms only concerning the noise, all that weight vectors of segments showing large contributions from these eigenspectrograms really reveals is the presence of noise in the segment (we could in theory identify these by looking at each eigenspectrogram in turn by eye, although there is no guarantee that it would be obvious). Clearly, this does not aid our understanding to any great extent, since we were aware of the noise being present in the waveforms in the first place.

For these reasons, the principal components of the noise are not used or even calculated. Further discussion of this matter is given in Chapter 7.

6.2.6 Remarks on noisy EMRI signals

Adding noise to the searchable waveforms affects the search for EMRI signals greatly, and this section has contained a large amount of information. It will be beneficial to summarise what we have learned before remarking on individual aspects highlighted by the searches.

In brief, we are unable to get a clear indication of an EMRI signal's presence in a searchable waveform unless the low frequency noise that LISA will see is suppressed *and* the eigenspectrograms and test spectrograms are whitened according to the shape of the new noise amplitude spectrum with this suppression. Neither whitening alone nor reshaping the noise amplitude spectrum alone will accomplish this, and if both these stages are omitted then we are unable to gain any useful information from the search whatsoever.

We have also seen that changing the whitening source is not hugely important as long as it has a similar amplitude spectrum, and that different noisy timeseries within the searchable waveforms do not change search results a great deal either if they have the same amplitude spectrum too.

The SNR measurements are only a guide for the moment. Like the prospect that a dip in the residual measurements of a searchable waveform (or a spike in the SNR) indicates the presence of an EMRI signal, we have only thus far found something to build on. A more thoughtful examination of how to register a potential detection is given in Section 6.5, but there are still other aspects to the searchable waveforms that must be explored and accounted for before we formalise our detection process.

Naturally, the previous tests do not cover every aspect of noisy searchable waveforms. The behaviour of the residual measurements in the presence of a significantly different amplitude spectrum, using a radically different whitening source, or the measurements that would be produced using a set of eigenspectrograms created from spectrograms inhabiting a different region of EMRI parameter space have not been included, but by this stage we feel sure that we do not require such exhaustive testing. So far, there has been every indication that the PCA-based search method is robust enough that we could process this data, and consistent enough that we can reproduce our successes if necessary.

One concern is that whitening the spectrograms as they are incorporated into the training set and used to generate the eigenspectrograms is taking unfair advantage of our prior knowledge, but it is a necessary step. Whitening at this stage is reshaping the EMRI signals, without which they would always differ from the whitened test spectrograms (even if the EMRI signal within that test spectrogram belonged to the training set) resulting in high residual measurements during searches. Naturally, the resulting eigenspectrograms are rather different from those produced without the whitening stage, but as mentioned previously, the unique components from each EMRI signal are still present despite the reshaping. Conceivably, these could be altered enough by the whitening that the residual threshold for test spectrograms R_T would have to be different from the threshold for non-whitened test spectrograms, but this is simply done; a repeat of a test like that in Section 4.3.3 would indicate an appropriate choice.

It is slightly concerning that the EMRI signal can be seen by eye in the spectrograms of low-frequency suppressed noisy timeseries, but we must bear in mind that

the PCA-based search method developed is able to search through many thousands of spectrograms in a consistent manner, producing a measure of its projection into a parameter space inhabited by EMRI signal (and beyond this, a measure of the SNR of these projections). Further, while we have only considered single isolated signals so far, we can extend the scope of the problem to include multiple signals for which tracking individual EMRI signals by eye will not be straightforward.

6.3 Multiple EMRI signals

Until now the EMRI signals in the searchable waveforms have been isolated, but it would be unwise to evaluate the PCA-based search method only in this simple situation. Therefore, we consider the additional complexities of a waveform containing a number of EMRI signals. Throughout, we consider a number of contributions to strain measurements from different CO - SMBH EMRI systems, rather than multiple COs orbiting around a single SMBH.

The presence of multiple signals renders the 'time to coalescence' measure in a searchable waveform insensible. Therefore, every waveform will be presented in terms of 'waveform time', like searchable waveforms without a signal. Further, we are attempting to examine the behaviour of the search only with regards to aspects involved in multiple signals and as such have temporarily removed the contribution expected from LISA noise. Consequently, the need for SNR calculations is eliminated too.

Finally, each of the waveforms described in this section were regenerated several times using different parameters for the EMRI signals, although each was taken from the set Υ in all parameters except for the duration of the waveform, producing near-identical results. This is to be expected, since we have a set of eigenspectrograms that spans the required eigenspace completely, and the projection of individual spectrograms into this eigenspace only alters the specific weight vector being considered.

6.3.1 The α waveform

Initially, we examine a searchable waveform α containing two widely-separated EMRI signals (shown in Fig. 6.23(a)), each taken from the set Υ. Although this is a somewhat backwards approach given the unrealistic nature of the EMRI signal 'switching on' described in Section 6.1, at this stage we simply want to check that

the search method will work in the expected manner; with no overlap between the two signals, and each reconstructible on their own, the search should show two dips in the residual measurements each centred on the coalescence time of the EMRI signal.

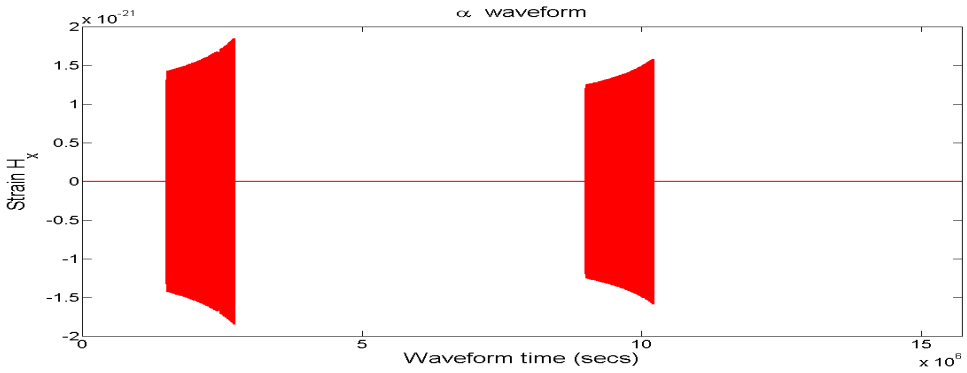
Fig. 6.23(b) shows that this is indeed the case. There is no signal confusion, and the two dips in the residual measurement have a minimum at the coalescence time for each EMRI signal. A closer look at the dips shows that each of the two T_c segments correspond to almost perfect reconstructions ($R \approx 1 \times 10^{-14}$ in both cases), again as expected. Naturally, our usual approach of checking the minimum residual measurement is not particularly useful since it will only select a single value (and corresponding coalescence time) and we would prefer something that highlights two distinct signals, but judging it only by eye for the moment we are confident that the presence of two EMRI signals does not cause surprising behaviour from the search method.

6.3.2 The β waveforms

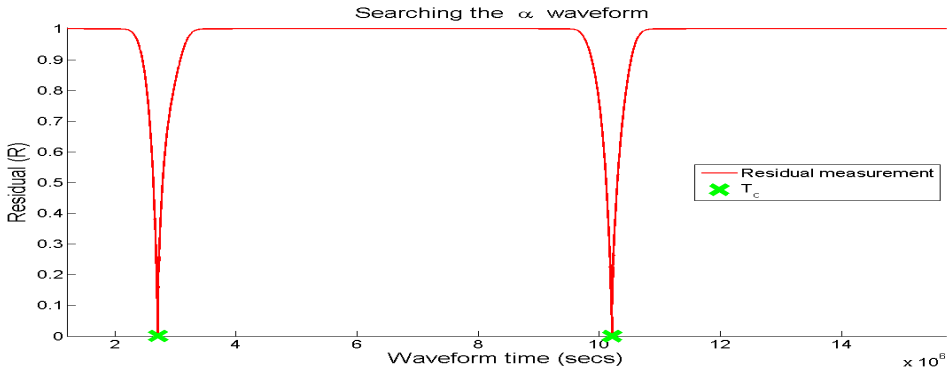
Able to handle the simplest of multiple-signal situations, we now construct a number of searchable waveforms designated β_i $(i = 1, 2, .., 6)$, where the time between the coalescence of two EMRI signals is altered. Unlike the α waveform signals, the β waveforms do not switch on, creating an overlap between the two EMRI signals, but apart from the duration of the signal present in the searchable waveform they share all parameters in common with two EMRI signals taken from the set Υ.

Fig. 6.24(a) and Fig. 6.24(b) show the residual measurements from the searches of the β waveforms, while Table 6.3 gives the coalescence times involved. The first signal (in waveform time) is kept constant, while the second appears to move backwards in time as subsequent tests are performed. As a result, what begins as residual measurements reminiscent of those produced from the α waveform gradually change from distinct dips to the creation of a general region of reduced residual measurements as the coalescence times get closer together.

As the overlap between the signals increases, so does the confusion between the two. The projection of the search segment spectrograms into eigenspace changes and it becomes increasingly difficult to distinguish between the two, but crucially this projection is not so alien as to produce very high residuals. In fact even at their closest (alternatively, at their maximum overlap) there remains a structure

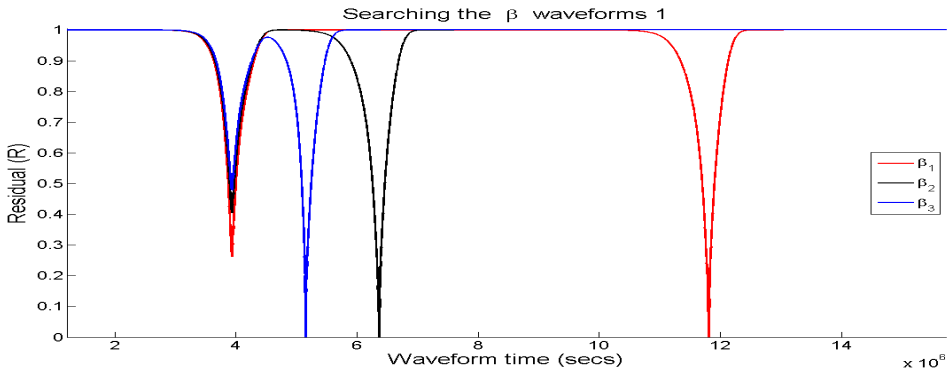


(a) The searchable waveform α.

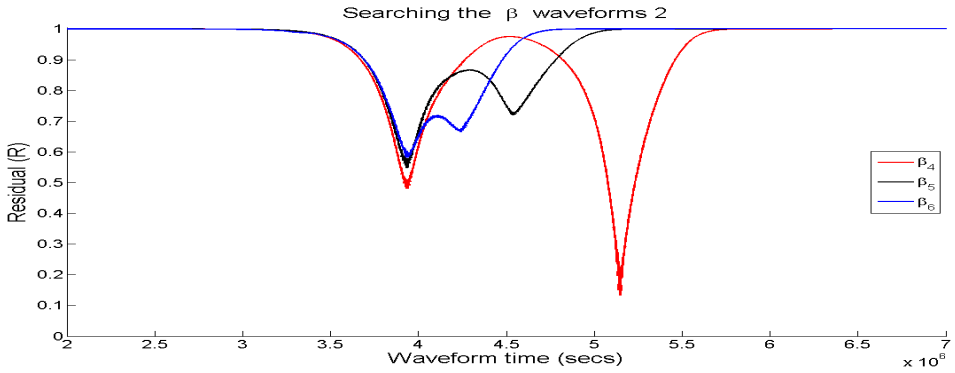


(b) Residual measurements of the searchable waveform α.

Figure 6.23: Searchable waveform α, and residual measurements. Two EMRI signals are present, widely-separated in time. With no overlap, the PCA-based search method is able to reconstruct them almost perfectly, strongly indicating their presence.



(a) Residual measurements of the searchable waveform $\beta_1 - \beta_3$.



(b) Residual measurements of the searchable waveform $\beta_4 - \beta_6$.

Figure 6.24: Residual measurements of the β searchable waveforms. As the overlap between the signals increases, the two signals become less distinct.

Searchable waveform name	t_c (EMRI signal 1)	t_c (EMRI signal 2)
β_1	3939600	11814600
β_2	3939600	6369600
β_3	3939600	5154600
β_4	3939600	5139600
β_5	3939600	4539600
β_6	3939600	4239600

Table 6.3: The coalescence times of the two EMRI signals in the β searchable waveforms. Coalescence times are given in terms of the waveform time (in seconds).

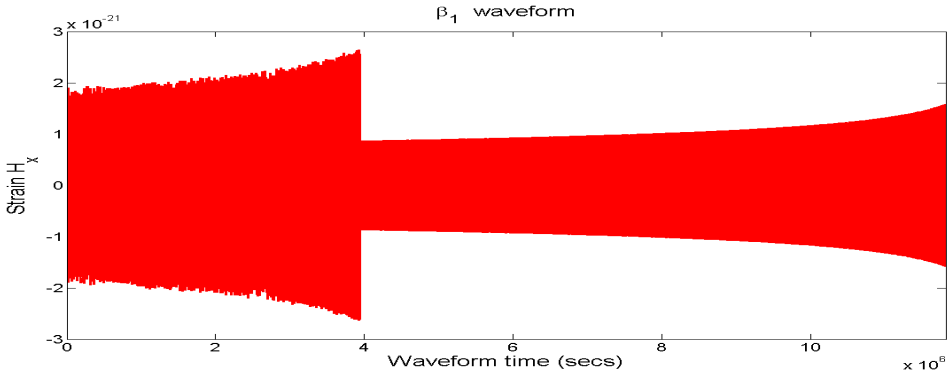
that suggests to the eye that two different signals are present, something that is considerably more difficult to see from looking at the searchable waveforms themselves. Fig. 6.25(a) and Fig. 6.25(b) show the searchable waveforms β_1 and β_6, demonstrating the signals with minimum and maximum overlap respectively. While we might infer that there are two signals in β_1 based on what now about isolated EMRI signal waveforms, β_6 looks different, with the region of overlap beginning to look like a single signal or perhaps a different type of signal altogether.

The coalescence times were chosen to represent a variety of circumstances. The first three are widely separated, but the third was chosen to produce a 'bridge' between the two dips; the two coalescence times here are separated by a period slightly greater than the segment length of a search spectrogram. β_4 separated the coalescence times by slightly less than one segment length, and β_5 and β_6 close this gap even further.

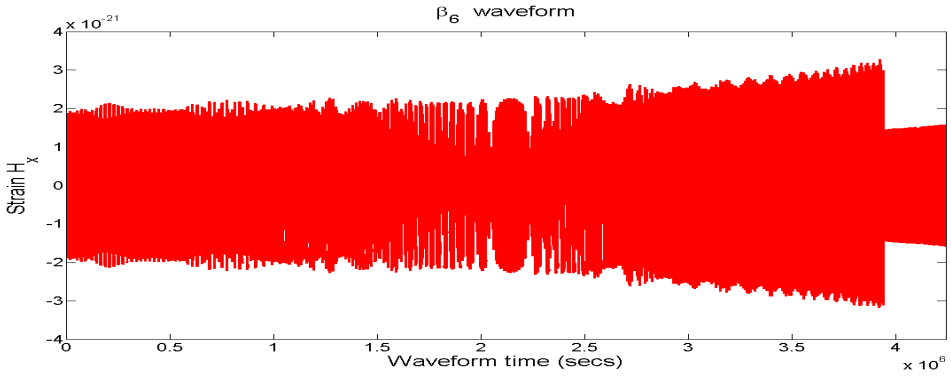
It is important to note that although the signal with the later coalescence time is being altered in terms of its placement, the earlier signal is not unaffected. Even when the dips are fairly distinct (as produced from $\beta_1 - \beta_3$), the residual measurements around the time of this earlier coalescence begin to climb higher as the second signals' coalescence 'approaches'.

6.3.3 The γ waveforms

The searchable waveforms γ are a special case of β-like waveforms, imagining a situation where several EMRIs coalesce at the exact same time ($t_c = 7869600$ secs). Between two and five different signals were added to the waveforms (each taken at



(a) Searchable waveform β_1.



(b) Searchable waveform β_6.

Figure 6.25: β.

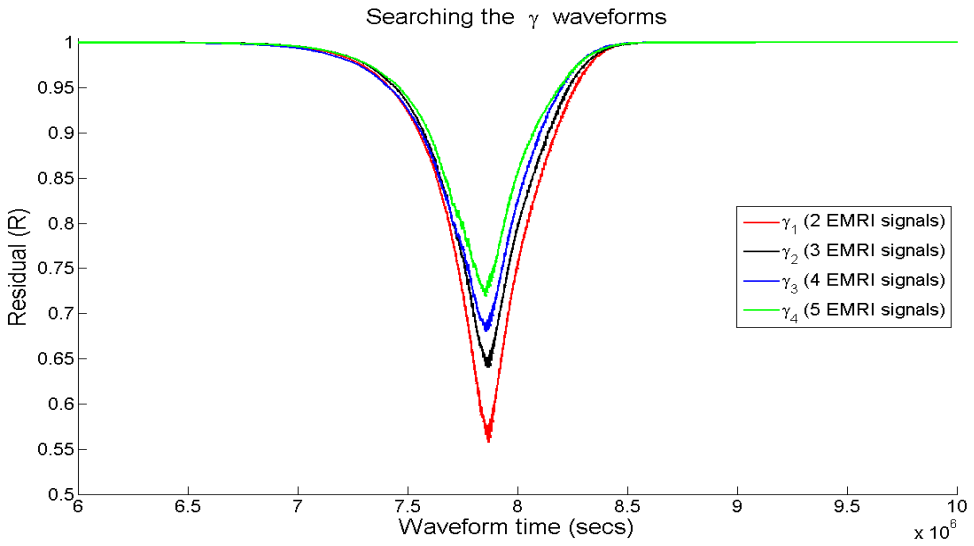


Figure 6.26: Residual measurements from the searchable waveforms $\gamma_1 - \gamma_4$.

random from the set Υ but evolved backwards for $\approx$ six months), but again the search proceeded as normal. Adding more signals increases residual measurment around the coalescence time, as Fig. 6.26 clearly demonstrates, but the shape of the residual measurements remains unchanged from what we have begun to deem characteristic of an *isolated* EMRI signal.

This is not entirely unexpected. The overlap in the different waveforms doesn't change, and the individual EMRI signals do look very similar, so we might reasonably predict that the resulting combined signal is also very similar. And despite the overall strain in the searchable waveform increasing as more signals are added, this alone would not make the signal increasingly difficult to reconstruct.

Instead, we have situation akin to that explored in Section 4.2.1.2, whereby the inter-modulation of two combined sinusoids result in a projection into the eigenspace that differs from the projection of two individual sinusoids which are then combined. In the γ waveforms, the signals being combined are more complicated but otherwise the situation is the same.

If each EMRI signal was identical however, the reconstructions would be close to perfect (and the residuals close to zero), since this would project into the eigenspace in a manner identical to one of the spectrograms belonging to the set Υ, but a different weight vector (scaled by some factor depending on how many EMRI signals

had been included). As discussed before, the PCA search method is not sensitive to amplitude this way, and would handle the increased contribution from each eigenspectrogram without concern. It is only the fact that each of the EMRI signals in the γ waveforms is different that causes such high residual measurements.

What the γ waveforms suggest then is that we will not be able to identify overlapping EMRI signals, and that looking at the residual measurements alone will not allow us to tell whether we are looking at multiple signals (and if we are, how many), or just a signal that is EMRI-like in appearance. This is unfortunate, since it hampers our ability to use the PCA-based search to identify numerous EMRI signals, and may restrict us merely to highlighting periods of time when EMRI signals are likely to be occurring.

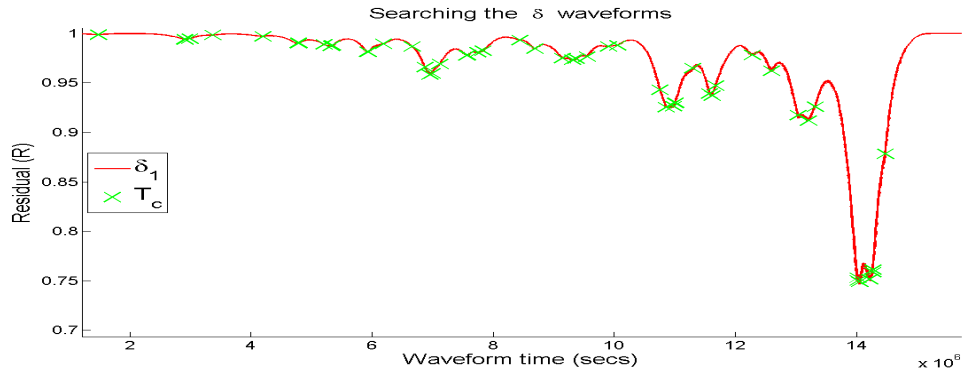
6.3.4 The δ waveforms

A final set of searchable waveforms present an extreme challenge for the PCA-based search method; 50 EMRI signals with randomly-generated coalescence times are combined and examined. Each signal is randomly selected from the set Υ although evolved backwards for ≈ 6 months, and the intention is to present an almost chaotic situations that the search will struggle with.

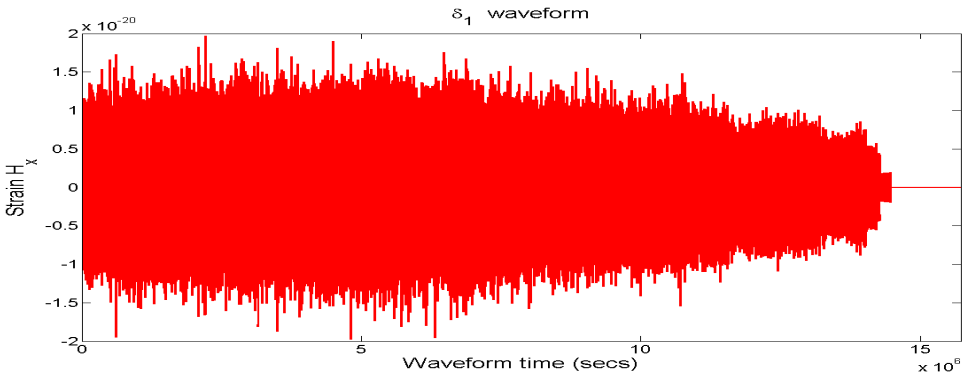
Fig. 6.27(a) and Fig. 6.27(b) show both the residual measurements from one such search and the searchable waveform. It is not possible to see the individual EMRI signals simply by eye (except the final one, which is addressed later) and it is difficult to determine anything meaningful about the searchable waveform in this manner. Furthermore, the residual measurements do not display distinct dips centred on the coalescence times of the signals.

Signal confusion reigns in this waveform, with the residual measurements for the beginning of the waveform being very high ($R \approx 1$). Further on in the waveform, sufficient time has passed to allow some EMRIs to coalesce and their contribution to the overall signal disappears, causing the signal confusion to drop off and resulting in a general trend of the residual dropping as time goes on. Nevertheless, it is only when the majority of the EMRIs have coalesced that the residual measurements drop by any great amount, and there is no point where there is a distinct dip revealing an isolated EMRI signal.

This behaviour occurs time and again. Fig. 6.28 shows three different δ searchable waveforms, each displaying similar behaviour. It is clear that this amount of



(a) Residual measurements for searchable waveform δ_1.



(b) Searchable waveform δ_1.

Figure 6.27: The searchable waveform δ_1 and the residual measurements from the search.

