A galactic binary detection pipeline

Tyson B. Littenberg

Maryland Center for Fundamental Physics, Department of Physics,
University of Maryland, College Park, MD 20742 and
Gravitational Astrophysics Laboratory, NASA Goddard Spaceflight Center, 8800 Greenbelt Rd., Greenbelt, MD 20771
(Dated: June 24, 2011)

The Galaxy is suspected to contain hundreds of millions of binary white dwarf systems, a large fraction of which will have sufficiently small orbital period to emit gravitational radiation in band for space-based gravitational wave detectors such as the Laser Interferometer Space Antenna (LISA). LISA's main science goal is the detection of cosmological events (supermassive black hole mergers, etc.) however the gravitational signal from the galaxy will be the dominant contribution to the data – including instrumental noise – over approximately two decades in frequency. The catalogue of detectable binary systems will serve as an unparalleled means of studying the Galaxy. Furthermore, to maximize the scientific return from the mission, the data must be "cleansed" of the galactic foreground. We will present an algorithm that can accurately resolve and subtract $\gtrsim 10000$ of these sources from simulated data supplied by the Mock LISA Data Challenge Task Force. Using the time evolution of the gravitational wave frequency, we will reconstruct the position of the recovered binaries and show how LISA will sample the entire compact binary population in the Galaxy.

I. INTRODUCTION

Just as traditional photon astronomy has profited from accessing the entire electromagnetic spectrum, so to would gravitational wave (GW) astronomy benefit from detectors covering many decades in frequency. With existing GW observatories, high-frequency gravitational waves are the target of ground based interferometers, and the LIGO/Virgo collaboration [1, 2] have developed astonishingly sensitive observatories which are poised to bring GW astronomy out of its infancy.

While small wavelength gravitational waves are, observationally, the most readily accessible, the richest signal space in the GW universe is at frequencies far too low for Earth-based instruments. A space-borne interferometer is the only foreseeable way of reaping the bounty of information transmitted at longer wavelengths, and allowing gravitational wave astronomy to reach its full potential.

Unique to the mHz frequency range is the existence of known sources, comprised of close binary star systems in the Galaxy, mostly the AM CVn-type binary stars [3]. These individual objects, discovered electromagnetically, are the near-by representatives of a much larger population of low frequency gravitational wave sources on our cosmic doorstep. Population synthesis models for the galaxy predict some 60 *million* binary star systems emitting gravitational radiation at frequencies between 0.1 and 10 mHz [4–6]. Because gravitational waves interact very weakly with intervening matter, a staggeringly large number of these objects, distributed throughout the entire Galaxy, are within reach.

Maximizing what can be gleaned from mHz GW data will require solving unique analysis problems in gravitational wave astronomy. In particular, the challenges posed by these galactic binaries are the sheer number of sources, and the degree with which they overlap one another in signal space, smearing together to form a confused blend of gravitational wave power which, depending on details of the detector, can be orders of magnitude larger than the instrumental noise floor. The total number of binaries which are individually resolvable out of this population is unknown and poses a very large dimension model selection problem. The harm in over-fitting the data, (i.e, tolerating large false alarm probabilities) or under-fitting the data (i.e., accepting large false dismissal rates) not only affects the science that can be done with the catalogue of resolved signals, but can also impact the data analysis efforts for more distant sources of gravitational radiation which share the same bandwidth.

To prepare for this challenge, we have set out to build a "detection pipeline" which can automatically solve every facet of this problem. To wit, we need to locate candidate sources in the data, select for the most parsimonious number, accurately estimate the physical parameters that describe the system, and cleanly regress the sources from the data. While the analysis software we have built is flexible with regards to the details of the instrument, we use the NASA/ESA LISA mission [7] with the intention of participating in round four of the Mock LISA Data Challenges (MLDC) [8]. MLDC datasets are released in pairs: one coming with the list of signal parameters within the data, and one without (henceforth, the "training data" and "blind data", respectively). For this study, we will focus on analyzing the training data in which all other types of sources have been removed, leaving behind only the galactic binaries and the instrument noise. The capabilities of the algorithm on this reduced dataset will serve as a realistic demonstration of what we could achieve on the blind data, as the only cosmological sources which can impact the number of resolvable binaries are the brightest of the binary black hole mergers and cosmic strings, neither of which pose a serious challenge to existing search algorithms.