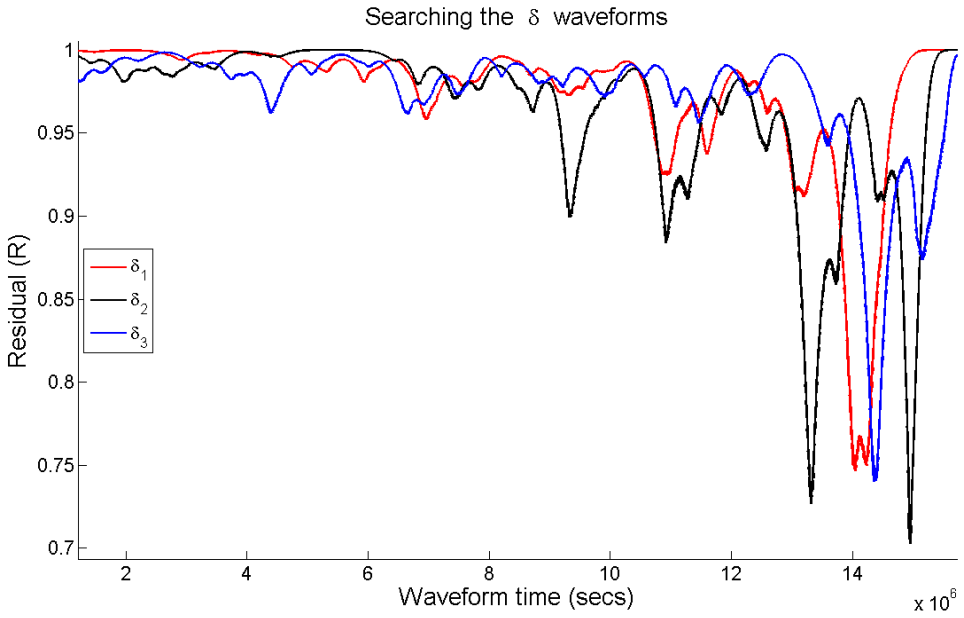
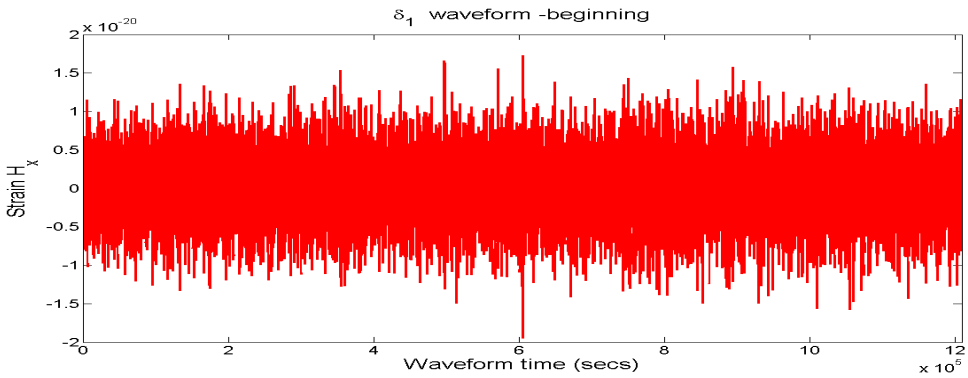


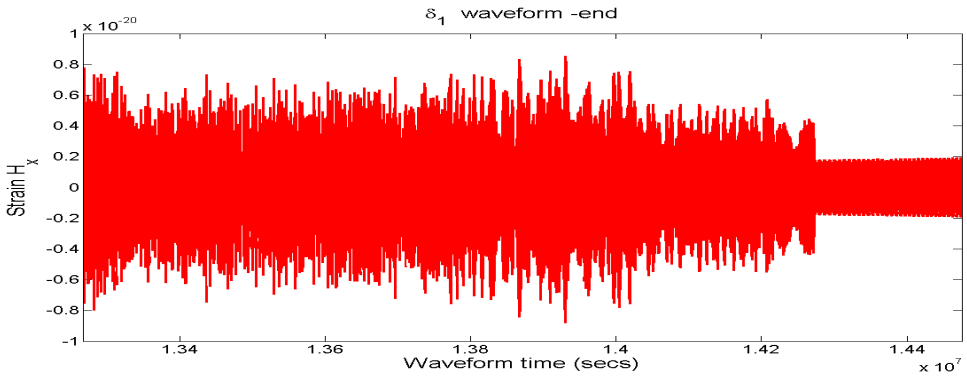
Figure 6.28: Residual measurements from the searchable waveforms $\delta_1 - \delta_3$.

overlap between different EMRI signals is extremely detrimental to any attempt to reconstruct test spectrograms. There may be a few pronounced dips in the residual measurements, but nothing that would strongly suggest that we are confident about suggesting a coalescence time.

Fig. 6.29(a) shows the first segment-length of the searchable waveform δ_1, a region of extreme confusion that projects far from the eigenspace, while Fig. 6.29(b) shows the final segment-length, when most of the EMRIs have coalesced and are no longer contributing to the overall signal. Fig. 6.30(a) and Fig. 6.30(b) then show the spectrograms that would be constructed from these segments of the searchable waveform. The signal confusion is self-evident in Fig. 6.30(a) (one might suggest there is more than one EMRI signal present, but it is not possible to see 50 distinct signals), while enough coalescences have occurred by the time of the test spectrogram in Fig. 6.30(b) for individual EMRI signals to be seen dropping out. Regardless of the improvement over time, the search is still looking at a spectrogram containing more than one EMRI signal at all time, and struggles to reconstruct it using the eigenspectrogram set available.

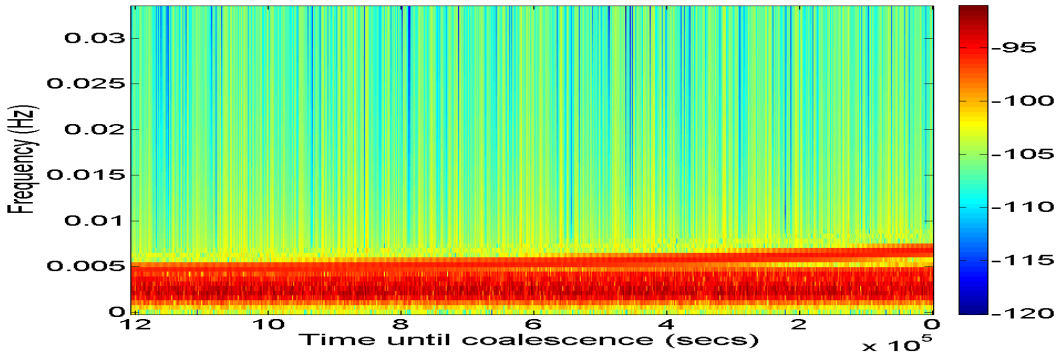


(a) The first segment-length of searchable waveform δ_1.

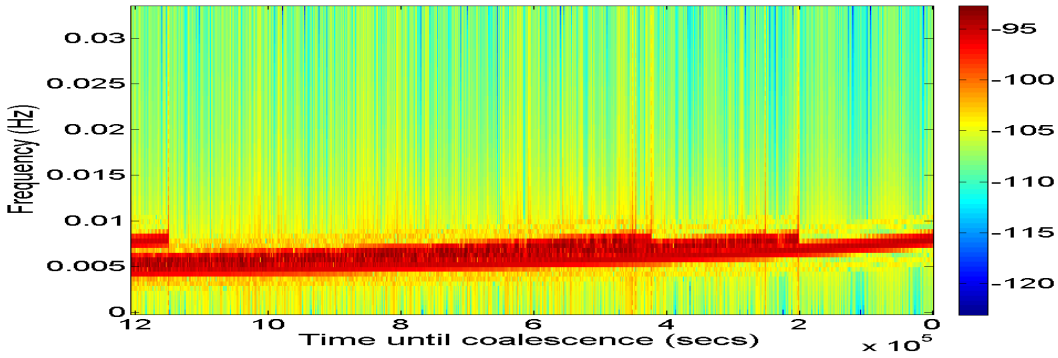


(b) The final segement-length of searchable waveform δ_1.

Figure 6.29: Searchable waveform δ_1 segments



(a) The first test spectrogram of searchable waveform δ_1.



(b) The final test spectrogram of searchable waveform δ_1.

Figure 6.30: Searchable waveform δ_1 spectrograms. The logarithm of the power is displayed to highlight the fine structure in the signal.

6.3.5 Mutliple signal searchable waveform remarks

The effects of overlapping EMRI signals are certainly a necessary consideration, but as presented the previous waveforms still represent a simplified situation. Principally, we have still restricted ourselves to a six-month period to search, and a limited eigenspace within which to project the signals.

Practical reasons prevent us from eliminating these restrictions, but they are still concerning. Although it seems unlikely, the combination of multiple EMRI signals within a parameter space may combine in a manner that mimics an EMRI signal outwith that parameter space, and we cannot rule this out without spanning the EMRI problem space in its entirety. Similarly, there may be other gravitational wave sources that project into an EMRI signal eigenspace like EMRI signals themselves. Further, our tests have thus far only highlighted situations where EMRI signals end during the period covered by the searchable waveform and ignore EMRI signals that may be many years from coalescence.

Nevertheless, the α,β,γ and δ waveforms do suggest that the presence of multiple EMRI signals can still be detected using the PCA-based search method under non-ideal circumstances, even though we may not always be confident about how many sources there are or their exact coalescence times.

6.3.6 EMRI populations and event rates

The waveforms β_1 and δ_1 represent little signal confusion and extreme signal confusion respectively, and while examining the behaviour of the search under these two extreme circumstances is useful, a searchable waveform with a more realistic amount of signal confusion is highly desirable. Therefore, what we require is an accurate estimate of the EMRI coalescence event rate.

Unfortunately, calculating such an estimate is far from trivial. Gair and Barack (41) explore the issue thoroughly, considering estimates of gravitational capture rates of SMBH and the space density of SMBHs amongst other factors, while Freitag (40) considers the inspiralling rates in a central galactic black hole. Common to both, however, is the acknowledgement that the calculations contain many uncertainties. Furthermore, the number of detectable signals of the EMRI signals in (41) is calculated using event rate estimates and their theoretical SNR, based on coherent integration.

The PCA-based search is a different prospect, and it is unclear quite how many sources may produce detectable signals, and at what rate, since we do not yet have a robust definition of 'detection'; so far, our best method for estimating EMRI signal coalecences within the framework of the PCA-based approach are gained by examining the SNR measurements by eye.

A cosmological model of EMRI signal sources, their distribution and their projection into an eigenspace is beyond the scope of this thesis but there is no denying that without this information it is not possible to accurately estimate the signal confusion due to multiple EMRI signals. The multiple-signal tests were not without merit, however; we were able to see the effects of overlapping signals (and it is inevitable that should more than one EMRI signal be present during the time LISA is operational there will be some degree of overlap), and the δ waveforms do show a situation in which the search method will struggle in its present form. Inevitably the PCA code will not work under all circumstances, and it may be that the presence of multiple EMRI signals represent a fundamental shortcoming. Some further discussion of this issue is given in Chapter 7.

6.4 The antenna pattern

Of all the approximations we have made in our analysis so far, the most obvious departure from reality is that we have ignored the effects of LISA's orbit and the response of the detector. In all previous tests, the detector is considered to be at the solar system barycentre for the duration of the searchable waveforms, and we have only examined the $h_\times$ polarisation of gravitational waveforms. Furthermore, the response of the detector to the gravitational wave sources is not included in our analysis. Naturally, lacking these features is concerning. Time constraints prevented their inclusion into the PCA-based search method outlined in this thesis, but we do have good reason to believe that the underlying theory and implementation are sound, and that we can improve on our treatment of the problem by incorporating new features as time goes on without fundamentally reworking our approach. The principal components of their signals will remain the basis of our attempts to detect EMRIs.

Originally, the 80640-sample long waveform segments from which spectrograms (and then eigenspectrograms) were constructed was chosen almost at random, being the length segment handled most efficiently by Matlab's ODE solving process during

the orbit evolution of the EMRI analytic kludge waveforms. In retrospect this was probably a sensible choice; the approximately-two-week period that each test spectrogram covers is much shorter than the orbital period of LISA (one year), so changes to gravitational strain measurements as a result of the detector response changing throughout the course of the orbit *within the period of a test spectrogram* should be fairly small. Untreated, these changes will be an additional source of noise that will not be combated effectively by the whitening process currently used since its spectral shape will change throughout the segment, in turn reducing the SNR of EMRI signals and thus reducing the possibility of successful signal detections. As yet the magnitude of this noise is not known.

Additionally, EMRI signals with identical parameters other than the time of their detection will project into the eigenspace in different ways due to the positioning of the detector changing, but it is possible that this could be taken into account. While the PCA approach has so far considered the entire EMRI parameter space, it would be possible to consider a subspace that did not include source location parameters. Multiple sets of eigenspectrograms could then be constructed, each corresponding to a separate region of the sky, and searchable waveforms could be be assessed using each of the sets. In principal, this could provide a crude method of extracting source location information based on the signals' projection into the different eigenspaces, with the assumption that the residual measurements within the set corresponding to the correct region of the sky for the source would be smaller than in other sets. This may not hold true, however; redundancy within the EMRI problem space may mean an EMRI signals in one location actually projects into an eigenspace in an extremely similar manner to a different EMRI signal in another location, to the extent that any differences between the two are overwhelmed by sources of noise. Against this uncertainty, the extra computational cost of generating -and using- each of the eigenspectrogram sets is perhaps not a great concern, but if this did prove to be a valid method real-time searches would be beyond the abilities of a single desktop computer.

6.5 How to detect the EMRI signals

Having examined major aspects of the waveforms that we will have to search through, we have reached a stage where we need a workable method to determine whether

or not we are actually detecting a signal or not. Until now, the residual measurements gained by analysing test spectrograms have only indicated the detectability of a signal, and our prior knowledge of where the signals are (since we created the searchable waveforms containing them) allowed us to determine whether or not we were on the right path.

The coalescence time of an EMRI signal should correspond to the peak of the SNR measurements around that time, since this is the point at which the signal projects most completely into the eigenspace. Naturally, this is subject to some variation because of noise within the searchable waveform, but for a strong signal the peak should at least provide a good estimate of the coalescence time.

However, the presence of multiple EMRI signals means that we cannot simply isolate the maximum SNR measurement in a searchable waveform, nor every SNR measurement above some threshold since this will typically include measurements leading up to and following the true coalescence time because of the spiked shape caused by the presence of an EMRI signal. Furthermore, as the β waveforms showed, the interference between overlapping EMRI signals can cause the SNR of the spikes to change (and of course increasing the amplitude of the EMRI signals by changing the source distance will also change the SNR). Nonetheless, in many cases estimating the number of EMRI signals and their coalescence times seems trivial from a human perspective; regardless of their SNR, it is fairly easy to pick out likely spikes from the noise and to see where they peak. We need a way to automate this process.

The first step is to recognise that the peaks of the SNR spikes are simply local maxima on a curve formed by all of the SNR measurements. We can therefore isolate them by calculating the gradient of this curve at each point and finding where the gradient changes from positive to negative. These points are then the estimates of the coalescence times for the EMRI signals. This is a very straightforward and computationally inexpensive process, but it is not infallible. The presence of noise means that measurement-to-measurement we are not dealing with a smooth curve, producing small variations that are also local maxima in a region where there is an overall trend in the gradient of the SNR measurements.

Our second step then is to remove these small variations by setting those SNR measurements below a threshold to zero, effectively smoothing out the line in regions of low SNR. A robust method to determine the threshold level is not immediately obvious; while isolated EMRIs can produce high-SNR sharp spikes, this sharpness changes depending on the search resolution and the signal itself. We have restricted

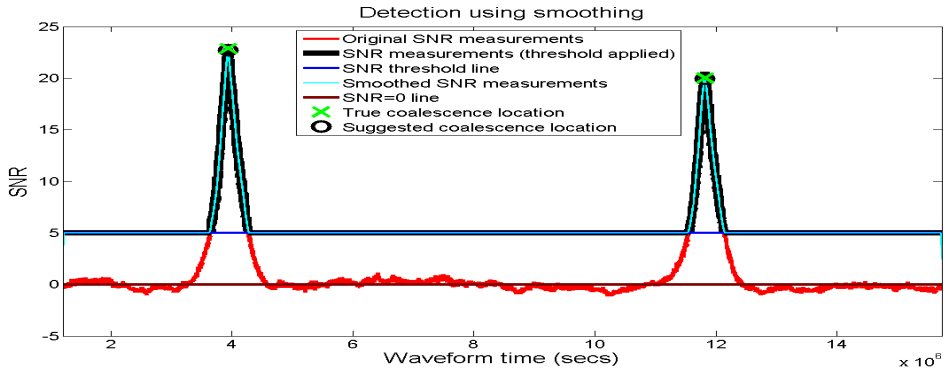
ourselves thus far to a limited subspace of the EMRI parameter space, and while the resulting waveforms do seem to produce an almost characteristic shape as they project into the eigenspace, we cannot be sure that this will hold true using different eigenspaces. Therefore, we cannot simply disregard estimates by demanding that the SNR rises and falls by a some amount within a certain time surrounding a local maxima (the grammar here betrays this fact; appropriate choices for 'some amount' and 'a certain time' are unknown).

Instead, we smooth the curve of the SNR measurements by convolving it with a gaussian window before determining the local maximum. This averages out the small variations that might cause problems, albeit at the risk of smoothing out some genuine spikes cause by the presence of an EMRI signal. Fig. 6.31(a) shows this method applied to a waveform, and the points of local maximum, which we take as our estimates of the EMRI signal coalescence times.

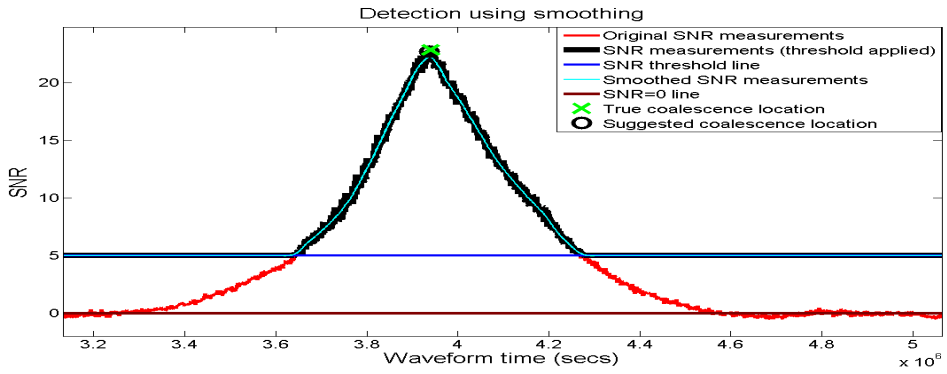
When applying this process, it became clear that using different window widths could distort the overall shape of the SNR curve, and that these windows would also be spanning a different period of time depending on the search resolution (in effect the resolution of the curve). Furthermore, the use of a gaussian shape for the window had been selected primarily for computational convenience (being a built-in function in Matlab), and it would be possible to achieve slightly different results using different shapes. Nevertheless, we had assumed that the method would be crude, and at this stage of the analysis we did not have sufficient time to exhaustively check the effects of changes to the parameters of this method, nor to fully explore different options (a more thoughtful approach in suggested in Section 7.3, but could not be implemented properly in time). Therefore our definition of 'a detection', and the application of the smoothing process to searchable waveforms in the following section were understood to represent a first attempt only due to the pressures of a deadline, undertaken with the acknowledgement that there were a number of sources of uncertainty that would have to be examined before it would be suitable even as a proof-of-concept.

6.6 Final test

We have not been able to exhaustively examine or include the effects of every complication that a realistic searchable waveform may contain, even those we are aware of, and we are doubtless ignorant of many more. Despite this, we have tested the



(a) SNR measurements of a searchable waveform containing two EMRI signals and noise.



(b) A closer look at the SNR measurements around the first coalescence time.

Figure 6.31: SNR curve and its smoothed counterpart, with the local maximum used as estimates of coalescence times. A threshold of SNR = 5 has been applied before the smoothing to eliminate candidate measurements caused by small fluctuations.

PCA-based search method in a variety of situations, and have a good notion of what we looking for as an indication of the presence of EMRI signals.

Therefore, we construct one last set of tests exploring circumstances under which we hope to justify the use of EMRI spectrogram principal components as a basis for a search for these sources within gravitational waveform measurements.

A noisy searchable waveform containing multiple signals

The δ waveforms demonstrated a situation in which the PCA-based search was breaking down, unable to produce distinct dips in residual measurements that would indicate the presence of an EMRI signal. Although it would be perfectly possible to use the smoothing method to obtain estimates of candidate coalescence times within such a waveform, the sheer number of overlapping signals make it difficult to see all of the EMRI signals even by eye.

Therefore, for a final test of searchable waveforms we constructed ourselves, we concentrate on β-like searchable waveforms containing two overlapping EMRI signals and noise. One hundred different noise timeseries, generated in the usual manner, were added to the signals (this number was decided on based on the available time and computational resources), and the same smoothing process was applied to each one in a limited Monte Carlo process that would provide some estimates of the coalescence times. The two signals were then moved closer together until they appeared to merge. Table 6.4 gives the coalescence times used, with each κ 'waveform' actually representing all 100 searchable waveforms with these coalescence times. Because of the computational requirements of producing and analysing 100 searchable waveforms, a search resolution of 1000 samples was used ($\approx$ 4.2 hours, see Table 5.2).

One intermediate stage was applied to κ_1; different SNR thresholds were applied prior to smoothing in order see how the number of estimated coalescences changed. Fig. 6.32 shows how the number of detections drops as the SNR threshold rises, settling on a threshold value of SNR=5. This value was then used for all of the other κ waveforms, since it isolated two rough periods of time containing the majority of the estimates *and* the true coalescence times.

Widely-separated coalescence times always produced two distinct estimates with little or no spread (in most cases all one hundred searchable waveforms would settle on two SNR measurements/test spectrograms as being the most likely candidates

Searchable waveform name	t_c (EMRI signal 1)	t_c (EMRI signal 2)
κ_1	3939600	4239600
κ_2	3939600	6369600
κ_3	3939600	5154600
κ_4	3939600	5139600

Table 6.4: The coalescence times of the two EMRI signals in the κ searchable waveforms. Coalescence times are given in terms of the waveform time (in seconds).

for the coalescence times), therefore the κ waveforms were specifically constructed to have significant overlap. κ_1 is identical in construction to β_6, when the signal confusion was certainly apparent in the noise-free residual measurements; only the resolution of the search applied is different.

Remarks on the κ waveforms and the results of the searches

Although crude, the smoothing method applied to fairly strong signals does appear to produce results close to what we might have expected if we simply looked at the SNR measurements by eye. Adding different noise timeseries to the EMRI signals causes test spectrograms to project into the eigenspace slightly differently each time, so the point of maximum SNR in each spike shifts slightly and hence the estimated coalescence times shift too. Available resources prevented us from repeating the searches with better resolutions or with more searchable waveforms, so it is hard to be confident that the spread in the histograms in Fig. 6.32 Fig. 6.33 is anything other than a very rough indication of what might be produced if we had the luxury of more time. However, the behaviour of the histograms as the true coalescence times get closer together is interesting. Distinct regions of estimates being to merge and it becomes first difficult and then impossible to tell them apart, replaced with a single region with many possible coalescence times (as shown in Fig. 6.33 with the κ_4 waveform).

Again, this is expected behaviour. If we consider the separation of the EMRI coalescence times, given a search resolution of 1000 steps and a step size of 15 seconds, only κ_1 has the signals separated by significantly longer than the width of the window used to smooth the measurements. κ_2 separations are only somewhat larger than this size, and κ_3 is approximately equal to the window width. By κ_4,

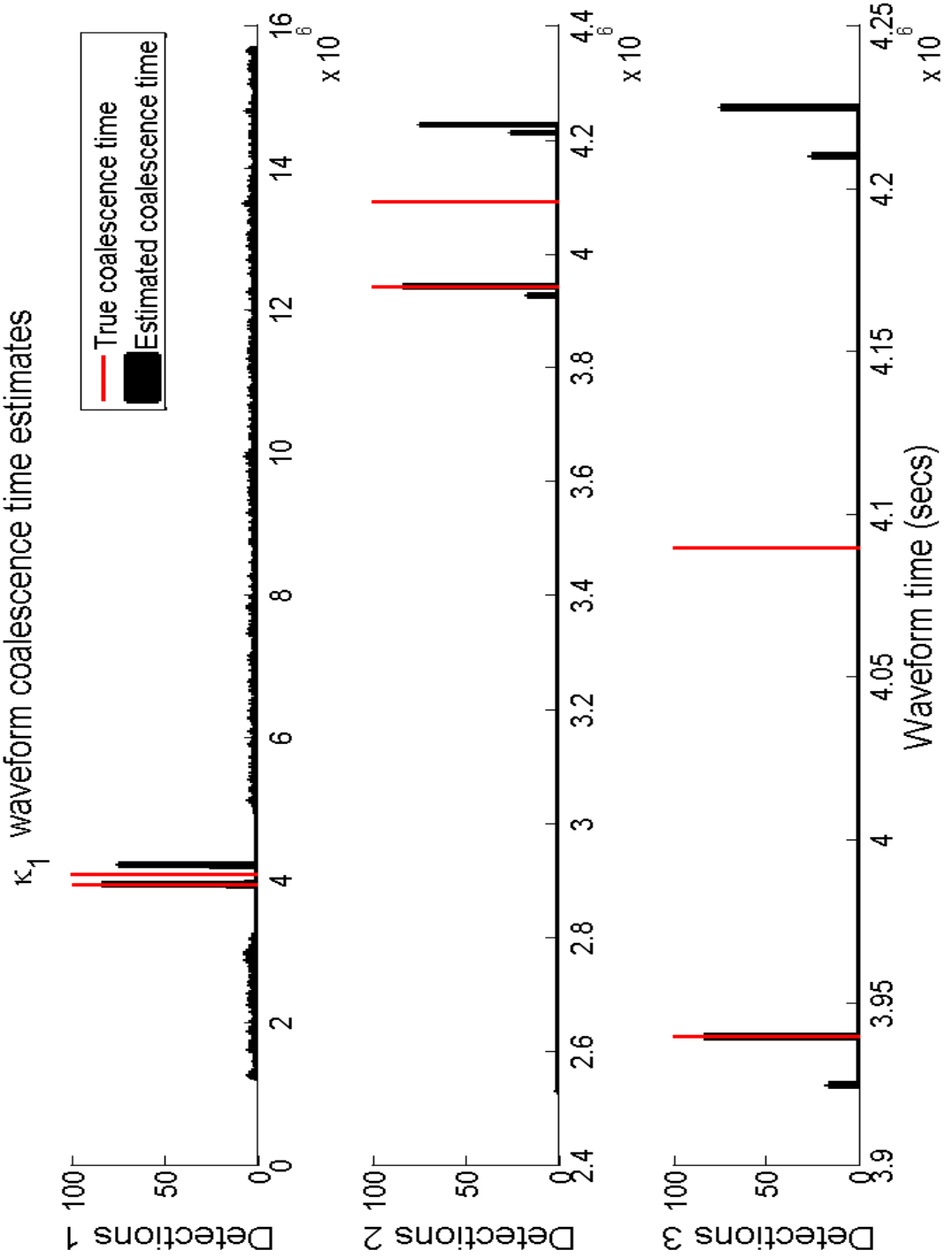


Figure 6.32: Histograms of the estimated coalescence times of the 100 κ_1 searchable waveforms, after smoothing the SNR measurements and applying an SNR threshold below which candidate detections were ignored. The number of estimates at a given time, labelled here as 'Detections' 1, 2 and 3, correspond to a threshold of SNR = 1, 3 and 5 respectively.

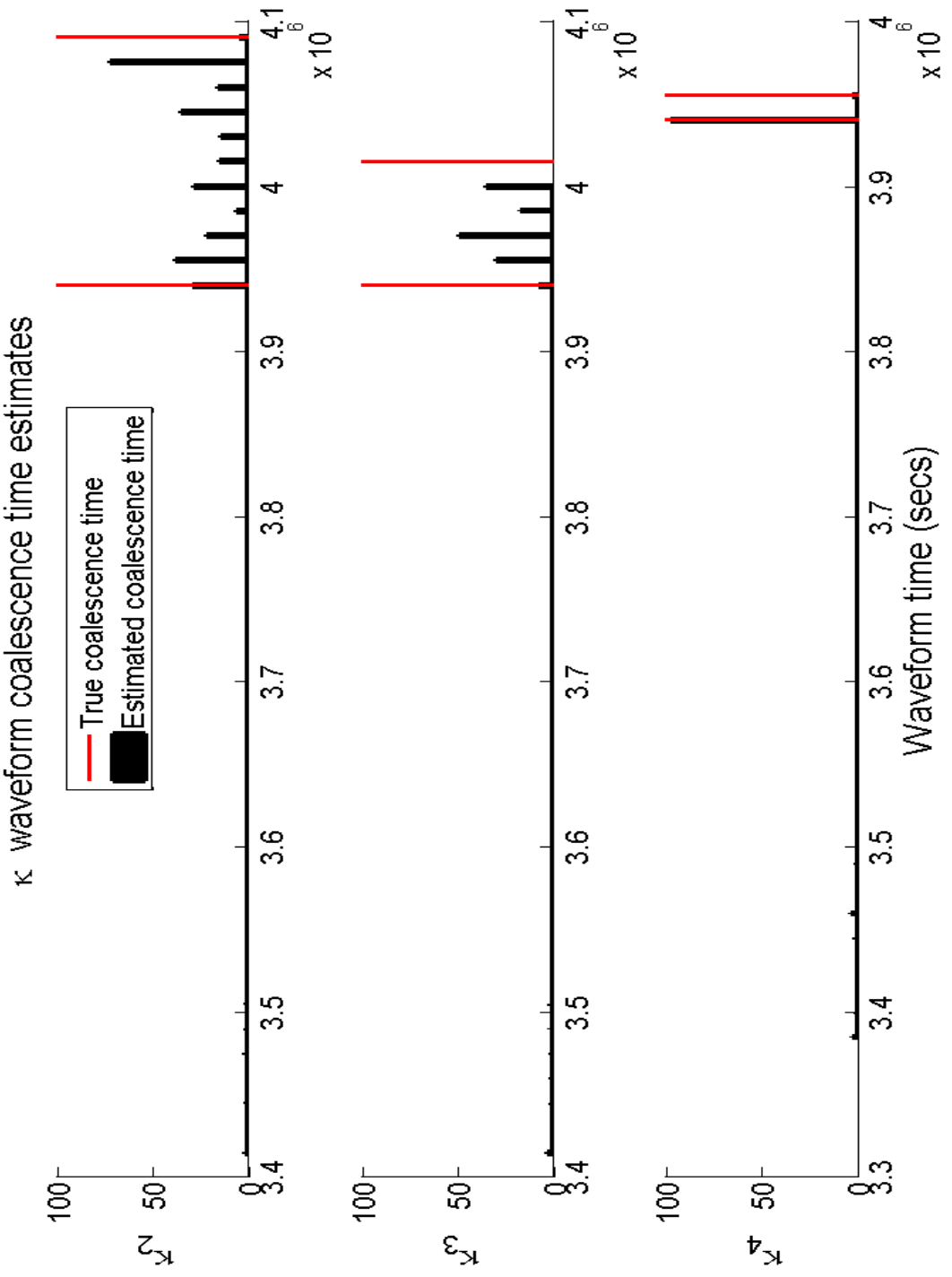


Figure 6.33: Histograms of the estimated coalescence times of the 100 κ_2, κ_3 and κ_4 searchable waveforms, after smoothing the SNR measurements and applying an SNR threshold of 5.

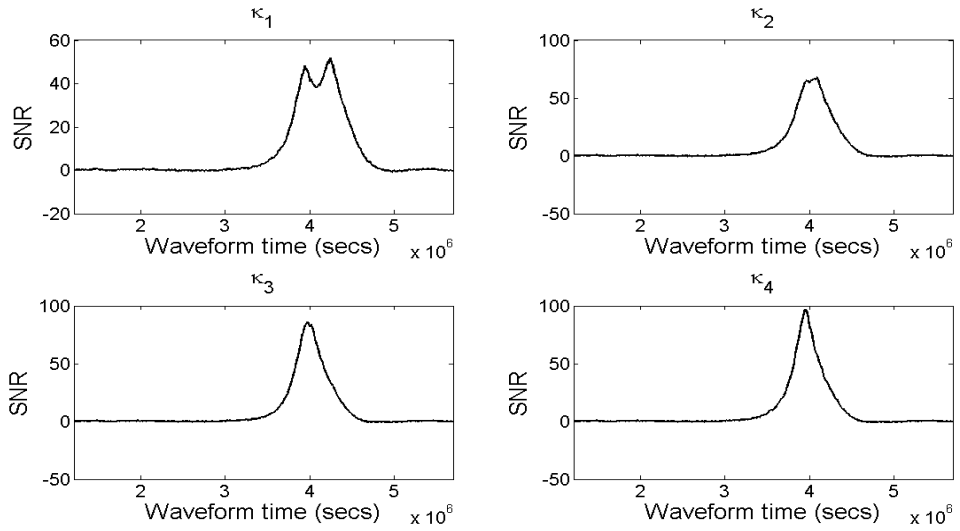


Figure 6.34: SNR measurements of the κ waveforms. As the overlap between the EMRI signals increases, the previously-distinct local maxima merge and become indistinguishable.

the two coalescences are separated by only 1000 data points in the timeseries, and it is therefore no surprise that the smoothing will see this as being part of the same feature. At this point, it seems clear that the spread in coalescence time estimates are caused by the changing noise used in each of the κ_4 searchable waveforms.

Certainly, our eyes can do no better by this stage. Fig. 6.34 shows the SNR measurements from one example of each of the κ waveforms. κ_1 clearly contains two distinct peaks, and they can be seen with difficulty in κ_2, but beyond this point there appears to be only one feature.

Thus we remain confident that the PCA-based search and the resulting measurements will allow us to make quick estimates of the coalescence time of EMRI signals provided that they are sufficiently separated in time, but it requires a more sophisticated effort to increase the accuracy of the estimates (conceivably, one simple step towards this goal would be to attempt the searches with better resolution). In addition, when faced with a cluster of points as given by the κ_4 histogram, we would be able to determine how likely it was that we were seeing a multitude of overlapping sources, or simply a couple of sources affected by noise.

Similarly, a better method of determining the existence and significance of local

maxima is desirable. Even above a certain threshold there are persistent estimates that (because we constructed the waveforms) we know are false positives, and the current SNR threshold was applied with little thought (and considerable prior knowledge about the expected output). There is also a bias towards the estimates being applied to the first of the coalescences when the signals are close together; as the first drops out the SNR falls, causing the spike to appear to peak at the first coalescence. These problems would be particularly troublesome in situations such as those in the δ waveforms, and if this is a likely scenario then there is a great deal more work necessary before the methods used so far could provide reliable estimates of the existence and coalescence times of EMRI signals.

Chapter 7

Conclusions

Previously, we stated that the analysis of gravitational wave data posed a formidable task, with some different challenge from those inherent in observing electromagnetic signals (in addition to the extreme difficulties of measuring the data itself). Thus, our intention was to investigate unconventional analysis techniques with an emphasis on quick search methods. Our goal was to determine what, if any, useful information could be extracted from simulated gravitational wave data from such searches.

From the beginning, there was no realistic expectation that our simple approaches would return all of the information available from each source, nor that the signals produced by these sources would necessarily be yield to the proposed search methods. While these predictions proved largely accurate, the analysis did produce some surprises and suggested that certain aspects of the work did merit further consideration.

The following chapter provides an overview of the results of the research in this thesis, and attempts to place them within the context of the wider gravitational wave data analysis community and their efforts. This is followed by a number of recommendations regarding future work on the research presented.

7.1 Thesis summary

The rapid search in Chapter 2 proved to be a rather inaccurate method for detecting SMBH inspiral signals, and required too much user input to be deemed robust. Thresholding the sum power measurements over a user-defined level, isolating a particular band of frequencies to examine and automatically merging the inspiral times for candidate detections within a certain period of time; the grammar itself

betrays the subjective nature of the process. In short, our approach was too simple. Nevertheless, the rapid search did highlight some of the difficulties in manipulating the gravitational wave data. Furthermore, it made us consider a visually-orientated approach to detection; we relied on a characteristic shape of a signal in the MLDC challenges which we could see by eye in the spectrograms of the datasets, at least for strong signals.

We carried this focus over in our consideration of the EMRI gravitational wave signals. With a much wider parameter space to consider, our least-squares fit method employed in the rapid searches of Chapter 2 was going to be computationally unfeasible, but the inability to see the waveforms within MLDC challenge spectrograms by eye demanded an alternative approach. It was the high signal-to-noise ratio of the SMBH binaries that allowed us to extract the shape of the signal from the background. Only by examining the analytic kludge waveform spectrograms themselves (such as those in Fig. C.22) and noting their visual similarity did we convince ourselves that there might be exploitable features within the EMRI signals.

The principal component analysis outlined in Chapter 4 provided a framework with which we could measure the similarity of EMRI spectrograms and construct a new parameter space spanning them. Although it would have been possible to immediately apply this theory to EMRI spectrograms, we could not predict the likely behaviour of the iterative PCA method constructed; it would have been difficult to locate faults in our algorithms. Thus we decided to implement our process using spectrograms of simple sinusoids; their construction was fast and easily understood, and a variety of different situations could be explored in a well-controlled manner. Additionally, we believed that they were sufficiently similar to the EMRI spectrograms (in particular the quasi-monochromatic sinusoids of Section 4.2.1.3). Only after we assured ourselves that the PCA process behaved as we believed it should did we feel confident about examining the EMRI spectrograms.

Results of the initial tests of PCA on EMRI spectrograms convinced us that the problem space could not be easily spanned, and that our focus would be necessarily restricted to a small region of parameter space (rather, in the context of PCA, a small number of training points from the problem space and the space spanned by the signals they contained). A large number of aspects of the EMRI signals required attention

With a firm grasp of the behaviour of the PCA process and the response of EMRI spectrograms to it, we were then able to turn our attention to the prospect of

actually searching for signals of this type within a dataset. Our search method also required testing under a number of circumstances, each producing slightly different output that we sought to understand.

Finally, Chapter 6 detailed attempts to make the searchable waveforms more realistic, and determine the performance of the search method in more challenging situations. Our efforts to ameliorate the difficulties encountered by the original iteration of the search method, and convert the output into a detection algorithm are included as well.

7.2 Using the results

The collaborative nature of gravitational wave data analysis has been stressed throughout this work, and the research within would certainly be impossible without the firm foundation of others within the wider community (in particular, the analytic kludge waveform generation method). Therefore, we must consider how best to use the results to contribute to the wider pool of knowledge.

We were unable to completely span the EMRI problem space with available computing resources, but an extension of the PCA applied to a large parameter space (as demonstrated in Fig. 4.47) which would completely span this space (to within a given residual threshold) could reduce the necessary computational power needed to perform different searches from the method describe in Chapter 5. Because the PCA reshapes the parameter space that has to be examined, it makes sense to conduct searches within this new space rather than attempting to convert our results back into the original parameter space.

Alternatively, our estimates of the coalescence times and the number of 'events' are most useful as starting points in a more thorough search. We have already ruled out exhaustive searches through EMRI parameter space, but they would be useful initial states of an MCMC search within this parameter space. Although its use is much more widespread within the field of social sciences, PCA has not received a great deal of attention in the gravitational wave data analysis community; as a result, PCA-based searches may be difficult to integrate with other types of analysis.

A significant drawback to the PCA is that by throwing out the redundant information, it is difficult to extract parameters from a signal based on its projection into a constructed eigenspace. The weights of a projection do provide a 'fingerprint' for

that signal, but it is not unique except in those cases where the eigenspace is as large as the original parameter space (making the PCA a rearrangement of the original information without discarding anything). In its current form, spectrogram classes can provide a starting point for all parameters defined by the original parameter space, but there is the difficulty of deciding which signals actually define the classes and how distant they should be from each other; if multiple signals widely separated in the original parameter space project into an eigenspace in an identical manner then there are multiple choices for the parameter values used to define that class.

Despite the challenges of constructing the EMRI eigenspace, the PCA-based search does provide some quick estimates, with a minimum amount of prior knowledge about the signals it is searching for. In these respects, we have maintained our stated focus of our research into gravitational wave data analysis. The results show that our approach is unlikely to supplant current analysis, but there is hope that it may prove complimentary to these efforts or can be utilised by them as initial estimates (see the next section for a roadmap of future work). Nothing in the work done suggests that PCA is a limited method, or that there is some inherent difficulty in applying it to EMRI spectrograms, but further work is required to prove that it is a useful tool for gravitational wave data analysis.

7.3 Future work

As presented, the PCA-based search algorithm is not a robust method for detecting EMRI gravitational wave signals in its present incarnation. However, the results from the initial tests do suggest that the signals are amenable to PCA and that the parameter space does contain redundancy that can be exploited. Thus, there is no fundamental reason to advocate abandoning this avenue of research in order to develop the method further. A number of limitations have already been identified in the main body of this thesis, and this section suggests some improvements that must be resolved before we can have real confidence in the viability of our approach.

It should be reiterated that PCA is all ‘up-front’ effort; the construction of the eigenspace is by far the most computationally expensive part of the process, and must be done iteratively. As the eigenspace expands, the process becomes slower and

more memory intensive. In Section 4.3.3, we stated that the necessary resources were beyond that of a standard desktop computer, but there is room for improvement to our procedures if we could access greater resources.

Generating the EMRI waveforms is particulary time consuming, but in fact all we require is that these waveforms are presented to the PCA algorithm in the correct format (that is, a timeseries with the correct sample frequency in a data format that Matlab was able to read and store as an array). Therefore, it would be possible to have a CPU dedicated solely to producing test waveforms and passing them to another CPU dedicated to producing the eigenspectrograms. It is also possible to have multiple CPUs generating their own smaller eigenspectrogram sets before combining them, in much the same way as shown in Fig. 4.41 but eventually combining multiple $\mathbf{E_4}$ sets to make a larger eigenspectrogram set. Only the redundant information is discarded at each stage, and there will be redundancy in the combined sets that must be removed by calculating their eigenspectrograms.

Eventually, we still require a system that can determine the eigenvectors of the L matrix (see equation 5.1 - 4.11). Further, projecting new signals into the eigenspace means that the entire set of eigenspectrograms must be able to be held in memory or loaded in piece by piece. The former method is faster but more memory-intensive, the latter uses less memory but is slower.

Because we were unable to completely span anything but small subspaces of the EMRI parameter space, the notion of spectrogram classes was not particularly useful. However, if we imagine a situation where the entire problem space is spanned by a constructed eigenspace, they can make a valuable contribution.

If we store individual EMRI spectrograms, we can re-project them into the now-complete eigenspace, creating fingerprints for each one. Many will be similar, and can be used to determine a number of spectrogram classes separated by a distance of at least ϵ^2 (see equation 4.16). As stated previously, new test spectrograms that project close to one of these classes in the eigenspace are not necessarily close in terms of the original parameter space. However, if we were interested in passing on starting values to an MCMC search in the original parameter space, we could advise that there are multiple possible starting points, each of which could be used to start a separate chain (in an extreme case where two signals project in the exact same way, either would be equally valid starting points).

The inclusion of spectrogram classes also reintroduces the possibility of identifying those spectrograms that would cause false alarms by projecting into an eigenspace in a reconstructible manner but far from a defined spectrogram class. Although our initial tests did not allow for this possibility, being able to identify such signals would be an important step towards making the PCA more useful.

Finally, it may be that there are small parameter ranges shared by spectrograms within an individual spectrogram class. This was not explored during the research presented, but it is possible that performing principal component analysis on the parameter values themselves would reveal common traits. In fact, this is a much more traditional approach to PCA; numerous observations of a few parameters (14, in this case), rather than in our setup where we have a few observation of an unknown number of 'parameters' (each time-and-frequency bin of the spectrograms).

So far, we have not made use of realistic noise in the searchable waveforms. The approximations used in Chapter 6 were useful to demonstrate the behaviour of the PCA search when a candidate signal was hidden in noise, but we saw that different noise profiles altered the way that a signal projected into a constructed eigenspace (the difference between the original power spectra and those with low-frequency suppression). Thus, basing a search method on that type of projection is not appropriate if the true noise profile is different. Further, when whitening the segment spectrograms we used our approximated noise profile: if we tried this when the true noise profile was different we risk obscuring an EMRI signal present in the timeseries rather than highlighting it.

Therefore, an understanding of the noise detected by LISA is crucial for constructing the eigenspace and searching timeseries for EMRI signals. Realistic noise may not be gaussian or stationary as we have assumed throughout this work, and when we include accurate orbital information for the spacecraft we will certainly see a periodic change in the noise strain measurements as they orbit the sun throughout the duration of the mission. There may even be changes to the expected instrumental noise as a result of design changes between now and the time the mission get underway.

However, the somewhat modular nature of the PCA-based search described in previous chapters is helpful here. The PCA does not know what it is using to contruct the eigenspace, it is up to the user to provide meaningful information. If we can generate more realistic noise timeseries, we can proceed in the same manner

as before, but it would be wise to repeat some (or all) of our initial tests to check that signals present in this new noise projected into eigenspaces in a meaningful manner (that is, distinct from the noise itself).

Another significant limitation is that we have not yet properly adapted our analysis to the actual strain measurements that will be recorded by LISA, instead limiting ourselves to the $h_\times$ strain produced by EMRIs as detected at the solar system barycentre. As well as the effects of the antenna pattern, we recognise that the strain measured by the spacecraft will combine the two polarisations of the gravitational wave.

In this matter, the way to proceed is not clear. If we could accurately account for the antenna pattern, we could analyse strain measurements at each spacecraft and look for coincident events. This would require determining the eigenspectrograms of combined polarisations of EMRI signals, and how this might be accomplished is not yet known. One crucial aspect of the PCA method is that the algorithm does not actually know what signals in the spectrograms actually are; as far as it is concerned they are simply an array of information (which is reshaped into a column vector). PCA reveals redundancy in the signals it is presented, so as long as we give it 'correct' spectrograms we will get meaningful output.

This has some significance beyond our concerns about what the signals look like to the detectors. The PCA method we use is essentially pattern-matching, and as a result if EMRI signals actually look different from the analytic kludge waveforms they will not project into whatever eigenspace we construct. If more accurate waveforms than the analytic kludge forms can be easily generated then they can simply replace the latter.

In a related manner, if the beginning or end of the EMRI signals exhibit behaviour that was not particularly redundant (the zoom-whirl behaviour towards the end might satisfy this concern, but this works as a general principle), this could actually be eliminated by simply not including it in the test spectrograms. As a result, the projection of a signal would be closest to the eigenspace for a section of a searchable waveform that did not include this behaviour. As a result, we would be providing estimates for when the signal began to display this behaviour rather than coalescence time. By ignoring this behaviour, we would surely be throwing away some useful information, but this might be preferable to not being able to span the eigenspace at all.

Of particular concern is the ability to recognise the presence of multiple signals within a searchable waveform based on on their residual measurements (and by extension, the SNR measurements) needs significant improvement. The smoothing procedure coupled with detecting local maxima is a crude first attempt, but struggles with multiple signals or those with significant overlap. However, there is considerable expertise within the gravitational wave data analysis community with regards to Bayesian analysis which might help implement a more meaningful search method. Currently, efforts concentrate on follow-up studies of candidate detections (54), and have applied MCMC searches to MLDC data sets (55), and there are a number of sophisticated refinements to the basic MCMC search method (see Section 1.2.4.3). Nevertheless, Simha (X..a) investigated a Bayesian approach to determining the number of discrete sinusoids within a data stream (with an eye towards applying the method to LISA data), a process that might be readily adaptable to the EMRI search SNR measurements, revealing how many EMRI signals there are within a searchable waveform. Though basic, it is valuable introductory material given the similarity of the problem it considers to the EMRI signals and provides an instructive initial framework for a search.

If we assume that the presence of an EMRI signal within a searchable waveform produces a 'characteristic' shape in the recorded SNR measurements (we have already seen the spikes, asymmetrical about the coalescence time, in our initial tests; see Fig. 6.20 and Fig. 6.19(a)), we can model this data $\mathbf{D}$ as a sum of $\mathbf{M}$ such shapes with different coalescence times and amplitudes, along with random Gaussian noise n. The model data for a particular combination of coalescence times and amplitudes, $\mathbf{F}$, is a summation of these shapes without the noise. To model the shape itself we would create a look-up table β showing the response of the PCA-based search to the presence of a single EMRI signal in a searchable waveform. Fig. 7.1 shows the normalised SNR of ten strong EMRI signals (reducing the effect of the noise on the shape), and the mean value of these ten. The relevant parameters will definitely include the amplitude of the shape's peak (A) and the coalescence time with respect to the waveform time (t_c), although there may be others that we have not considered.

Therefore, we have

$$\mathbf{D} = \sum_{k=1}^{M} A_k \beta(t - t_{c_k}) + n_k \tag{7.1}$$

and

$$\mathbf{F} = \sum_{k=1}^{M} A_k \beta(t - t_{c_k}). \tag{7.2}$$

We want to determine the probability distribution of $\mathbf{M}$, by calculating the probability of a particular value of $\mathbf{M}$ within the SNR measurement dataset. In terms of Bayes' theorem (see Section 1.2.4.3), we want to calculate

$$p(\mathbf{M}|\mathbf{D}, I) = \frac{p(\mathbf{M}|I)p(\mathbf{D}|\mathbf{M}, I)}{p(\mathbf{D}|I)}. \tag{7.3}$$

We can immediately assign a uniform prior since we have no preference for a given number of EMRI signals within a dataset, and we can express the likelihood as a marginal integral over our parameters and utilising the probability product rule

$$\begin{aligned} p(\mathbf{D}|\mathbf{M}, I) &= \int ... \int p(\mathbf{D}, A_k, t_{c_k}|\mathbf{M}, I) d^M A_k d^M t_{c_k} \\ &= \int ... \int p(A_k, t_{c_k}|\mathbf{M}, I) \times p(\mathbf{D}|A_k, t_{c_k}, \mathbf{M}, I) d^M A_k d^M t_{c_k}. \end{aligned} \tag{7.4}$$

Again, we have no preference for a particular amplitude or coalescence time, so we assign a uniform probability distribution to $p(A_k, t_{c_k}|\mathbf{M}, I)$ for amplitudes between 0 and some value A_{max}, and for a coalescence time between 0 and the period covered by the searchable waveform t_{max}.

The last hurdle is to determine $p(\mathbf{D}|A_k, t_{c_k}, \mathbf{M}, I) = p(\mathbf{D}|\{\mathbf{F}\}, \mathbf{M}, I)$, which we do by assuming that each SNR measurement is independent and that the noise in each is representable using a gaussian distribution. Hence,

$$p(\mathbf{D}|\{\mathbf{F}\}, I) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-\left(\mathbf{F}-\mathbf{D}\right)^2}{2\sigma^2}\right] \tag{7.5}$$

and therefore

$$p(\mathbf{D}|\{\mathbf{F}\}, M, I) \ \alpha \ \exp\left[\frac{-(\chi)^2}{2}\right] \tag{7.6}$$

with

$$\chi^2 = \sum \left(\frac{\mathbf{F} - \mathbf{D}}{\sigma}\right)^2 \tag{7.7}$$

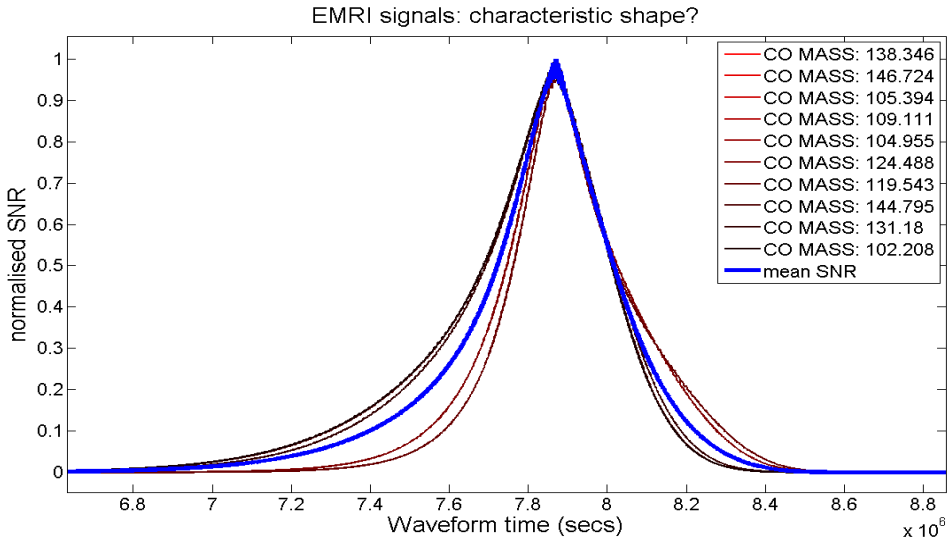


Figure 7.1: Normalised SNR measurements of ten searches for EMRI signals with randomly-chosen CO mass values (all other values correspond to those in the set Υ). The coalescence time for each signal is the same ($t_c = 7869600$ seconds), and the eigenspace they are projected into is that constructed from the set Υ. There is no definite characteristic shape to the measurements, but they are very similar and all ten searches are not distinct. The mean normalised SNR is also included (in blue), a first approximation of a characteristic shape.

where the summation is over the **i** SNR measurements (from the **i** segments of the searchable waveform).