The challenge presented by the galactic foreground has been addressed with iteratively more sophistication in past MLDCs [9]. Challenge 2 was the first to simu-

late a complete galaxy and in response to this, Crowder and Cornish developed the `BAM` algorithm [10] which, to date, has been the most successful attempt at solving the entire problem. The `BAM` codes themselves have been lost to the sands of time, and did not model the evolution of the binary's orbital period (nor did Challenge 2). Other existing algorithms which have worn their teeth on the MLDC data sets include an F-statistic maximization scheme using a Nelder-Mead simplex algorithm [11], a hierarchical cleaning algorithm (also employing the F-statistic) [12] and an MCMC search algorithm featuring a Markovian delayed rejection proposal distribution [13]. Beyond MLDC entries there have been numerous proof of principal studies including (but not limited to), Refs. [14–16].

Our goal for this paper is to build a new analysis pipeline to handle the signal from the population of galactic binaries. We are, in effect, attempting to extend the `BAM` algorithm by including frequency evolution in our model of the waveforms, and technical advancements in GW data analysis made since 2006.

The paper is organized as follows: §IIAnalysis §IIIModel §IVAlgorithm §VCatalogue §VIDiscussion

II. DATA ANALYSIS BASICS

There are two desired products of the data analysis procedure. For each model of the data ($\mathcal{M}$) under consideration, we need to find the "best fit" model parameters and have some sense of how well these parameters are constrained, as well as determine which model is most strongly supported by the data.

The fundamental piece of math at work here is Bayes' theorem which, when cast as an inference problem, takes the form

$$p(\vec{\theta}|s, \mathcal{M}) = \frac{p(s|\vec{\theta}, \mathcal{M})p(\vec{\theta}|\mathcal{M})}{p(s|\mathcal{M})} \qquad (1)$$

The left hand side of equation 1 is the posterior distribution function for parameters $\vec{\theta}$ given the data s, from which parameter estimation conclusions can be drawn within model $\mathcal{M}$ (see, for instance, [17] for a thorough introduction). The numerator of the right hand side contains the product of the likelihood – our "goodness of fit" measure – and the prior, encoding our knowledge before the new data were collected. The denominator is the evidence, or marginalized likelihood for $\mathcal{M}$. Comparisons of $p(s|\mathcal{M})$ between different models reveal which is most strongly supported by the data and, assuming uninformative priors on the models themselves, should be taken as the preferred representation.

In the Bayesian framework, one needs only to define the likelihood and prior distributions, and the rest of the analysis is reduced to an oft time-consuming calculation. We prefer the Markov Chain Monte Carlo family of algorithms to perform the computation, although

Nested Sampling, and its offspring `MultiNest`, have been used to similar affect (Ref.XXX,XXX). There is no shortage of data analysis literature describing the concept of MCMCs. However, for the sake of introducing vocabulary and notation, we briefly sketch the algorithm.

We begin with some random position in parameter space $\vec{\theta}_x$ as the first "link" in the chain by evaluating its likelihood $p(s|\vec{\theta}_x, \mathcal{M})$ and prior probability $p(\vec{\theta}_x|\mathcal{M})$. Next, we suggest a trial position, $\vec{\theta}_y$, from a proposal distribution $q(\vec{\theta}_y|\vec{\theta}_x)$. The new likelihood and prior probability are evaluated and $\vec{\theta}_y$ is adopted as the next sample in the chain with probability $\alpha_{\vec{\theta}_x \leftrightarrow \vec{\theta}_y} = \min[1, H_{\vec{\theta}_x \leftrightarrow \vec{\theta}_y}]$ where $H_{\vec{\theta}_x \leftrightarrow \vec{\theta}_y}$ is the Hastings ratio

$$H_{\vec{\theta}_x \rightarrow \vec{\theta}_y} = \frac{p(s|\vec{\theta}_y, \mathcal{M})p(\vec{\theta}_y|\mathcal{M})q(\vec{\theta}_x|\vec{\theta}_y)}{p(s|\vec{\theta}_x, \mathcal{M})p(\vec{\theta}_x|\mathcal{M})q(\vec{\theta}_y|\vec{\theta}_x)}. \qquad (2)$$