Finally, we now have

$$p(\mathbf{M}|\mathbf{D}, I) \;\alpha\; \frac{1}{(A_{max}t_{max})^M} \int \ldots \int \exp^{\left(\frac{-\chi^2}{2}\right)} d^M A_k d^M t_{c_k}. \tag{7.8}$$

To find the probability distribution for the number of sources, we perform the integrations over the defined parameter ranges for different values of **M**, and plot the probabilities. The peak of this plot should reveal the most probable number of EMRIs within the searchable waveform.

A number of problems immediately suggest themselves. Firstly, our noise is not independent since the measurements are based on searchable waveform segments

with a great deal of overlap. However, the major drawback is our uncertainty about the suitability of our model data. We do not have a clear understanding about its form, and the mean characteristic shape in Fig. 7.1 would only be a crude approximation. Further, we are basing even this on results from our initial tests; a spanned eigenspace might create an entirely different shape when spectrograms are projected into into it. It may be that a characteristic shape does not exist at all, or that several further parameters are required to defining it, increasing the scale of the numerical integration required (it seems likely that some sort of 'width' parameter for the characteristic shape would be required).

Unfortunately, even if we are comfortable with these uncertainties, the computational effort of numerically integrating over the desired parameters may be far too large to handle. The range of amplitudes itself is likely to be very large (as seen in Fig. 6.20) and the coalescence times could be at any point in the data stream (recall, approximately one year of LISA data $\approx 2^{21}$ data points at the design sample rate). If these are indeed the only two significant parameters, we still require an integral sum over $2\mathbf{M}$ parameters.

This is not the end of the matter, however, since we could carry out an MCMC search throughout our parameter space to determine the correct probability distribution. Once again, techniques already developed by gravitational wave data analysts could help make performing this kind of search more effective. Several improvements to a basic MCMC have been explored, including reverse jump MCMC (where the characteristic shape model itself is a parameter and the search can jump between different model), and parallel tempering (changing the likelihood surfaces to allow easier access to hard-to-reach areas of the parameter space); (56) and (24) provide an introduction to these techniques and how to implement them, while (57) and (58) discuss the use of the Blocked-Annealed Metropolis Hasting algorithm (BAM) and Delayed Rejection schemes respectively. Nevertheless, even in a rudimentary form, this approach seems extremely computationally expensive, and a more careful analysis of the situation would be necessary before we attempt to implement the outlined approach. This would also provide an indication of which improvements to the basic MCMC search could be usefully applied, since the aforementioned techniques are just some of many available. Finally, we may even decide to consider other types of algorithm that examine multi-modal likelihood surfaces such as MULTINEST (59).

A roadmap for future work

So far, we have only been able to apply our PCA-based approach to a small number of situations involving EMRI signals, as well as our basic tests on sinusoid-based spectrograms. Whether dealing with limited parameter spaces or approximations of noise present in our timeseries, we have been conscious that our techniques, however promising, will require further work before we can use them to produce astrophysical estimates. This final section presents a brief guideline for how to go from our current efforts towards this goal.

- Firstly we must extend the parameter space from which we extract our trial spectrograms and create an eigenspace. Doing so will show the long term behaviour of the residual measurements and a more accurate estimate of the number of eigenspectrograms needed to span the EMRI problem space that that provided by examining Fig. 4.47. To do so we will need a more efficient storage system, the patience to undergo a large number of iterations, or a larger memory capacity than available on a standard desktop computer. Ideally, we will be able to draw parameters for trial spectrograms from a significant fraction of the problem space.

- Next, we must check the behaviour of our PCA-based search method using this expanded eigenspace and a large number of timeseries containing different EMRI signals. Determining whether it follows the behaviour presented by our limited searches will allow us to feel confident about the basis of those searches, or force us to reexamine how to search for signals using the eigenspectrograms. Thus far the PCA-based searches presented in Chapter 5 appear promising, but we cannot be sure that EMRI signals will project in a distinguishable manner (compared to noise signals, say) when drawn from a larger problem space.

- If no major problems have been posed by the previous steps, we would desire definitions of spectrogram class for different EMRIs. This may allow us to determine correlations between particular parameters and their projection into the eigenspace (all high-eccentricity EMRIs may project in a similar manner, for example). If this does not occur it will inform us that it is not possible to determine characteristics of an EMRI system based on the principal components

of the resulting waveform. However, if there are clear spectrogram classes, we can use projections of new spectrograms to determine possible starting points (or significant ranges in parameter space) for other search methods, such as MCMCs.

- To make our searches more realistic, we can repeat the eigenspace generation and searches using more accurate waveforms (if they become available) as well as more accurate representations of the noise detected by LISA, and the effects of the antenna pattern on signals. This will allow us to present a more realistic scenario for our search method to work on, and more authentic representation of spectrograms' projection into an eigenspace (as well as a more complete eigenspace). Further, we would want to use the information from both gravitational wave polarisations rather than discarding one.

- We would then look at the EMRI signals and their principal components under a number of circumstances, to see how our techniques worked in different regimes (other than our test case, where we examined a two-week period leading up to the coalescence of the compact object and the supermassive black hole). Amongst others, we would look at the principal components of spectrograms covering different lengths of time (shorter, and longer, and at different times leading up to the coalescence), examine the effects of changing the sample frequency, and more thoroughly investigate the effect of signal-to-noise on the projection of signals into the eigenspace. It would also be worth applying PCA to gravitational wave signals from different types of sources: the same approach may be able to be used to search for these sources, and there may be some overlap of eigenspectrograms with those of the EMRIs, which would need to be examined carefully.

- Finally, we would look at our 'detection' method again, in an attempt to find a better way to determine coalescence times and numbers of sources more accurately. This may involve refining the smoothing technique, or adopting a Bayesian framework of some kind. This would be a significant change to our current situation, since our detection method is in a rather rudimentary stage.

Ultimately, the desired end product would be a bank of eigenspectrograms that spans the EMRI problem space. Additionally, we would be able to account for

or adapt to the distortions to EMRI signals caused by noise and the effects of LISA's orbit. Then the projections of LISA data into the eigenspace could be used to indicate the presence, and number of, EMRI signals in that data, and provide estimates for the coalescence times of the signals. Depending on the nature of the projection, we may also be able to suggest possible parameters for the EMRIs, or at least regions of the parameter space they might inhabit.

This information would likely be passed to a matched-filtering search, using templates generated by selecting parameters suggested by a MCMC search through the original parameter space (or an alternate search technique based on moving around the parameter space in such a manner that we home in on the correct parameter values). The PCA-based search would only be a part of a hierarchical search for EMRI signals, but would prevent us from searching exhaustively through the original parameter space.

There are significant obstacles to turning our PCA-based search method into a viable means to detect the presence of EMRI signals within LISA data, still more if we want to produce estimates for anything beyond the coalescence times of our detections. However, we still believe that the basic procedure is worth taking further, and will be able to contribute to the wider gravitational wave data analysis community's efforts confirm the existence of gravitational waves. The inherent difficulties in making such a confirmation should not dampen our enthusiasm for pursuing that goal.

Appendix A

Recalculating noise

In Chapters 2 and 6.2 we required the ability to add noise to particular datasets, without an obvious method for generating a suitable timeseries from scratch. Therefore, we attempted to generate noise datasets based on existing examples of noise.

We required a noise curve; strain amplitude measurements at given frequencies covering the frequency range under consideration ($0-33$ mHz in the case of the MLDC datasets). Thankfully the MLDC provided an accompanying noise-free timeseries for each of the noisy training datasets (useful for testing analysis techniques in more idealised situations), allowing us to extract a noise-only timeseries.

It was then possible to extract a noise curve from a spectrogram of this noise-only timeseries, averaging the strain power throughout the period covered by the timeseries, from which we could then determine the average strain amplitudes easily. The stationary nature of the LISA noise allowed us to do so without worrying that we were averaging over some modulating effect.

The noise was also designed to be gaussian in nature, and the absolute values of our strain amplitudes provided a measure of the standard deviation σ of the noise curve at each frequency. Therefore, we generated a vector of normally-distributed random numbers with standard deviation equal to the σ values. A partner vector of imaginary numbers representing randomly chosen phases was also generated at this time, in the same manner. The two vectors were combined into a single array with each frequency having a real and imaginary component to the strain amplitude, and an inverse Fourier transform of this array produced a new noise timeseries, with the same spectral shape as the original. To construct a new noisy timeseries containing

a signal, we had only to add the noise-free timeseries with that signal to our new noise timeseries.

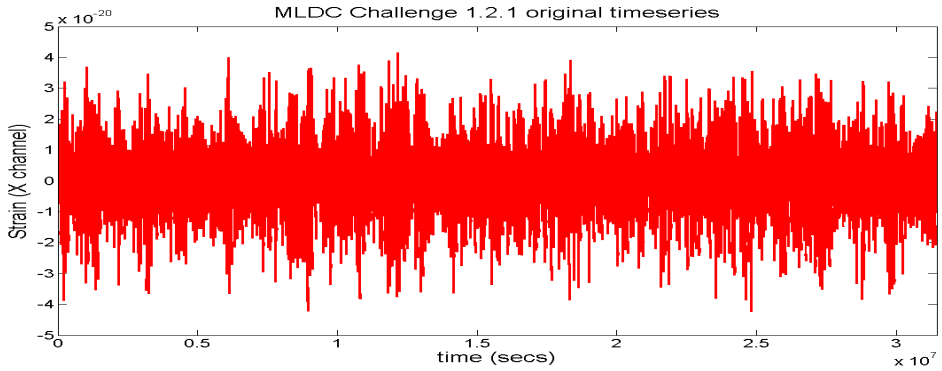
The key benefit was flexibility. The algorithm used merely required a noise curve, but was indifferent as to its origins. Therefore, the difference between the MLDC Challenge datasets 1.2.1 and 2.2 for example simply required regenerating some noise used the relevant noise-only dataset (in this manner, the presence of additional signal sources such as the white dwarf binaries in Challenge 2.2 did not trouble us since they formed the signal part of the operation, and were removed from the original noisy timeseries in exactly the same way as the signal timeseries lacking these sources in Challenge 1.2.1).

Fig. A shows an original timeseries (that of MLDC Challenge 1.2.1) and a newly-generated noise. The spectrograms of those timeseries is shown in Fig. A.

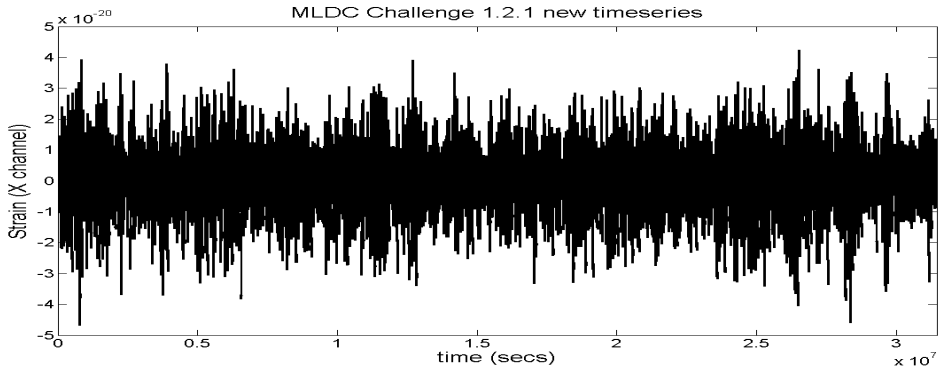
In Chapter 2, the noise in the MLDC datasets was originally generated by the *LISA Simulator* ((W..t),(60)) or *Synthetic LISA* ((W..r), (61)) programs, but at the time of creating the Rapid Search Method we were unable to integrate either of these programs into our search algorithms. Thankfully, the inclusion of noise-free datasets along with the noisy datasets allowed us to determine noise-only datasets by simply subtracting the former timeseries from the latter, after which we were able to follow the 'recipe' above.

When we were required to generate noisy timeseries in Section 6.2, the approach was fundamentally the same as before, except that we did not have to extract the noise-only timeseries from MLDC datasets. Instead we used a timeseries generated by using the 'Sensitivity Curve Generator for Spaceborne Gravitational Wave Observatories' ((W..q), which produces strain amplitude measurements, (62) and (63) describe the procedure in greater detail).

Being given the strain amplitude measurements to begin with removes the need to deal with averaging strain power measurements, but the response of the detector has been averaged over source direction and gravitational wave polarisation, and therefore provides only an approximation to the real noise timeseries that might be detected by LISA. However, we have already introduced approximations such as ignoring the contribution of gravitational wave signals from other non-EMRI sources, and treated the $h_\times$ and h_+ polarisations of the gravitational wave as completely separate. Thus, we had already resigned ourselves to the fact that analysing searchable

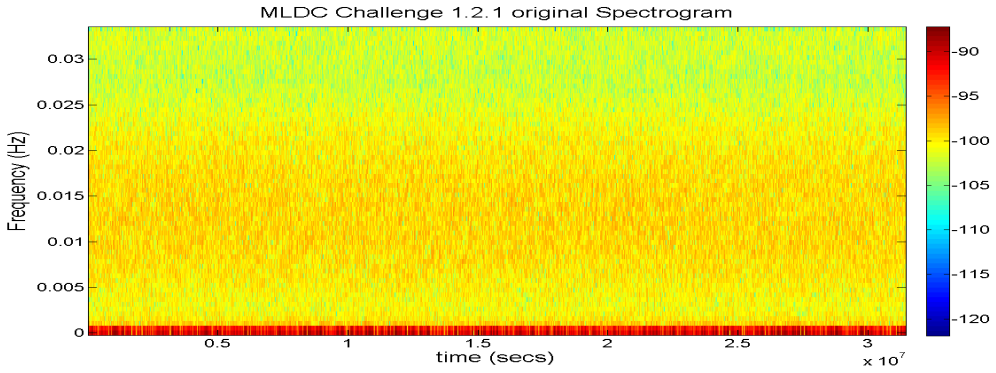


(a) Original MLDC Challenge 1.2.1 noise timeseries

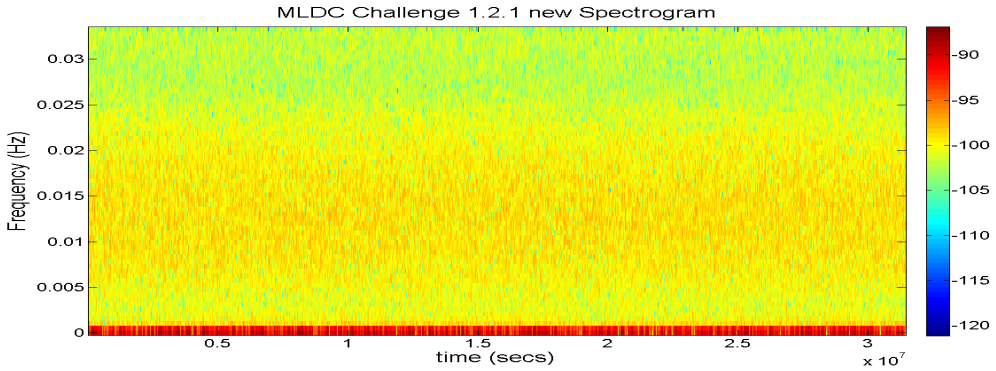


(b) Newly-generated noise timeseries

Figure A.1: The noise timeseries given by MLDC Challenge 1.2.1 (red), and a new noise timeseries generated strain from this (black).

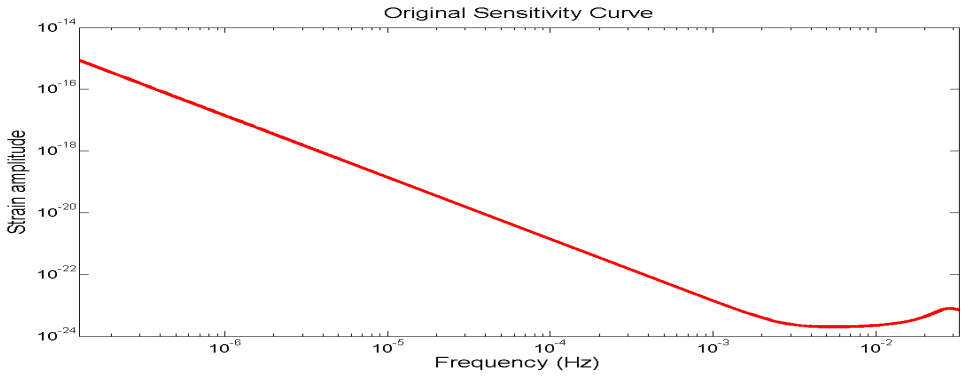


(a) Original MLDC Challenge 1.2.1 spectrogram

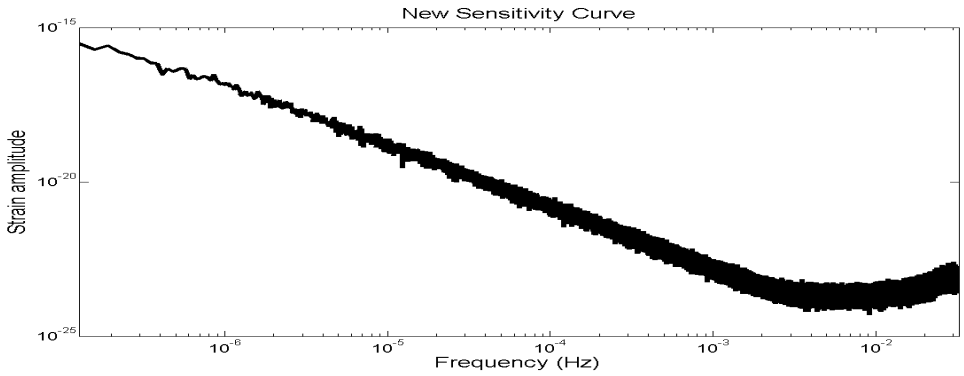


(b) Newly-generated spectrogram

Figure A.2: A spectrogram of the original noise timeseries given in Fig. A.2(a), and a spectrogram of the regenerated noise timeseries in Fig. A.2(b).



(a) Original strain amplitude sensitivity curve



(b) Newly-generated strain amplitude sensitivity curve

Figure A.3: A strain amplitude sensitivity curve from the Sensitivity Curve Generator (red) is shown in Fig. A.3(a), while Fig. A.1(b) displays an average of ten regenerated strain amplitude spectrums, showing that the two match closely.

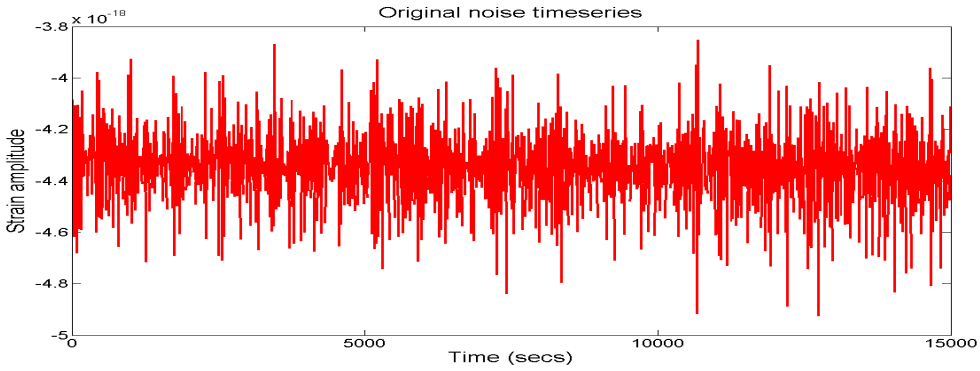
waveforms including EMRI signals and noise was only going to provide an estimate of the output we might produce if the PCA-based search was applied to real data.

The final difference between the regeneration in this situation was that the original amplitude strain did not provide the desired frequency resolution, requiring us to use linear interpolation to achieve this. While this is another source of uncertainty, we felt that it would not be a significant one.

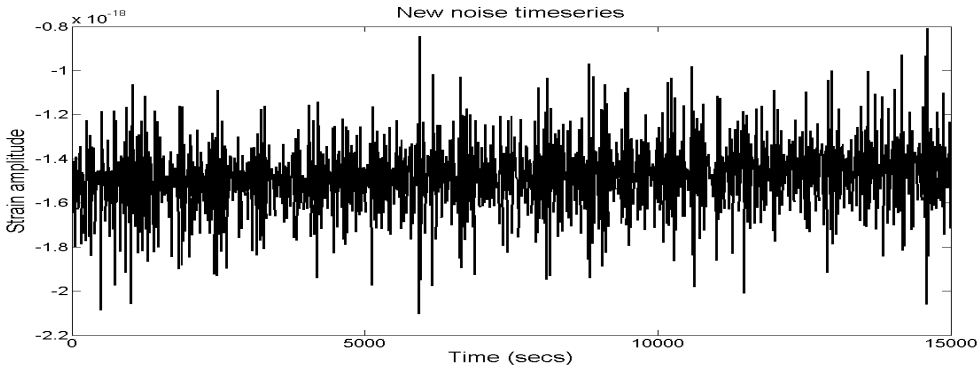
Fig. A shows the strain amplitude noise curve used to add noise to the EMRI signal searchable waveforms, and the average of 10 noise curves constructed from newly-generated timeseries, while Fig. A shows a timeseries generated using the original noise curve according to Larson's generator as well as one using the method described.

The drawback to regenerating and adding the noise in this manner is that there is an element of treating the process like a black box. It was beyond the scope of this thesis to thoroughly investigate the noise in LISA; its inclusion was designed to provide support for proof-of-concept of the SMBH rapid search and the PCA-based EMRI search. In this regard, the approach used provided suitable approximations of real noise consistent with those previously used, but it does not speak to the appropriateness of those original sources.

Nevertheless, one of the strengths common to both searches is their modular construction. Changes to the expected noise, whether a result of new TDI combinations or an alteration in the strain sensitivity, can be easily incorporated into the algorithms. Although some work was done regarding the effects of whitening the noise (and noisy signals) and low-frequency noise suppression, a more exhaustive examination is required to provide strong support the two search methods outlined in this thesis. Furthermore, the expertise of those involved in constructing the original noise timeseries allows us to be confident in the suitability of our methods.



(a) Original sensitivity curve timeseries



(b) Newly-generated sensitivity curve timeseries

Figure A.4: A noise timeseries generated using the Sensitivity Curve Generator (W..q) (in red), and a new noise timeseries from the regenerated strain amplitude spectrum (in black). The first 1000 points are displayed.

Appendix B

The 'eig' function in Matlab

As presented in this thesis, the calculation of the eigenvectors and eigenvalues using Matlab is treated as a black box; we trust it to return meaningful, accurate output but do not consider the precise method of operation. While the alternative - designing a solving algorithm from scratch - was impractical, it seemed prudent to include some additional details in the event that concepts within the main body of work were reused or continued with different software.

The intent here is *not* to suggest that the author has a full understanding of the algebraic techniques involved in the solving algorithm. Time constraints prevented a more thorough investigation of the algorithm and the sources of error that prevented residual measurements of training spectrograms to better than $R \approx 10^{-14}$; the assumption being that the theory behind the methods presented in the thesis were not at fault, and that the next step would be to look more closely at the precise methods of calculation.

From version 6.0 (R12), the Matlab software uses the Fortran90-based LAPACK routines library (Linear Algebra PACKage) for a number of linear algebrabic manipulations of data; the 'eig' function which produces the eigenvalues and eigenvectors of relevant matrices itself calls different subroutines depending on the form of the matrix. Specifically, our inputs were designated the domain of the Symmetric Eigenproblems (SEP) subroutines. At this stage the documentation is a little unclear, but the specific subroutine is labelled DSTEQR, an application of the QR algorithm.

The QR algorithm itself is a process to provide a Schur factorisation of a matrix,

from which the eigenvalues and eigenvectors can be extracted easily. This involves iteratively describing the matrix as the product of an orthogonal matric and an upper triangular matrix until there is a convergence towards one triangular form for the matrix, and specifically this process involves first transforming the matrix to upper-Hessenberg form using Givens rotations. Latini ((W..u))gives an overview of the process and additional material regarding the QZ method used in the generalised eigenvalue problem.

Appendix C

Spectrograms and Eigenspectrograms

C.1 Monochromatic sinusoid eigenspectrograms

The complete set of eigenspectrograms constructed from the set of spectrograms in section 4.2.1.1 are displayed in Figs. C.22 - C.10. The visual representation of the eigenspectrograms has been included for completeness' sake, and the meaningful information it imparts is discussed in the main body of work. To aid clarity, 'eigenspectrogram' has been contracted to 'espec' and the power measurements of each group of eigenspectrograms are displayed by the final colour bar.

C.2 Quasi-monochromatic sinusoid eigenspectrograms

The complete set of eigenspectrograms constructed from the set of quasi-monochromatic spectrograms with a uniform frequency change as described in section 4.2.1.3 are displayed in Figs. C.11 - C.20. The rate of change of frequency, $\dot{f} = 0.0977\text{Hzsec}^{-1}$ in this case. The visual representation of the eigenspectrograms does not provide

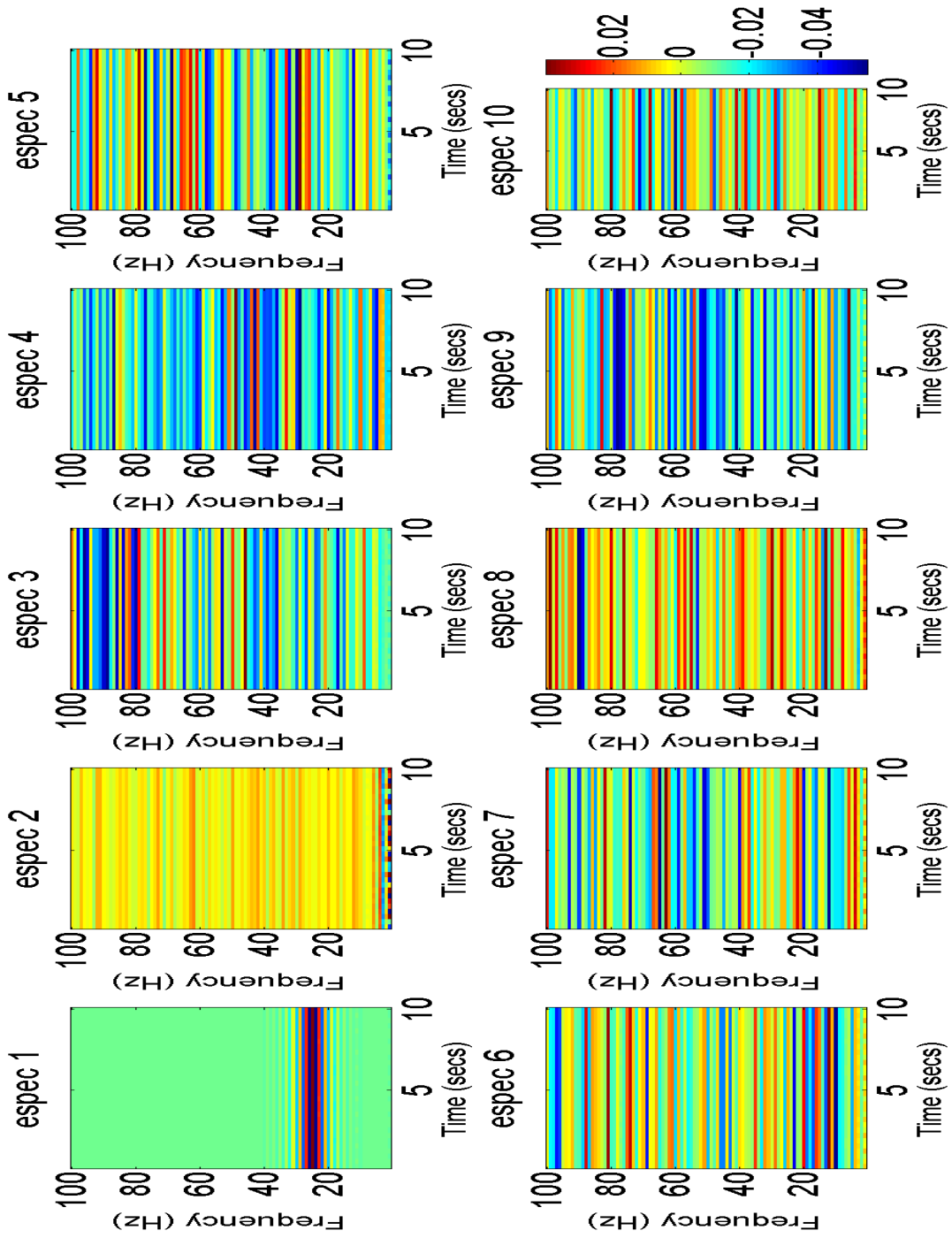


Figure C.1: The first ten eigenpectrograms generated by the set of single-frequency sinusoids (see 4.2.1.1)

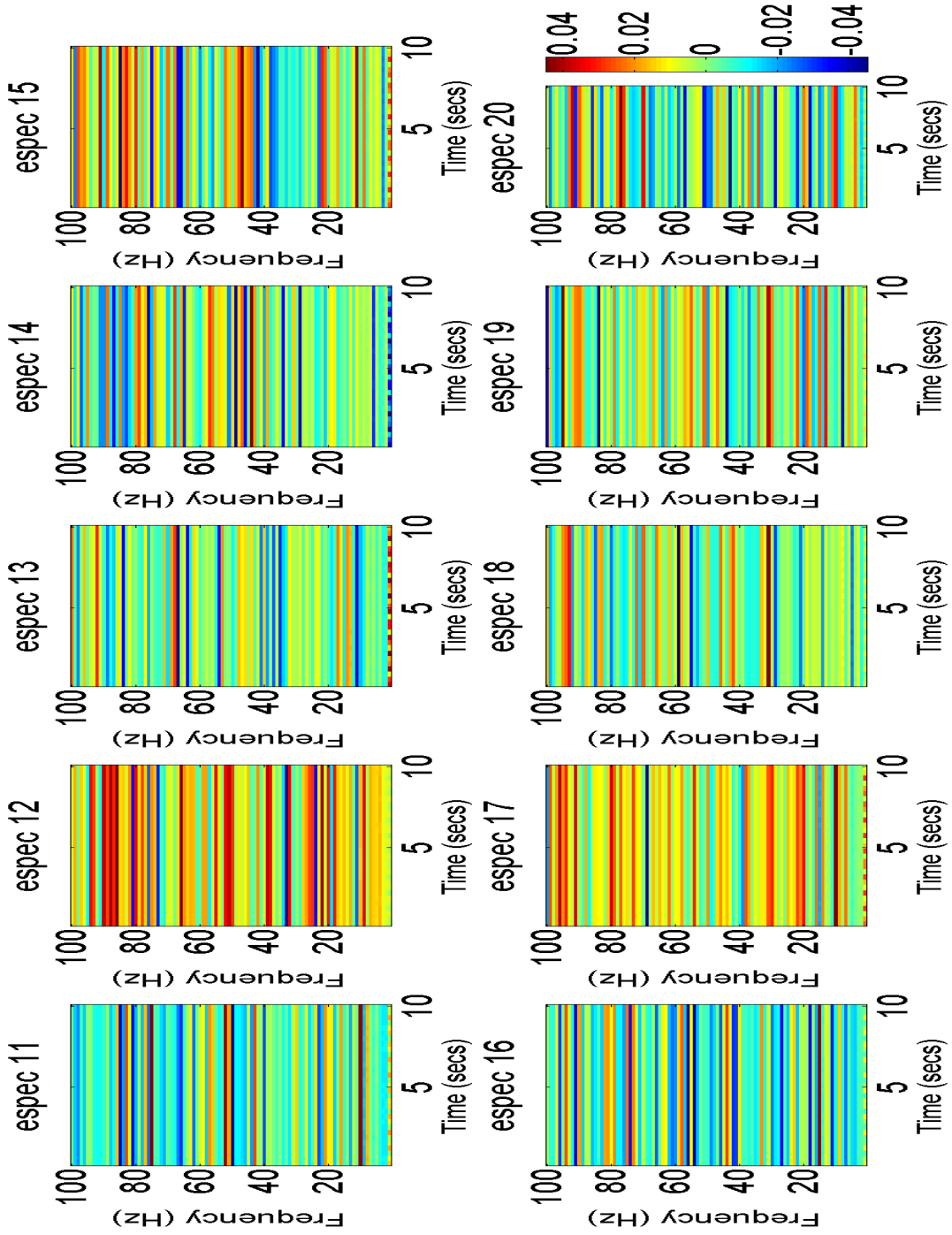


Figure C.2: Single-frequency sinusoid eigenspectrograms 11-20.

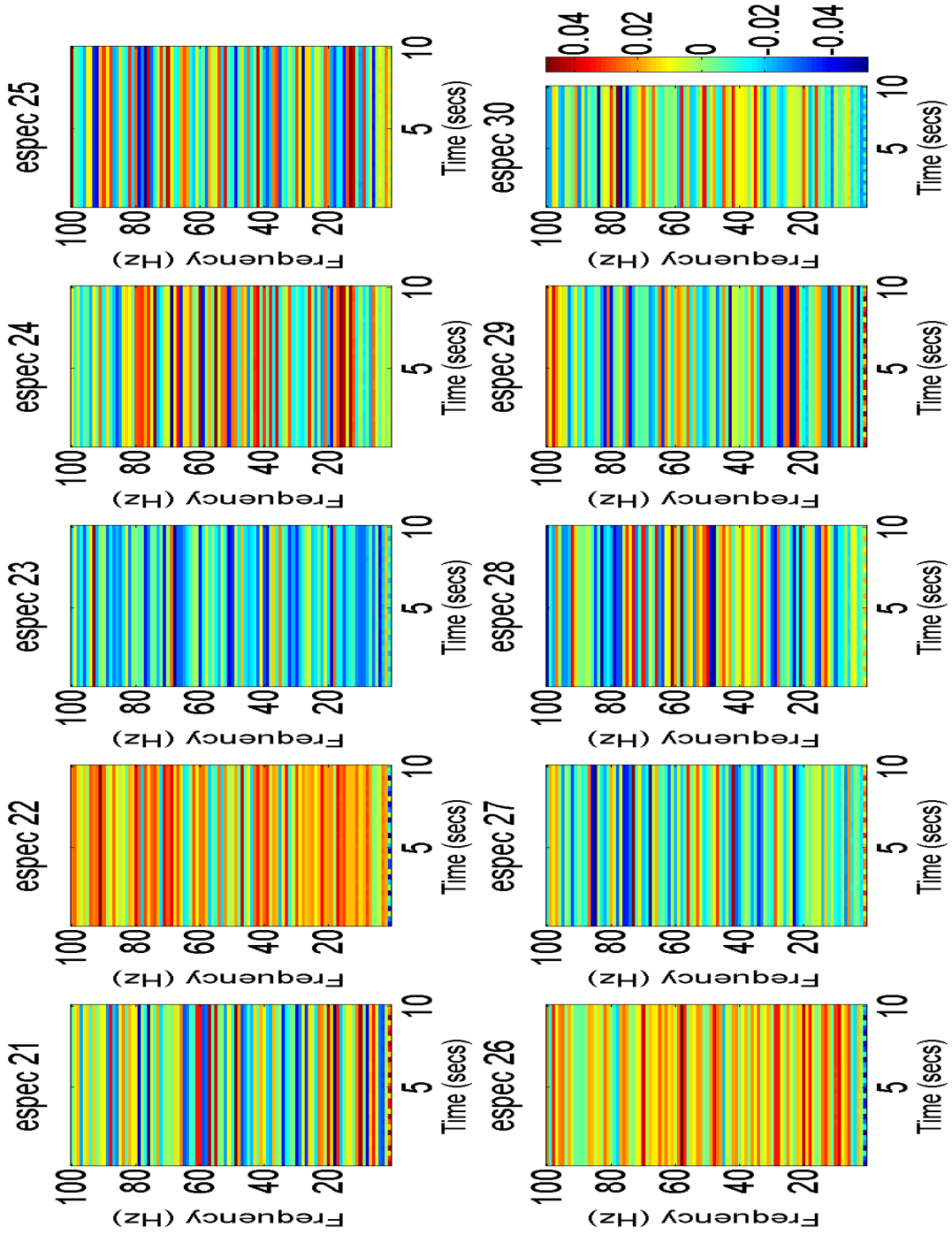


Figure C.3: Single-frequency sinusoid eigenspectrograms 21-30.

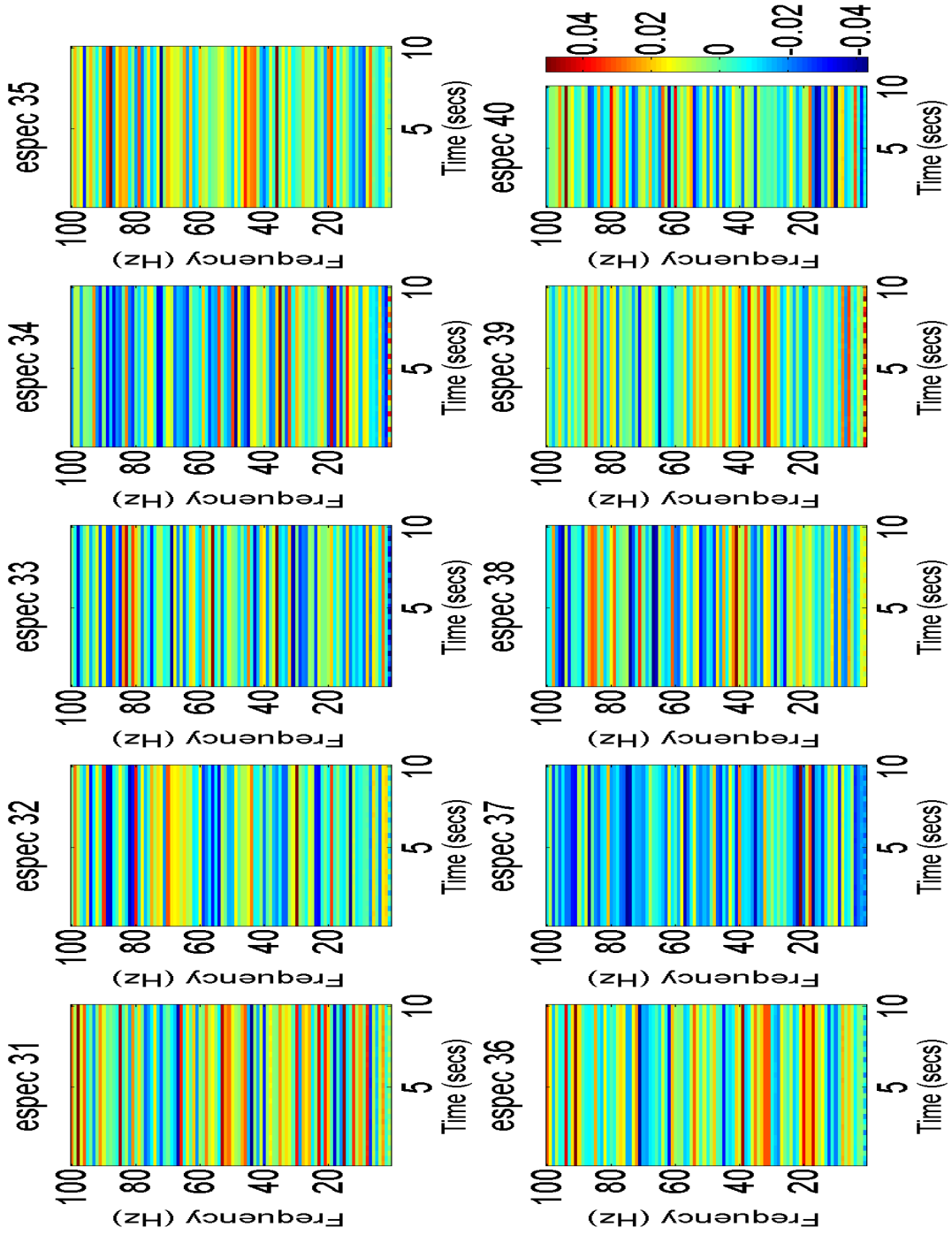


Figure C.4: Single-frequency sinusoid eigenspectrograms 31-40.

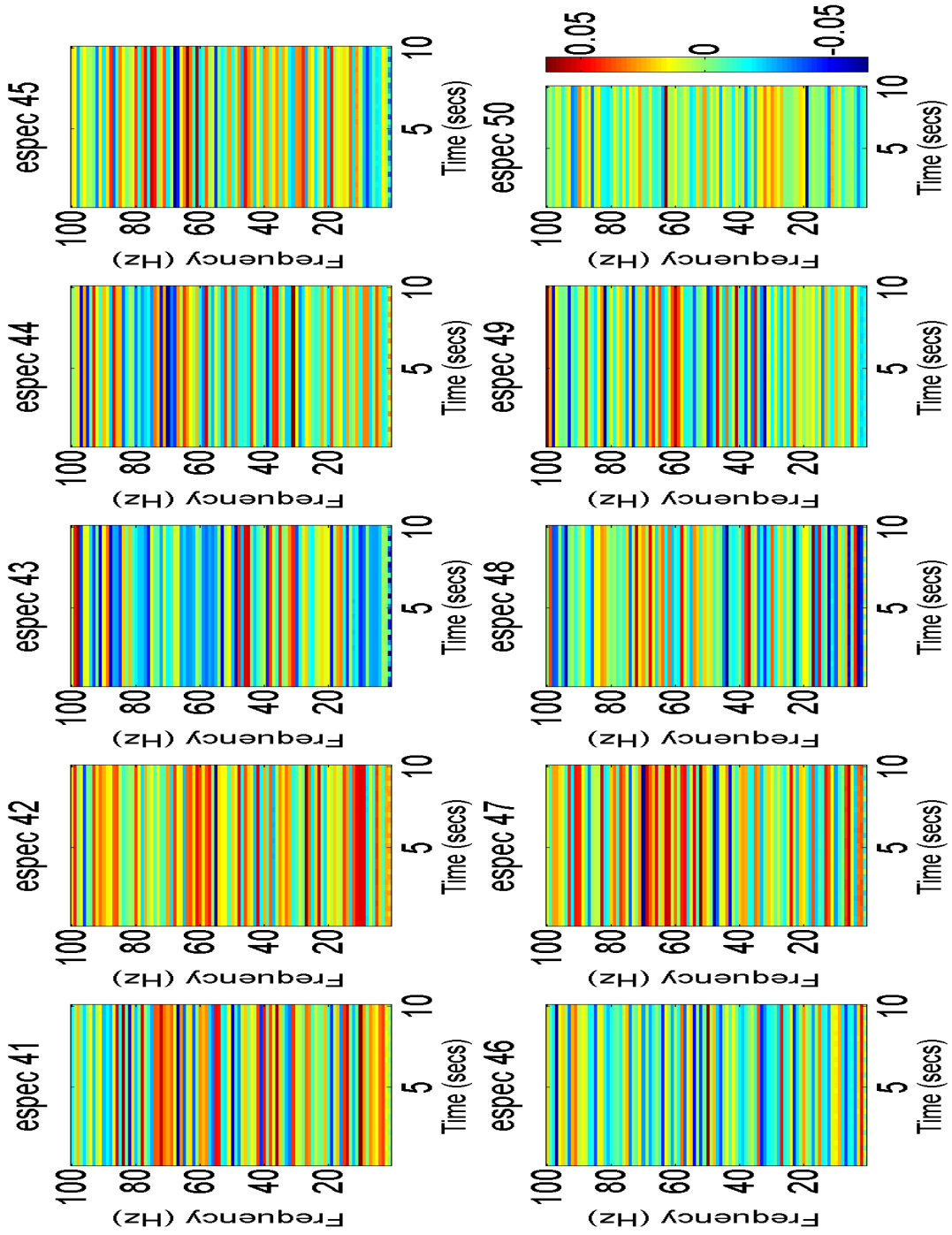


Figure C.5: Single-frequency sinusoid eigenspectrograms 41-50.

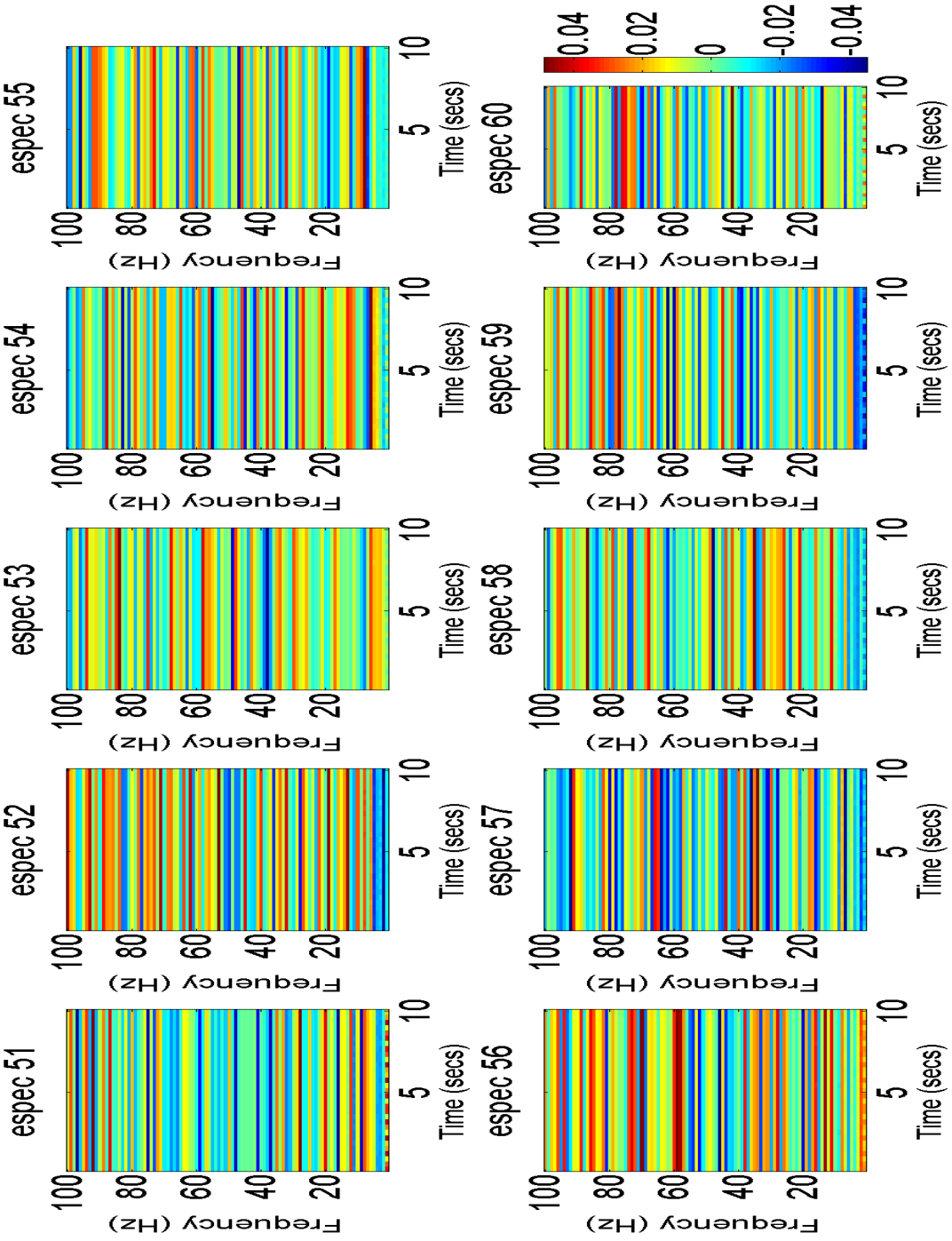


Figure C.6: Single-frequency sinusoid eigenspectrograms 51-60.

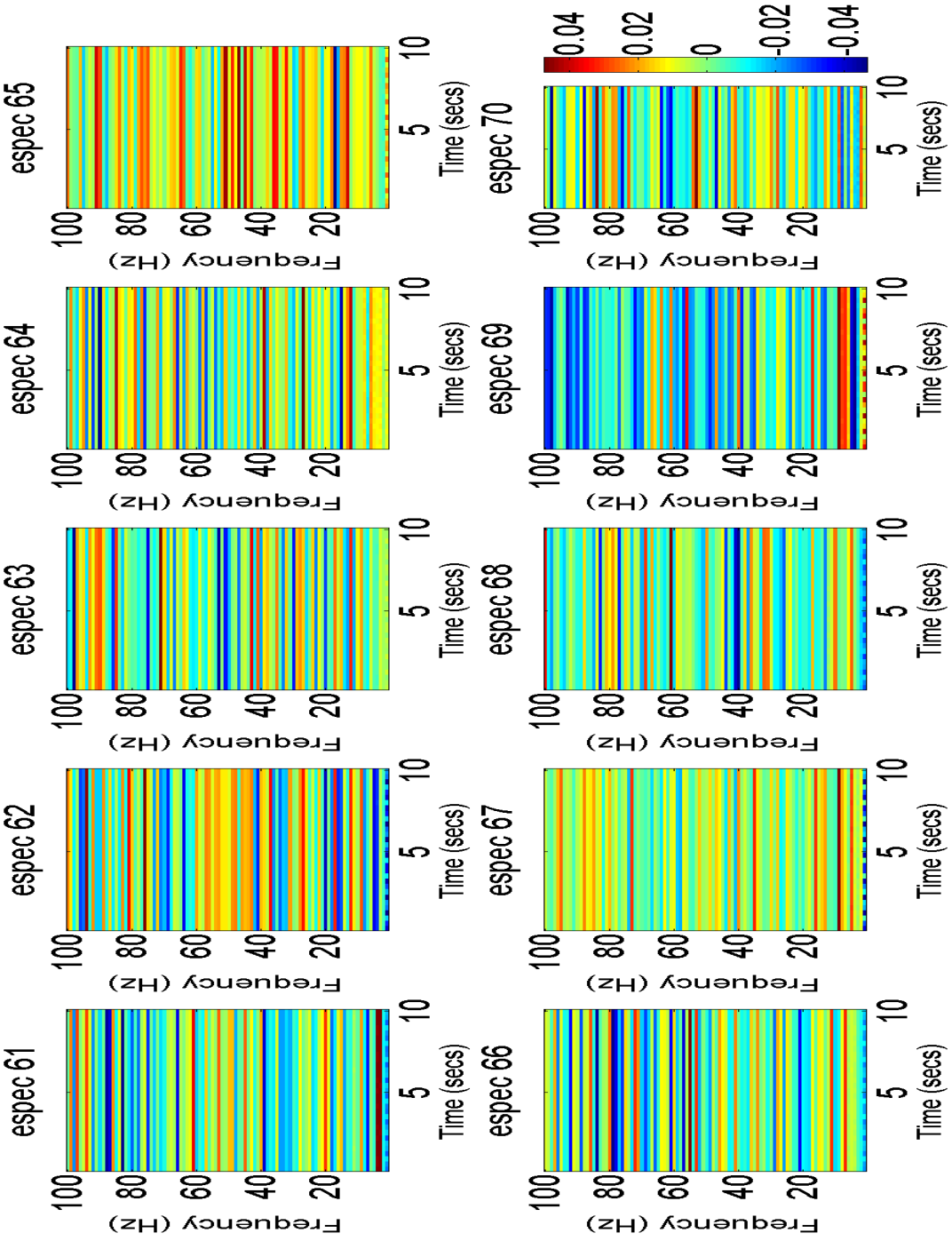


Figure C.7: Single-frequency sinusoid eigenspectrograms 61-70.

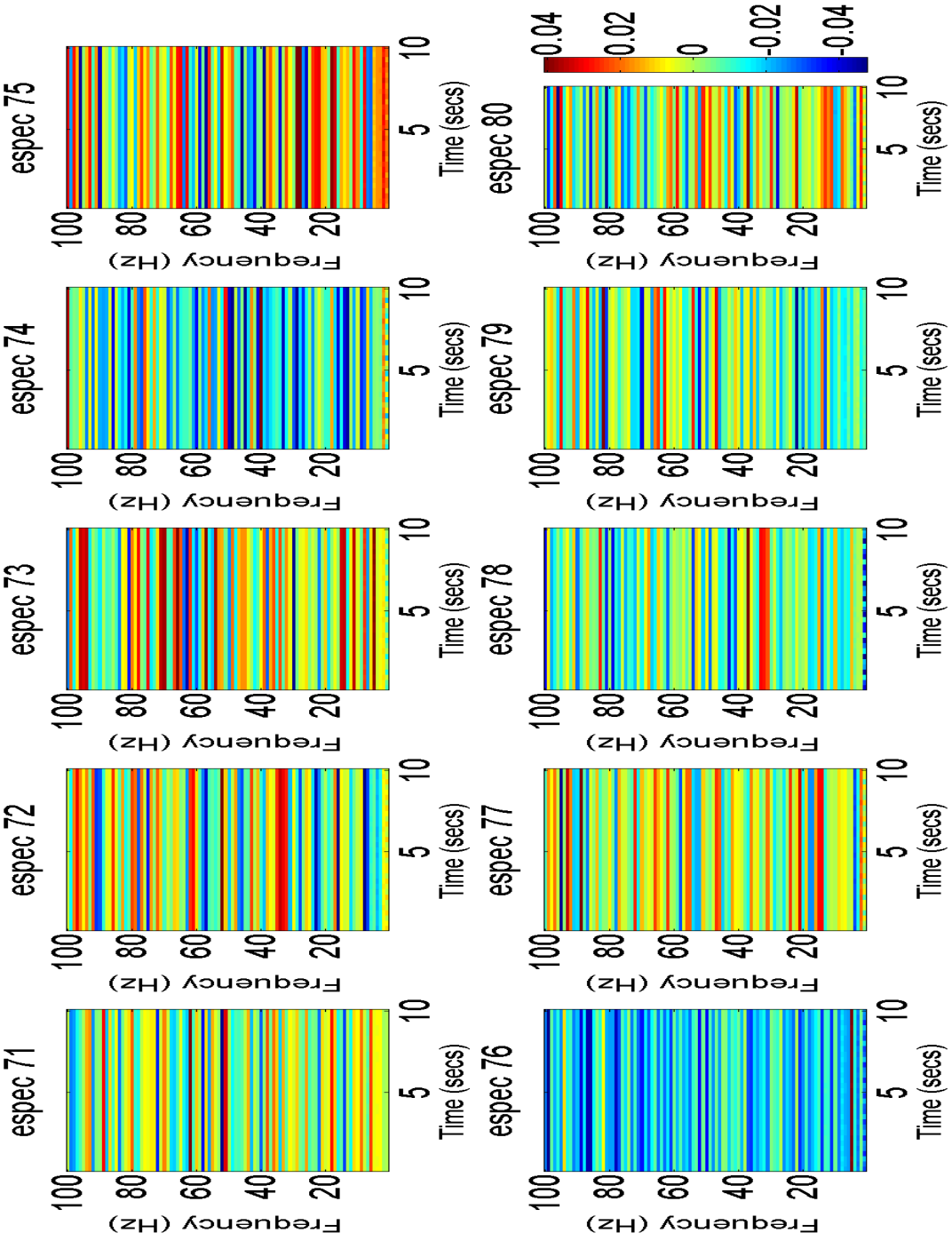


Figure C.8: Single-frequency sinusoid eigenspectrograms 71-80.

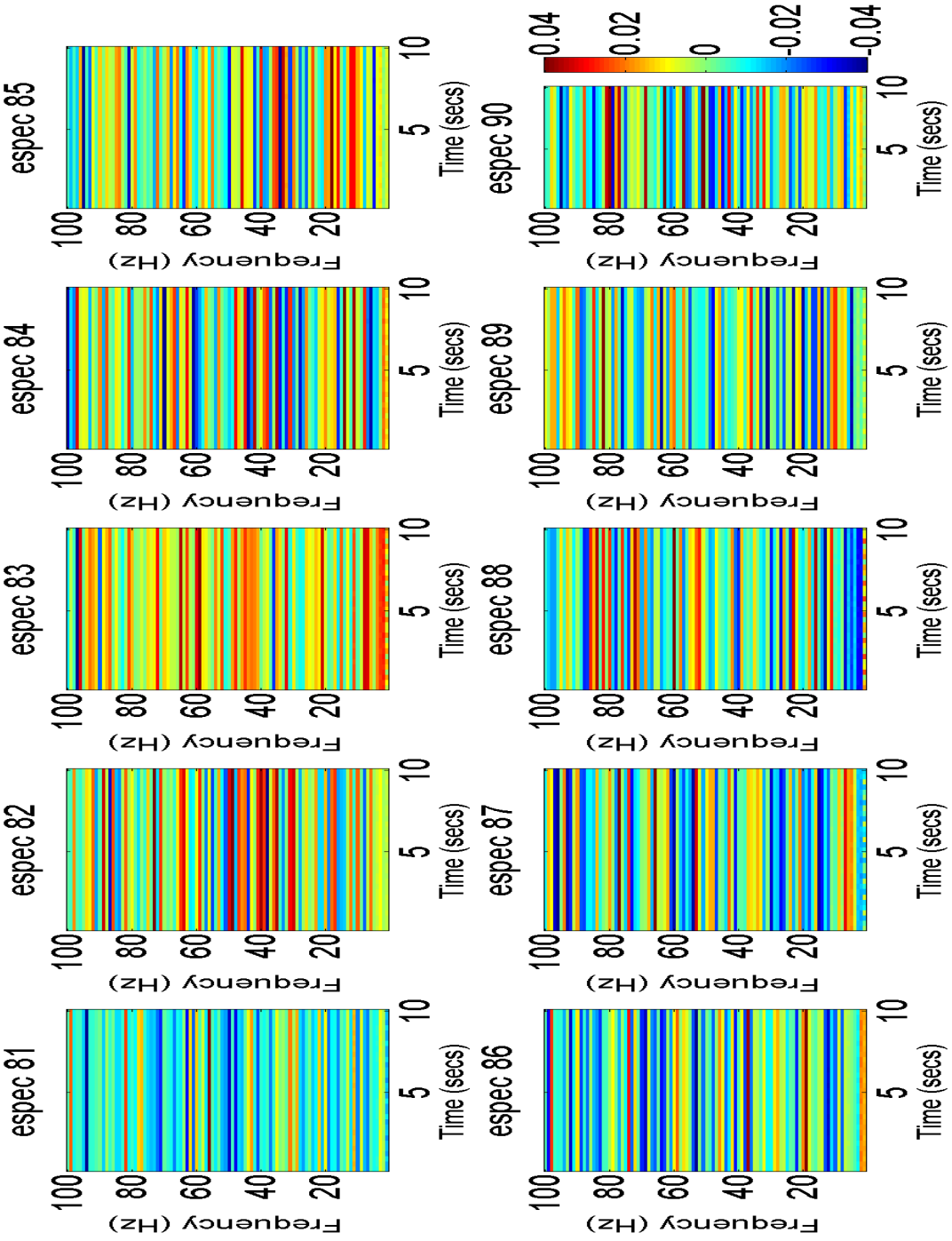


Figure C.9: Single-frequency sinusoid eigenspectrograms 81-90.

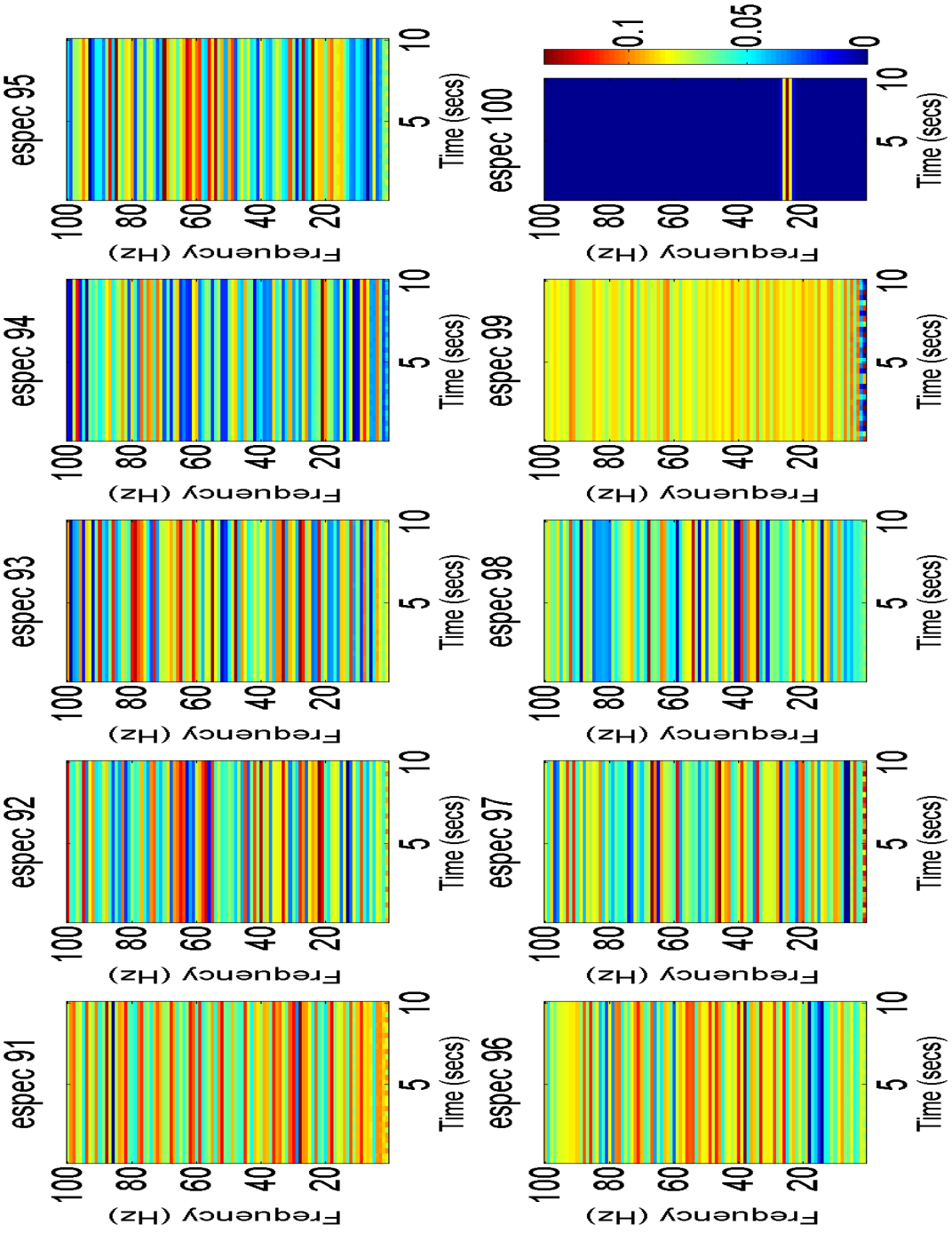


Figure C.10: Single-frequency sinusoid eigenspectrograms 91-100.

much meaningful information, and has been included for completeness' sake, although they do appear similar in many respects to the eigenspectrograms from the monochromatic spectrograms (see C.1). To aid clarity, 'eigenspectrogram' has been contracted to 'espec' and the power measurements of each group of eigenspectrograms are displayed by the final colour bar.

C.3 Spectrogram samples

A set of eight spectrograms (see Fig. C.21) created from approximately 2-week long analytic kludge waveforms in the manner described in section 4.3. The parameters at time t_0 are provided in Table C.1, and were drawn randomly from the same region of parameter space as those in Section 4.3.3 using the same parameter resolution. Eight randomly-selected eigenspectrograms from the set produced by the PCA method are shown in Fig. C.22.

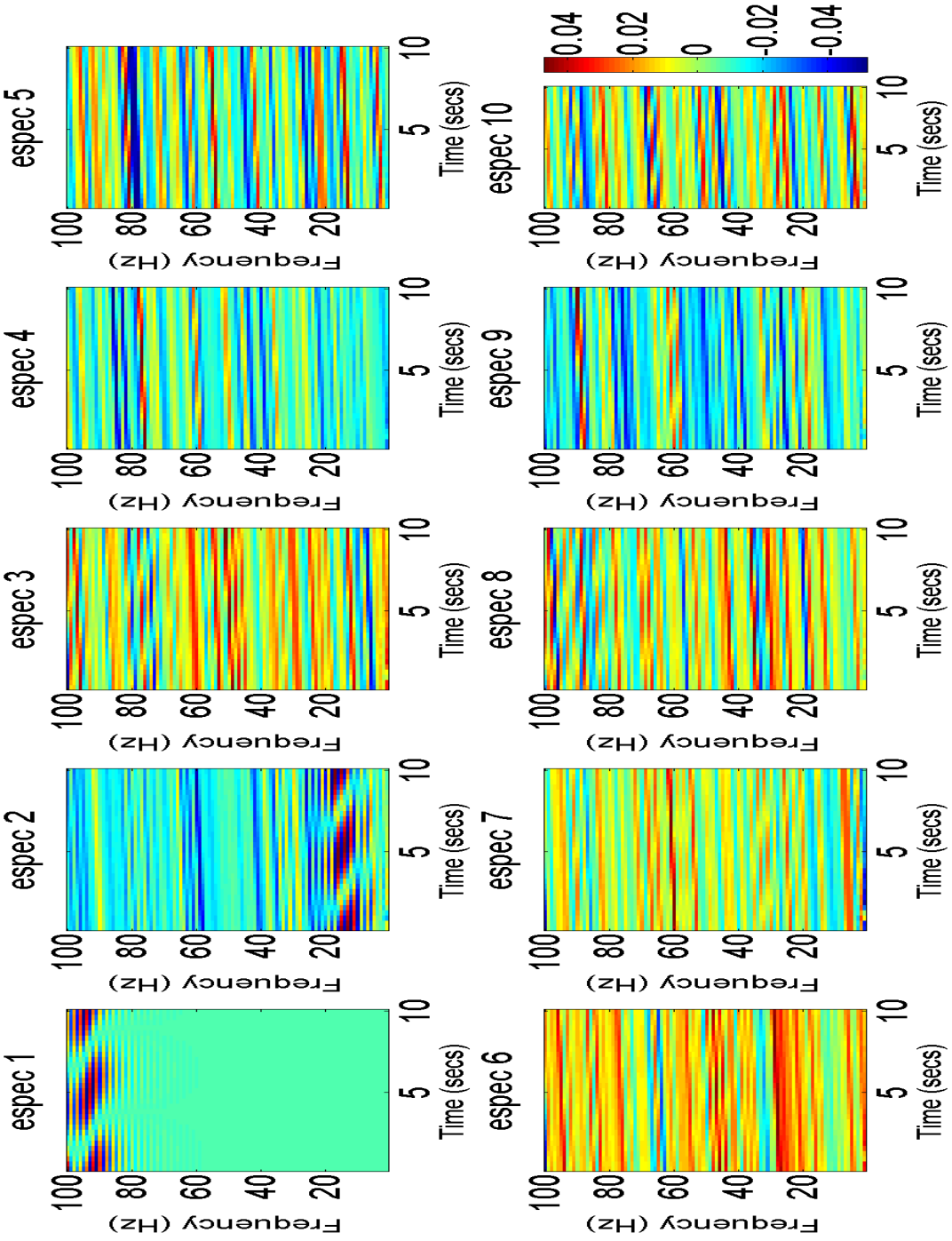


Figure C.11: The first ten eigenpectrograms generated by a set of quasi-monochromatic sinusoids (see 4.2.1.3).

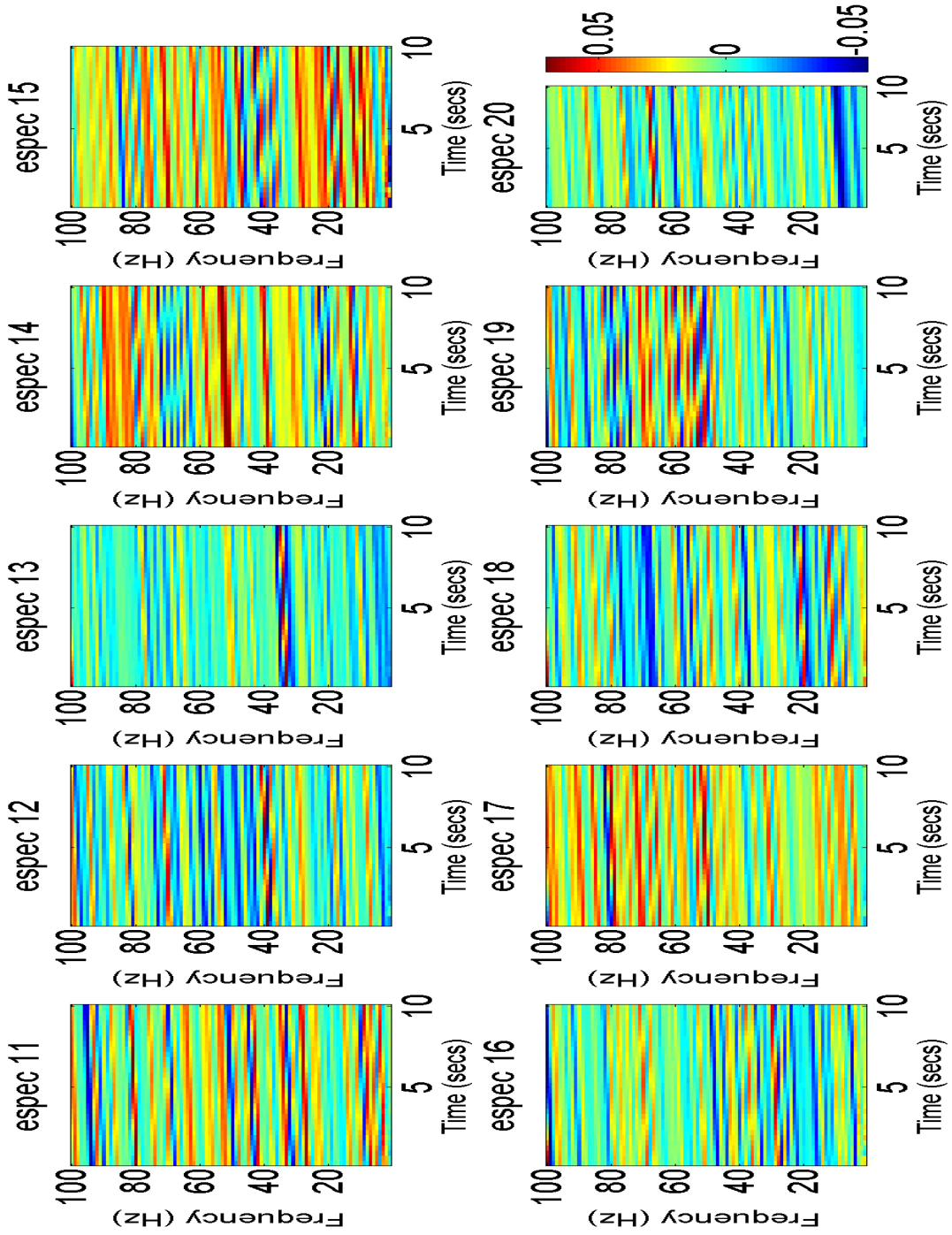


Figure C.12: Quasi-monochromatic sinusoid eigenspectrograms 11-20.

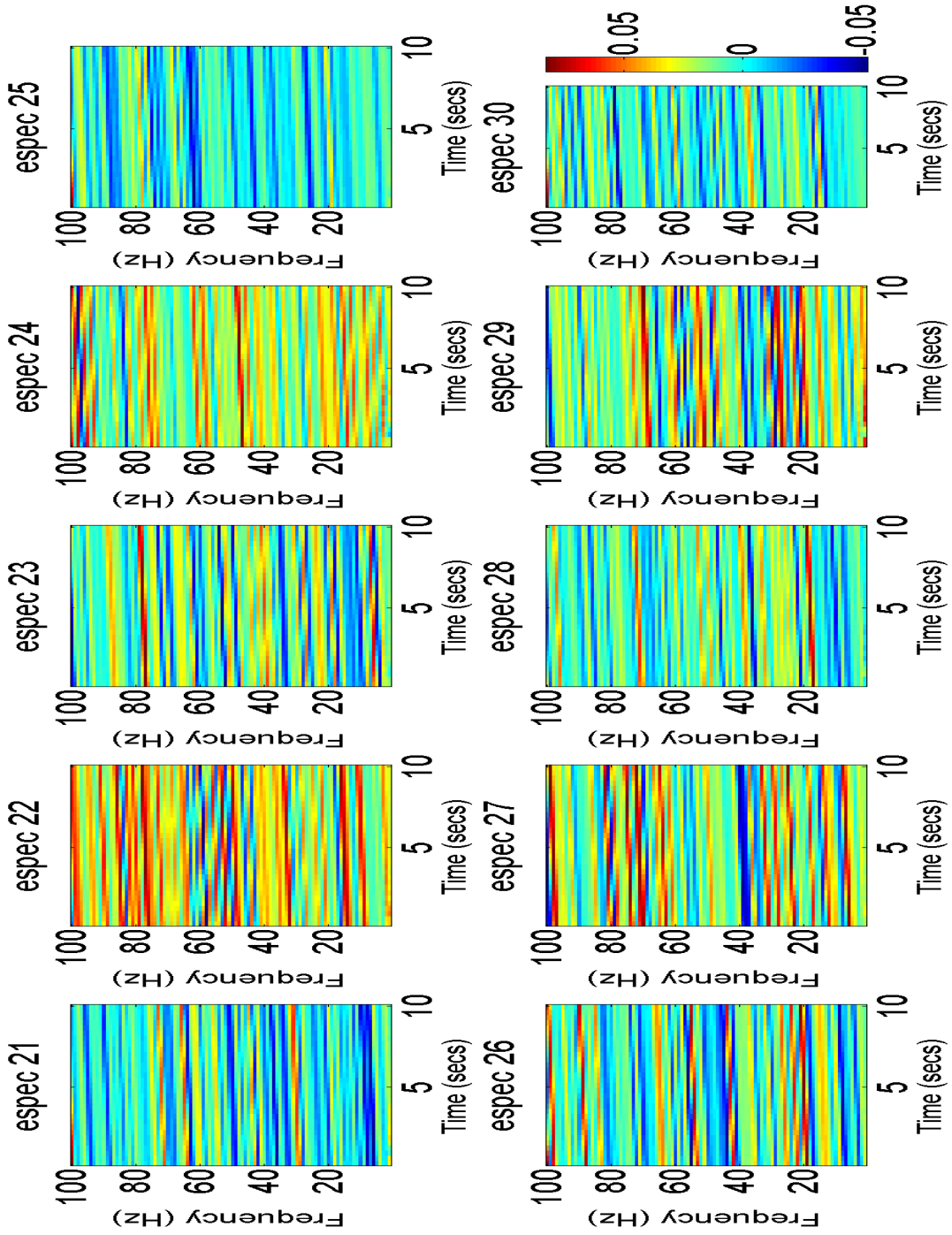


Figure C.13: Quasi-monochromatic sinusoid eigenspectrograms 21-30.

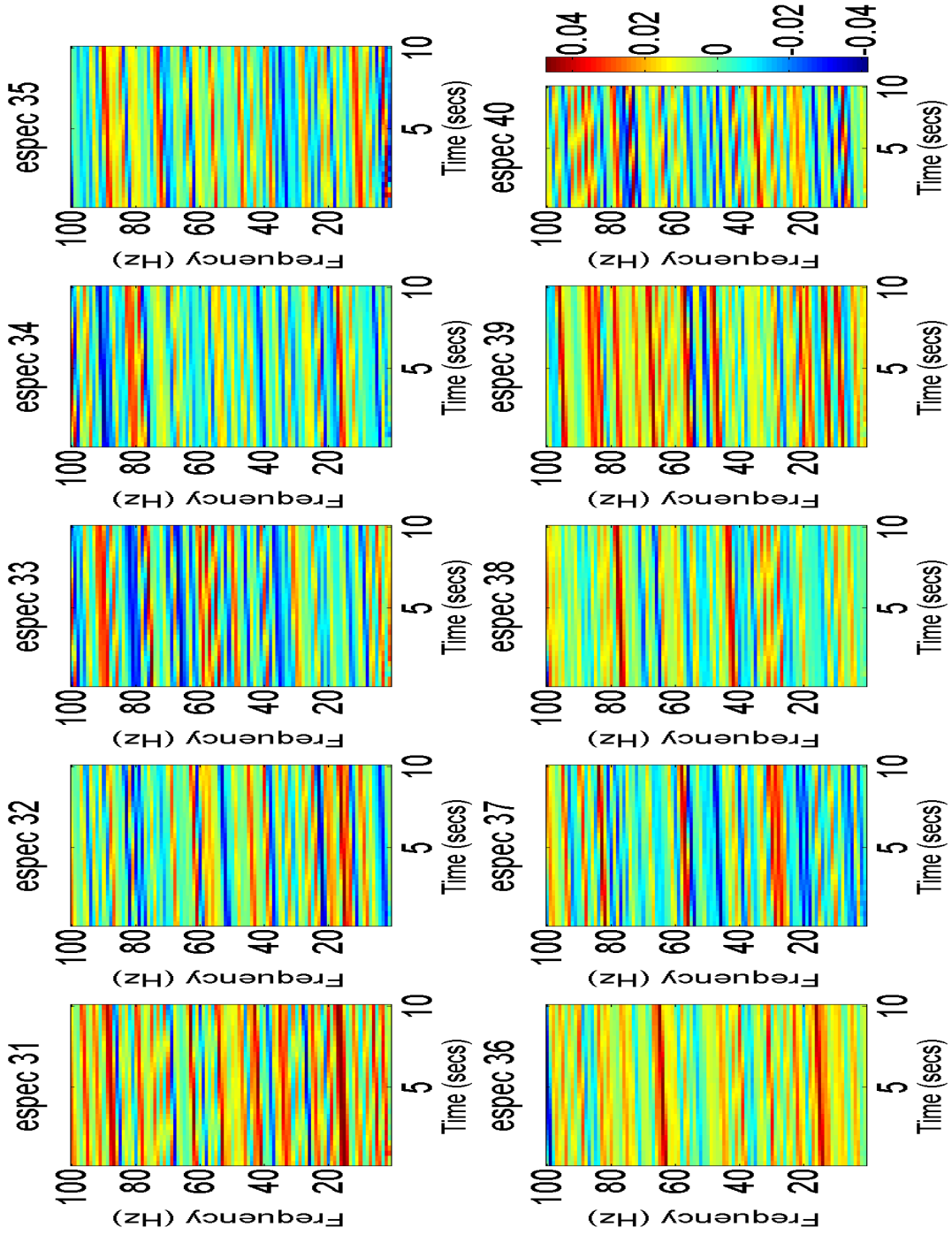


Figure C.14: Quasi-monochromatic sinusoid eigenspectrograms 31-40.

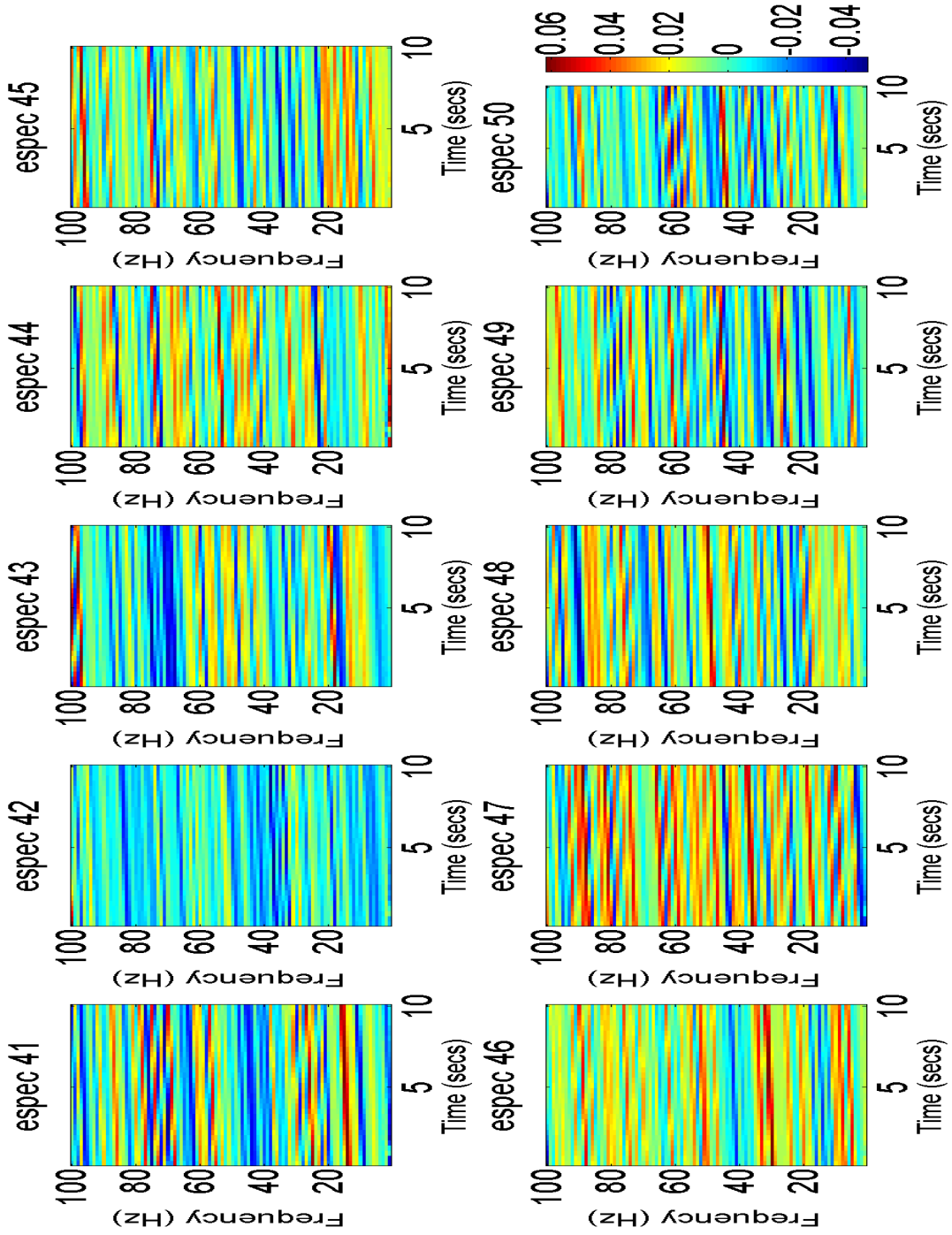


Figure C.15: Quasi-monochromatic sinusoid eigenspectrograms 41-50.

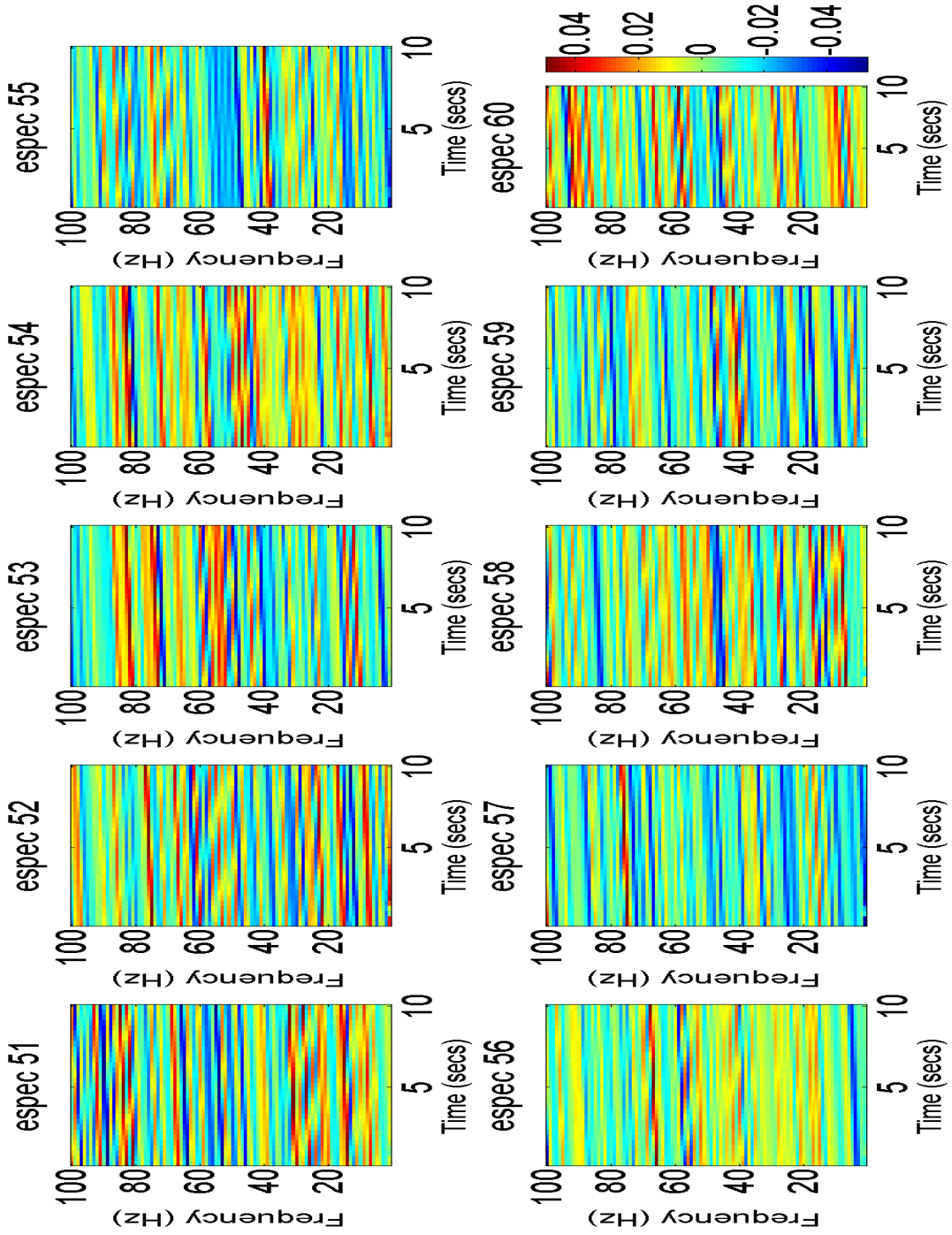


Figure C.16: Quasi-monochromatic sinusoid eigenspectrograms 51-60.

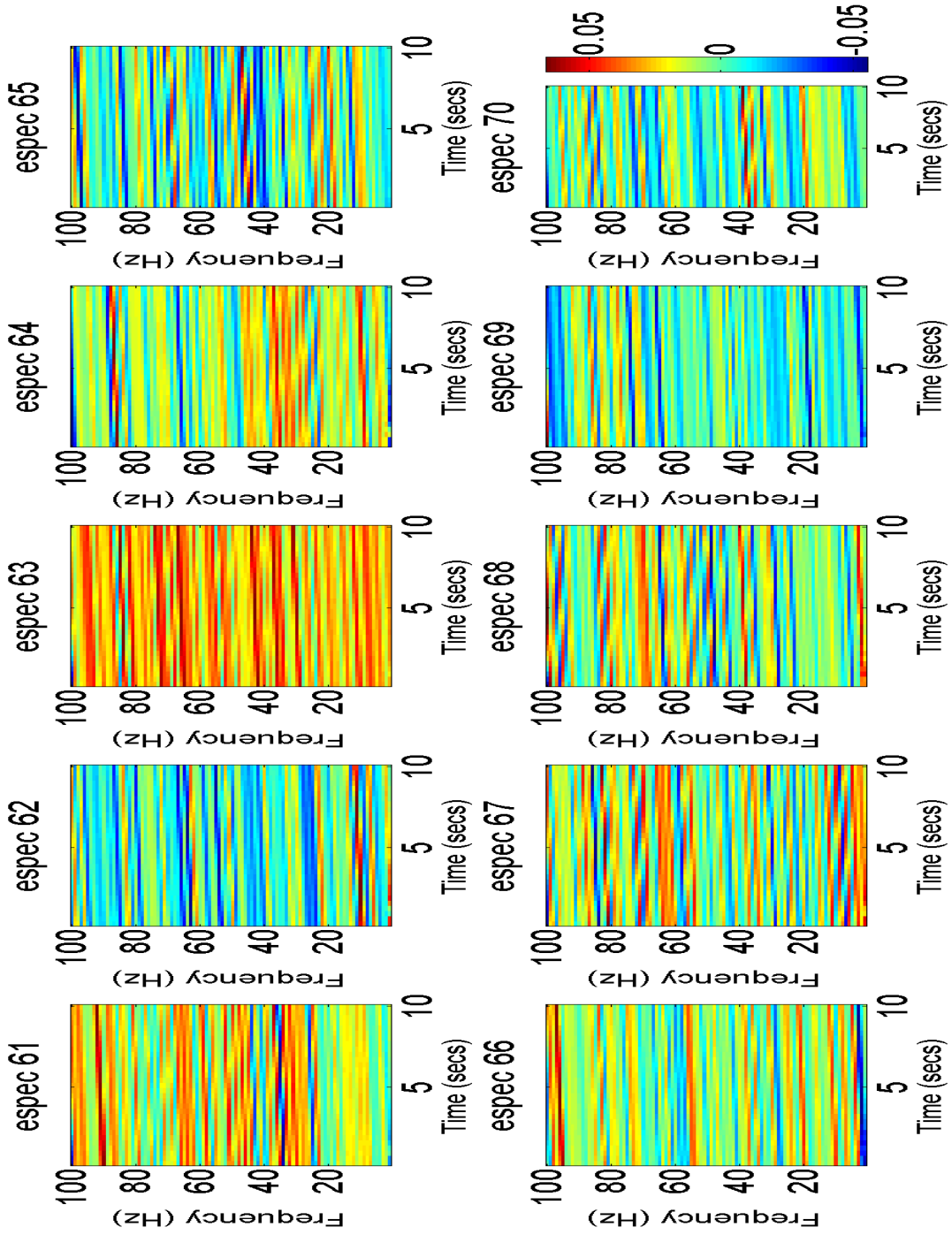


Figure C.17: Quasi-monochromatic sinusoid eigenspectrograms 61-70.

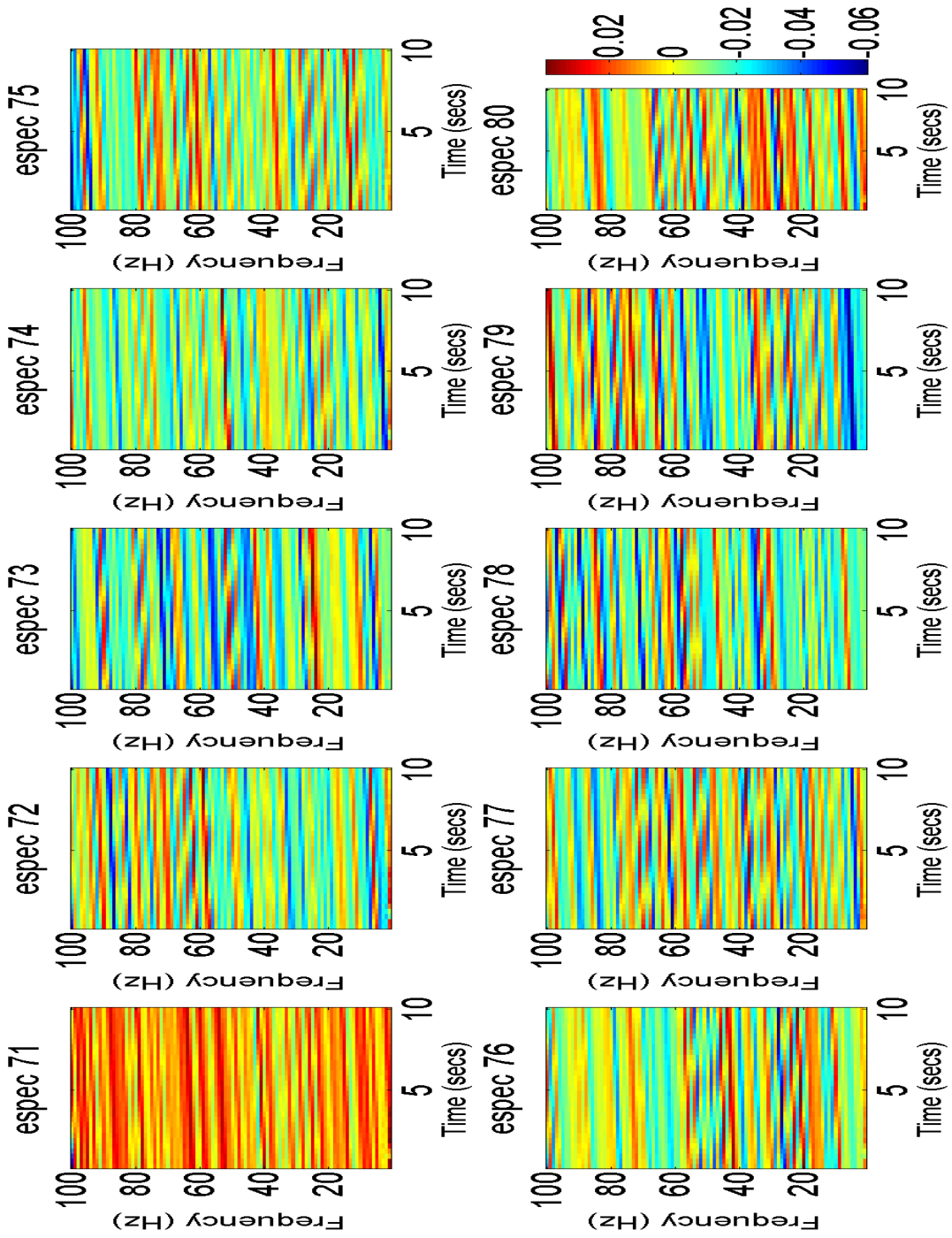


Figure C.18: Quasi-monochromatic sinusoid eigenspectrograms 71-80.

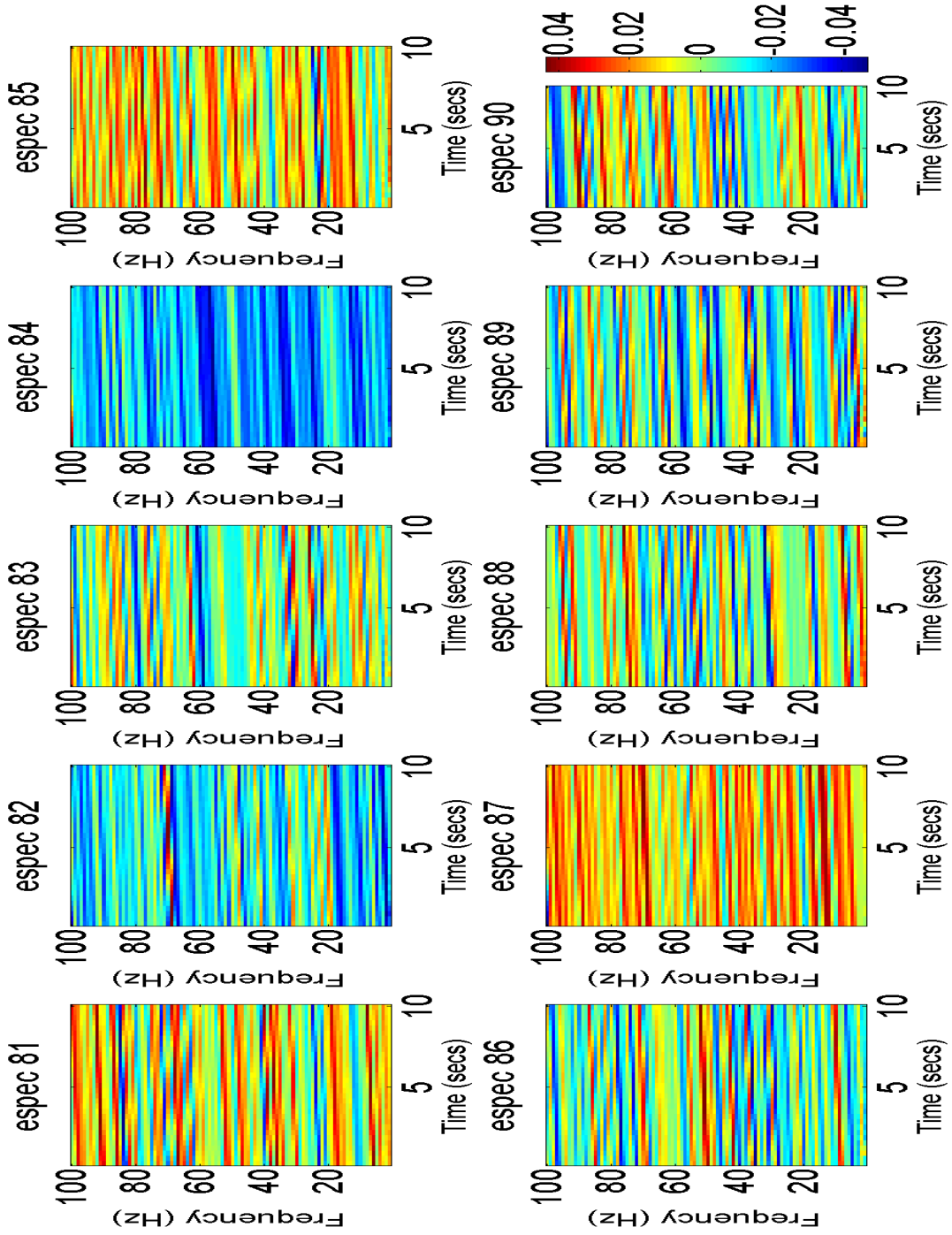


Figure C.19: Quasi-monochromatic sinusoid eigenspectrograms 81-90.

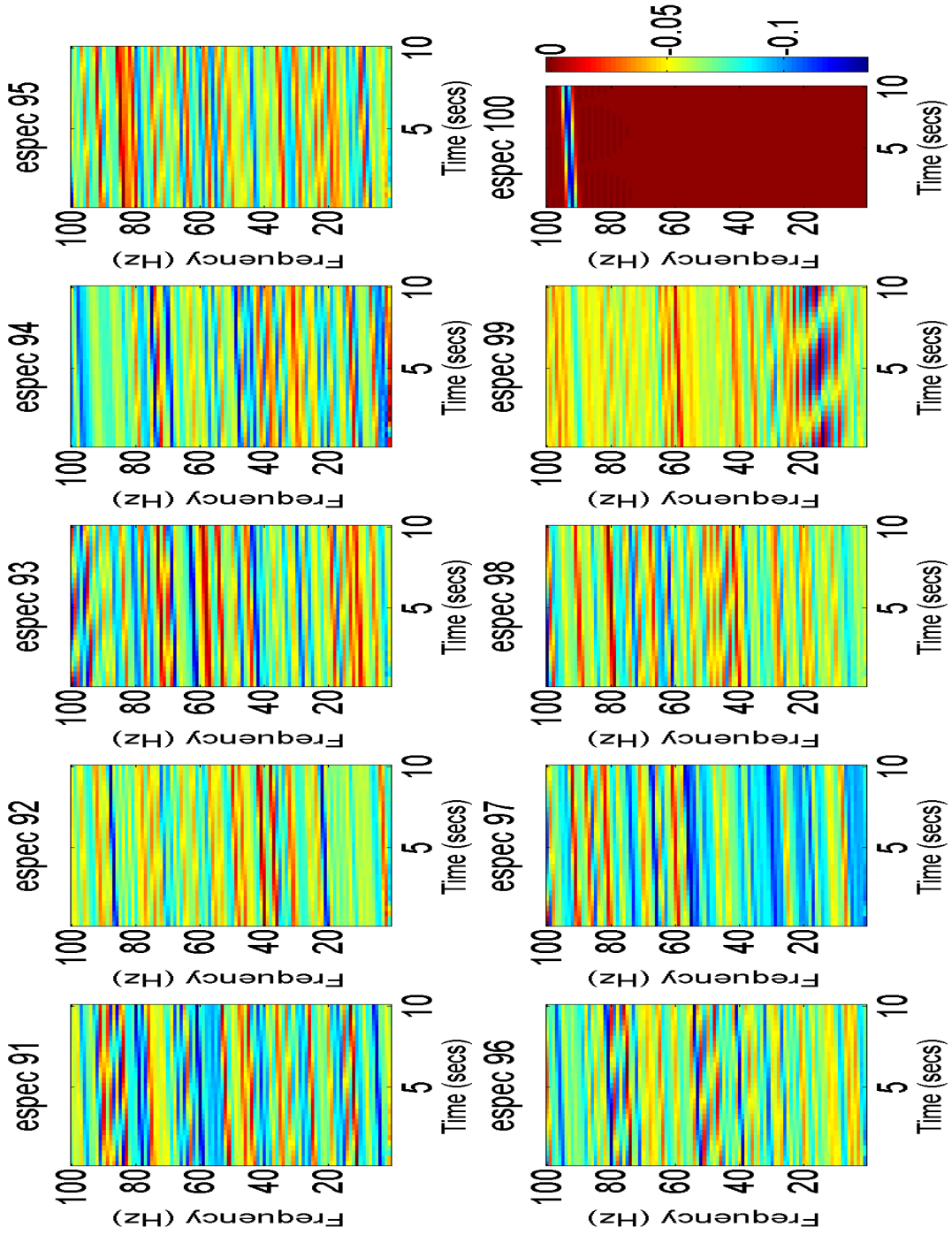


Figure C.20: Quasi-monochromatic sinusoid eigenspectrograms 91-100.

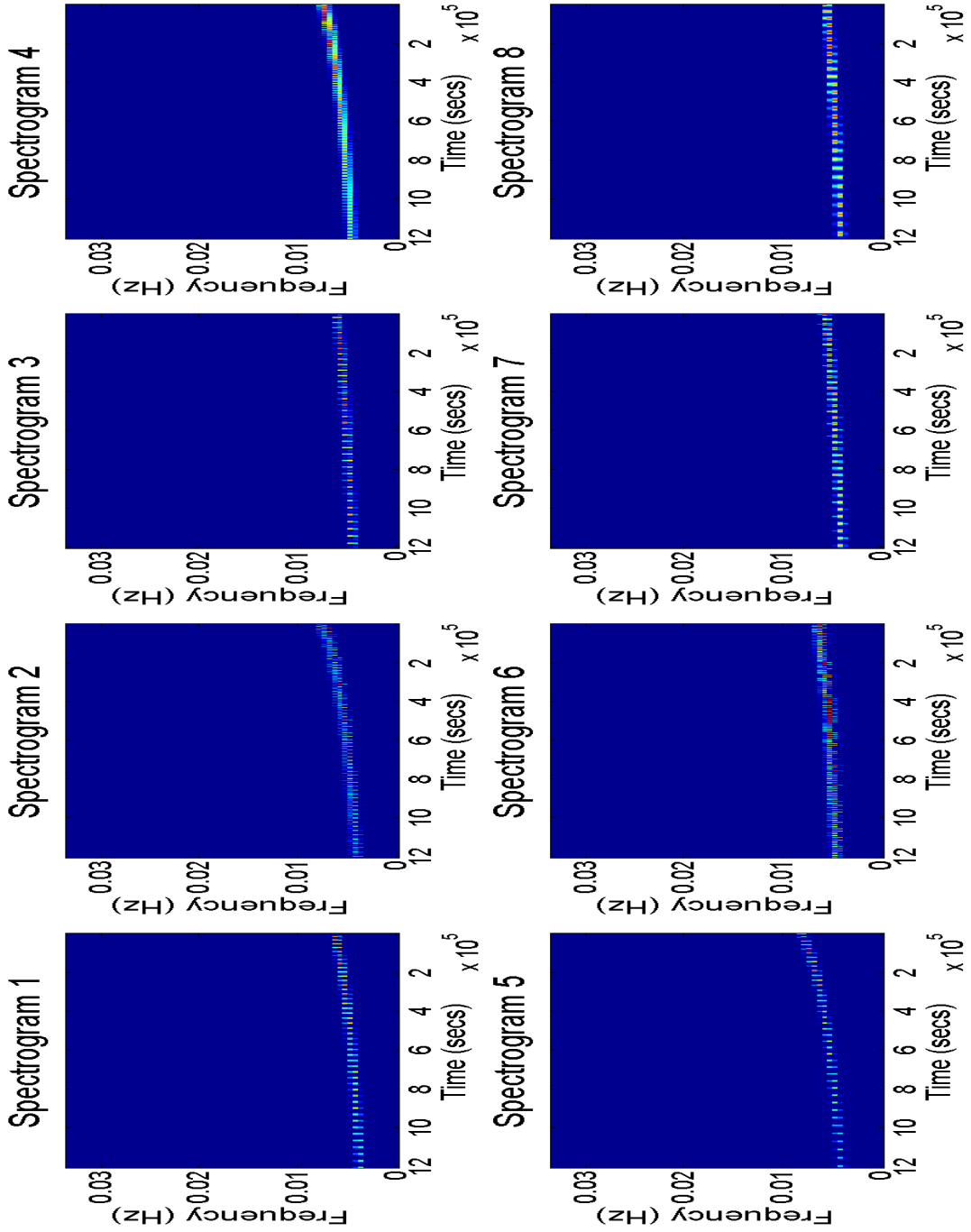


Figure C.21: Eight randomly-selected spectrograms taken from a large parameter space (see Table.C.1 for parameter values).

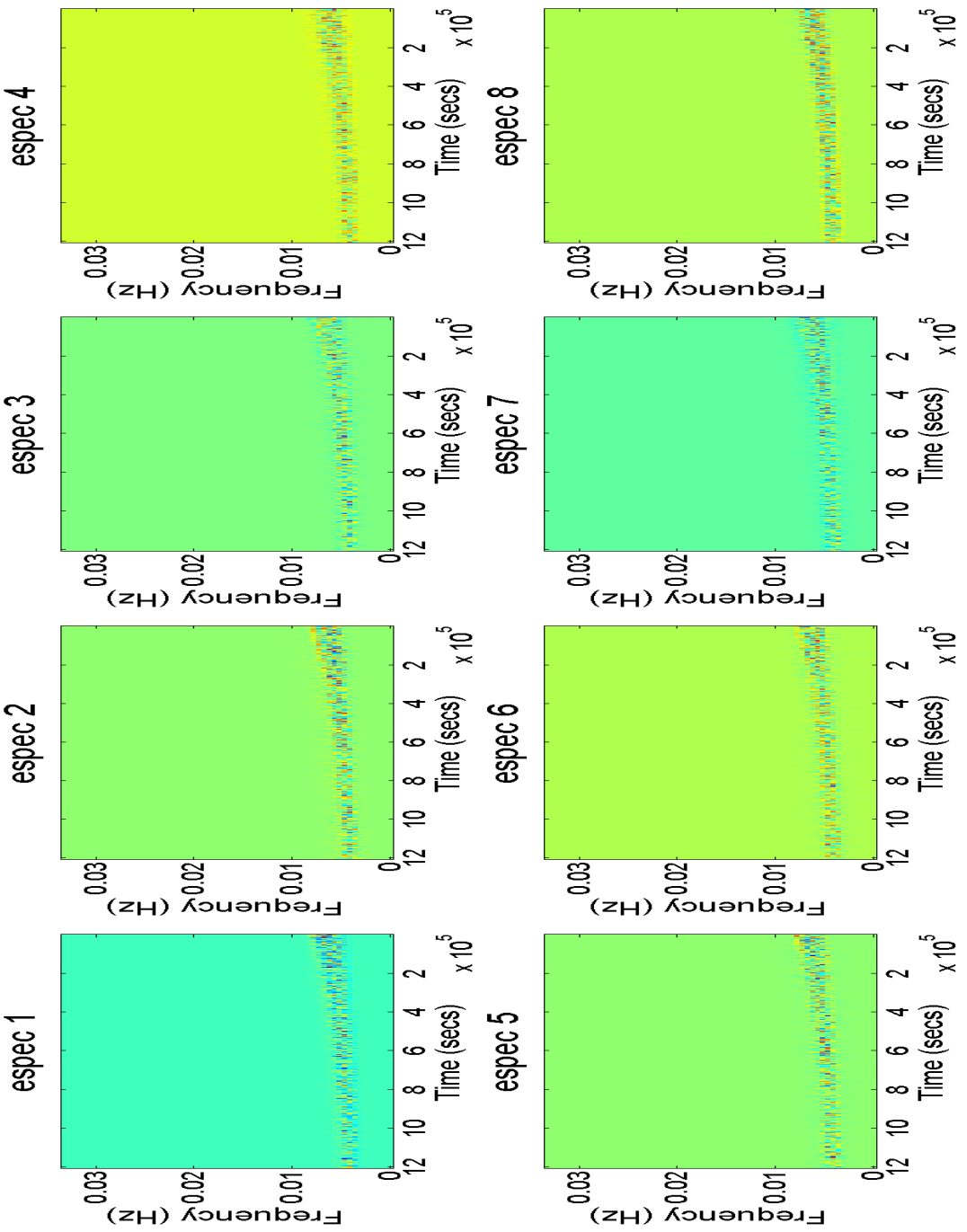


Figure C.22: Eight randomly-selected eigenspectrograms taken from the attempt to span the large parameter space described in Table.4.3.

Parameter	Spectrogram number							
value	1	2	3	4	5	6	7	8
t_0	-900	-900	-900	-900	-900	-900	-900	-900
μ	162	160	110	179	181	119	188	148
M	1486000	1278700	1298000	1014500	1255000	1371600	1430100	1087300
S/M^2	0.35	0.25	0.25	0.13	0.26	0.46	0.27	0.27
e_0	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
$\tilde{\gamma}_0$	3.7525	6.2308	4.0492	5.0265	0.3665	2.7402	6.0039	2.7053
Φ_0	3.4208	3.0369	1.1345	3.9619	4.5902	0.2967	6.2134	4.5204
θ_s	0.8203	4.311	1.7802	6.1785	5.6723	4.119	0.9948	4.8346
ϕ_s	3.4732	3.0194	0.8552	0.9948	2.9147	4.9218	0.0349	5.8294
λ	0.2623	1.1699	0.8907	0.9779	0.751	1.3445	0.402	1.1874
α	1.6406	1.3439	0.9076	1.9548	0.9774	4.9393	2.1817	2.0944
θ_k	3.3336	1.5708	5.3931	0.5411	6.0563	2.7227	1.4661	5.5327
ϕ_k	5.1138	0.6283	0.192	5.5152	5.5152	2.1991	4.7473	3.3161
D	1.501	1.856	1.233	1.304	1.408	1.82	1.449	1.169

Table C.1: The parameter values for the EMRIs. For simplicity, the source distance D is given in Gpc

References

[1] Bernard Schutz. *A first course in general relativity; 2nd ed.* Cambridge Univ. Press, Cambridge, 2009.

[2] J. H. Taylor and J. M. Weisberg. Further experimental tests of relativistic gravity using the binary pulsar psr 1913 + 16. *ApJ*, 345:434–450, October 1989.

[3] Bernard F. Schutz B.S. Sathyaprakash. Physics, astrophysics and cosmology with gravitational waves. *Living Reviews in Relativity*, 12(2), 2009.

[4] Kimberly C.B. New Chris L. Fryer. Gravitational waves from gravitational collapse. *Living Reviews in Relativity*, 6(2), 2003.

[5] J. M. Weisberg and J. H. Taylor. Relativistic binary pulsar b1913+16: Thirty years of observations and analysis. *arXiv.org:astro-ph/0407149*, 2004.

[6] The LIGO Scientific Collaborat Abbott. Beating the spin-down limit on gravitational wave emission from the crab pulsar. *arXiv.org:0805.4758*, 2008.

[7] J. Weber. Gravitational radiation. *Phys. Rev. Lett.*, 18(13):498–501, Mar 1967.

[8] J-P Zendri et al. Status report and near future prospects for the gravitational wave detector auriga. *Classical and Quantum Gravity*, 19(7):1925, 2002.

[9] P Astone et al. Explorer and nautilus gravitational wave detectors: a status report. *Classical and Quantum Gravity*, 25(11):114048, 2008.

[10] F Acernese et al. First joint gravitational wave search by the auriga/explorer/nautilus/virgo collaboration. *Classical and Quantum Gravity*, 25(20):205007, 2008.

[11] Peter R. Saulson. *Fundamentals of Interferometric Gravitational Wave Detectors*. World Scientific Pub Co Inc, Syracuse University, USA.

[12] Sheila Rowan Jim Hough. Gravitational wave detection by interferometry (ground and space). *Living Reviews in Relativity*, 3(3), 2000.

[13] H. Vahlbruch, M. Mehmet, S. Chelkowski, B. Hage, A. Franzen, N. Lastzka, S. Gossler, K. Danzmann, and R. Schnabel. Observation of squeezed light with 10-db quantum-noise reduction. *Phys. Rev. Lett.*, 100, 2008.

[14] I. Ciufolini and F. Fidecaro, editors. *Gravitational Waves: Sources and Detectors*, 1997.

[15] F. J. Raab and the Ligo Scientific collaboration. The status of laser interferometer gravitational-wave detectors. *Journal of Physics Conference Series*, 39:25–31, May 2006.

[16] F. Acernese et al. Status of Virgo detector. *Classical and Quantum Gravity*, 24:381–+, October 2007.

[17] Stephen Fairhurst, for the LIGO Scientific Collaboration, Hirotaka Takahashi, and for the TAMA Collaboration. Status of the joint ligo–tama300 inspiral analysis. *Classical and Quantum Gravity*, 22:S1109, 2005.

[18] D.G Blair et al. The science benefits and preliminary design of the southern hemisphere gravitational wave detector aigo. *Journal of Physics: Conference Series*, 122.

[19] Nicolas Arnaud, Matteo Barsuglia, Marie-Anne Bizouard, Philippe Canitrot, Fabien Cavalier, Michel Davier, Patrice Hello, and Thierry Pradier. Detection in coincidence of gravitational wave bursts with a network of interferometric detectors: Geometric acceptance and timing. *Phys. Rev. D*, 65(4):042004, Jan 2002.

[20] Massimo Tinto Sanjeev V. Dhurandhar. Time-delay interferometry. *Living Reviews in Relativity*, 8(4), 2005.

[21] Benjamin J. Owen. Search templates for gravitational waves from inspiraling binaries: Choice of template spacing. *Phys. Rev. D*, 53(12):6749–6761, Jun 1996.

[22] Andrzej Krlak Piotr Jaranowski. Gravitational-wave data analysis. formalism and sample applications: The gaussian case. *Living Reviews in Relativity*, 8(3), 2005.

[23] C. Messenger, R. Prix, and M. A. Papa. Random template banks and relaxed lattice coverings. *arXiv.org:0809.5223*, 2008.

[24] Phil Gregory. *Bayesian Logical Data Analysis for the Physical Sciences*. Cambridge University Press, New York, NY, USA, 2005.

[25] D.S. Sivia. *Data Analysis: A Bayesian Tutorial*. Oxford University Press.

[26] A. Eckart and R. Genzel. Observations of stellar proper motions near the galactic centre. *Nature*, 383:415–417, oct 1996.

[27] John Kormendy and Karl Gebhardt. Supermassive black holes in nuclei of galaxies. *arXiv.org:astro-ph/0105230*, 2001.

[28] K. A. Arnaud et al. Report on the first round of the mock lisa data challenges. *arXiv:gr-qc/0701139v3*, 2007.

[29] K. A. Arnaud et al. An overview of the second round of the mock lisa data challenges. *arXiv:gr-qc/0701170v4*.

[30] Stanislav Babak et al. The mock lisa data challenges: from challenge 1b to challenge 3. *CLASS.QUANT.GRAV*, 184026, 2008.

[31] S. Babak et al. Report on the second mock lisa data challenge. *CLASS.QUANT.GRAV*, 114037, 2008.

[32] (Ed.) Krishnan, B., (Ed.) Papa, M. A., and (Ed.) Schutz, Bernard F. Gravitational wave data analysis. Proceedings: 11th Workshop, GWDAW-11, Potsdam, Germany, Dec 18-21, 2006. Prepared for 11th Annual Gravitational Wave Data Analysis Workshop (GWDAW-11), Potsdam, Germany, 18-21 Dec 2006.

[33] Curt Cutler. Angular resolution of the lisa gravitational wave detector. *Physical Review D*, 57:7089, 1998.

[34] Curt Cutler and Éanna E. Flanagan. Gravitational waves from merging compact binaries: How accurately can one extract the binary's parameters from the inspiral waveform? *Phys. Rev. D*, 49(6):2658–2697, Mar 1994.

[35] G. Heinzel et al. Spectrum and spectral density estimation by the discrete fourier transform (dft), including a comprehensive list of window functions and some new flat-top windows, 2002.

[36] Kirsty J. Rhook and J. Stuart B. Wyithe. Realistic event rates for detection of supermassive black hole coalescence by lisa. *arXiv.org:astro-ph/0503210*, 2005.

[37] Kristen Menou, Zoltan Haiman, , and Vijay K. Narayanan. The merger history of supermassive black holes in galaxies. *The Astrophysical Journal*, 558(2):535–542, 2001.

[38] A. Sesana, F. Haardt, P. Madau, and M. Volonteri. The gravitational wave signal from massive black hole binaries and its contribution to the lisa data stream. *The Astrophysical Journal*, 623:23, 2005.

[39] S. Drasco. Strategies for observing extreme mass ratio inspirals. *Classical and Quantum Gravity*, 23:769–+, October 2006.

[40] Marc Freitag. Monte carlo cluster simulations to determine the rate of compact star inspiralling to a central galactic black hole. *Classical and Quantum Gravity*, 18(19):4033, 2001.

[41] J. R. Gair and L. Barack. Event rate estimates for LISA extreme mass ratio capture sources. *Classical and Quantum Gravity*, 21:1595–+, October 2004.

[42] Patrick R. Brady and Teviet Creighton. Searching for periodic sources with ligo. ii. hierarchical searches. *Phys. Rev. D*, 61(8):082001, Feb 2000.

[43] Yasushi Mino, Misao Sasaki, and Takahiro Tanaka. Gravitational radiation reaction to a particle motion. *Phys. Rev. D*, 55(6):3457–3476, Mar 1997.

[44] Theodore C. Quinn and Robert M. Wald. Axiomatic approach to electromagnetic and gravitational radiation reaction of particles in curved spacetime. *Phys. Rev. D*, 56(6):3381–3394, Sep 1997.

[45] S. Teukolsky. Perturbations of a rotating black hole. i. fundamental equations for gravitational, electromagnetic, and neutrino-field perturbations. *Astrophys. J.*, 185(635-648).

[46] S. Teukolsky. Perturbations of a rotating black hole. ii. dynamical stability of the kerr metric. *Astrophys. J.*, 185(649-674).

[47] L. Barack and C. Cutler. LISA capture sources: Approximate waveforms, signal-to-noise ratios, and parameter estimation accuracy. *Phys Rev D.*, 69(8), April 2004.

[48] P. C. Peters and J. Mathews. Gravitational radiation from point masses in a keplerian orbit. *Phys. Rev.*, 131(1):435–440, Jul 1963.

[49] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991.

[50] W.J Krzanowski and F.H.C. Marriott. *Multivariate Analysis, Part 1: Distributions, Ordination and Inference*. Hodder Arnold, 1994.

[51] I.T Jolliffe. *Principal Component Analysis: Second Edition*. Springer, 2002.

[52] Thomas A. Whitelaw. *Introduction to Linear Algebra, Second Edition*. Chapman & Hall , CRC, 1991.

[53] Gerald Goertzel. An algorithm for the evaluation of finite trigonometric series. *The American Mathematical Monthly*, 65(1):34–35.

[54] John Veitch and Alberto Vecchio. Assigning confidence to inspiral gravitational wave candidates with bayesian model selection. *Classical and Quantum Gravity*, 25:184010, 2008.

[55] Miquel Trias, Alberto Vecchio, and John Veitch. Markov chain monte carlo searches for galactic binaries in mock lisa data challenge 1b data sets. *arXiv:0804.4029v2*.

[56] Tyson B. Littenberg and Neil J. Cornish. A bayesian approach to the detection problem in gravitational wave astronomy. *Physical Review D*, 80:063007, 2009.

[57] Jeff Crowder and Neil J. Cornish. Extracting galactic binary signals from the first round of mock lisa data challenges. *arXiv:0704.2917v1*, 2007.

[58] Miquel Trias, Alberto Vecchio, and John Veitch. Delayed rejection schemes for efficient markov-chain monte-carlo sampling of multimodal distributions. *arXiv:0904.2207*, 2009.

[59] Farhan Feroz, Jonathan R. Gair, Michael P. Hobson, and Edward K. Porter. Use of the multinest algorithm for gravitational wave data analysis. *Classical and Quantum Gravity*, 26:215003, 2009.

[60] Neil J. Cornish and Louis J. Rubbo. Lisa response function. *Phys. Rev. D*, 67(2):022001, Jan 2003.

[61] Michele Vallisneri. Synthetic lisa: Simulating time delay interferometry in a model lisa. *Physical Review D*, 71:022001, 2005.

[62] Shane L. Larson, William A. Hiscock, and Ronald W. Hellings. Sensitivity curves for spaceborne gravitational wave interferometers. *Physical Review D*, 62:062001, 2000.

[63] Massimo Tinto, F. B. Estabrook, and J. W. Armstrong. Time-delay interferometry for lisa. *Phys. Rev. D*, 65(8):082003, Apr 2002.

Webliography

[W..a] 11th Gravitational Wave Data Analysis Workshop
http://gwdaw11.aei.mpg.de/.

[W..b] AURIGA Bar Detector
http://www.auriga.lnl.infn.it/.

[W..c] Australian International Gravitational Research Centre
http://www.gravity.uwa.edu.au/.

[W..d] Australian International Gravitational Research Centre
http://www.gravity.uwa.edu.au/.

[W..e] Document for Challenge 1 Draft v1.0
svn.sourceforge.net/viewvc/lisatools/Docs/challenge1.pdf.

[W..f] GEO600 home page
http://geo600.aei.mpg.de.

[W..g] GW Notes
http://brownbag.lisascience.org/lisa-gw-notes/.

[W..h] LAPACK Users' Guid Third Edition
http://www.netlib.org/lapack/lug/lapack_lug.html.

[W..i] LCGT project overview page
http://www.icrr.u-tokyo.ac.jp/gr/LCGT.html.

[W..j] LIGO home page
http://www.ligo.caltech.edu.

[W..k] LIGO sensitivity curve data
http://www.ligo.caltech.edu/~jzweizig/distribution/LSC_Data/.

[W..l] LISA MLDC Taskfore (listwg1b)
http://www.tapir.caltech.edu/dokuwiki/listwg1a:home.

[W..m] Matlab Spectrogram function documentation
http://www.mathworks.com/access/helpdesk/help/toolbox/signal/index.html?/access/helpdesk/help/toolbox/signal/spectrogram.html.

[W..n] MLDC Round 4 dataset graphical representation
http://www.tapir.caltech.edu/dokuwiki/_media/listwg1b:mldc4.pdf.

[W..o] Mock LISA Data Challenge
http://astrogravs.nasa.gov/docs/mldc.

[W..p] Rome Gravitational Wave Group
http://www.roma1.infn.it/rog/.

[W..q] Space Based Interferometer Sensitivity Curve Generator
http://www.srl.caltech.edu/~shane/sensitivity/MakeCurve.html.

[W..r] Synthetic LISA
http://www.vallis.org/syntheticlisa/.

[W..s] TAMA project office home page
http://tamago.mtk.nao.ac.jp/spacetime/index.html.

[W..t] The LISA Simulator Page
http://www.physics.montana.edu/lisa/.

[W..u] The QR algorithm
http://www.acm.caltech.edu/~mlatini/research/qr_alg-feb04.pdf.

[W..v] VIRGO home page
http://wwwcascina.virgo.infn.it/.

[W..w] Virgo sensitivity curve data
http://www.virgo.infn.it/senscurve/.

[X..a] Bayesian evidence as a diagnostic for testing models of gravitational wave data
http://www.astro.gla.ac.uk/users/vsimha/.

[X..b] LIGO Press and Media: Chronology page
http://www.ligo.caltech.edu/LIGO_web/PR/scripts/chrono.html.

[X..c] LISA Pathfinder mission overview
http://www.esa.int/esaSC/120397_index_0_m.html.

[X..d] NASA: LISA mission page
http://lisa.nasa.gov/.