This process of stochastically stepping through parameter space repeats until some convergence criteria are satisfied. Equation 2 is derived from the detailed balance condition, which requires that the probability of being at state $\vec{\theta}_x$ and transitioning to state $\vec{\theta}_y$ is the same as being at $\vec{\theta}_y$ and moving to $\vec{\theta}_x$. If satisfying this condition when adopting new solutions in the chain, the number of iterations spent in a particular region of parameter space, normalized by the total number of steps in the chain, yields the probability that the model parameters have values within that region.

The choice of $q(\vec{\theta}_y|\vec{\theta}_x)$, by construction, does not alter the recovered posterior distribution function. The proposal distribution does, however, dramatically affect the acceptance rate of trial locations in parameter space and, therefore, the number of iterations required to satisfactorily sample the posterior. To help mitigate the multi-modal structure of the target distributions, we include parallel tempering [18] – a set of chains, running simultaneously, each at a higher "temperature" – as is becoming standard in GW applications of MCMCs [? ? ?].

We use a trans-dimensional (or "reverse jump") Markov Chain Monte Carlo (RJMCMC) [19, 20] as the tool for sampling the target posterior distribution function in model space. The RJMCMC is novel in its ability to move between models, with the number of iterations the chain spends in a particular model proportional to the marginalized likelihood for that model. This class of MCMC algorithm has previously been used to study the galactic binary detection problem on toy problems [21] and smaller data sets [15, 16], but has yet been turned loose on the full galaxy.

III. THE DATA MODEL

We model the data s as having two contributions:

$$\tilde{s}_\kappa = \tilde{n}_\kappa(\vec{\eta}_\kappa) + \sum_i^M \tilde{h}_\kappa^i(\vec{\lambda}^i). \qquad (3)$$

The instrument noise, n, is assumed to be stationary and gaussian with colored spectral density parameterized by $\vec{\eta}_\kappa$. The gravitational wave component of the data is the superposition of M gravitational wave templates in, each parameterized by $\vec{\lambda}$ which will be described in detail later. The subscript κ denotes the different interferometer channels synthesized from the LISA phase-meter data. We use the usual noise orthogonal AET channels [22]. Because we expect all of the galactic binaries to be well below the LISA transfer frequency $f_* = c/2\pi L$ we can neglect the T channel which is, in effect, GW-free for $f < f_*$.

1. The waveform model

Galactic binaries in the LISA band are expected to exhibit relatively little frequency evolution during the lifetime of the mission. Thus, the phase of the GWs emitted from the binary can be safely approximated as $\Phi(t) = \varphi_0 + 2\pi f_0 t + \pi \dot{f}_0 t^2 + ...$ where higher order derivatives of f can be neglected for binaries below ~ 9 mHz during a five-year-long mission [23]. For this work, we fix $\dot{f}_0 = 0$, as is the case in the MLDC data simulations.

Given these assumptions about the phase evolution of the binary, as well as restricting the templates to circular orbits (perhaps a dubious constraint Ref.XXX), We can fully describe a GB waveform with eight parameters:

$$\vec{\lambda} \to \left(f_0, \dot{f}_0, \theta, \phi, \mathcal{A}_0, \iota, \psi, \varphi_0 \right) \qquad (4)$$

where the subscript 0 indicates the value taken at the first time-sample in the data. Parameters $\{\theta, \phi\}$ describe the sky-location of the binary in ecliptic coordinates, and $\{\iota, \psi, \varphi_0\}$ are angles that fix the orientation of the binary. The amplitude

$$A_0 = \frac{2\mathcal{M}^{5/3}\left(\pi f_0\right)^{2/3}}{D_L} \qquad (5)$$

couples the chirp mass $\mathcal{M}$ and luminosity distance D_L preventing the independent measurement of either quantity. If the binary orbital evolution is driven purely by the emission of gravitational waves (as opposed to, for instance, mass transfer between the individual stars in the binary), then the linear term in the frequency evolution depends only on the frequency and chirp mass via:

$$\dot{f} = \frac{96}{5}\pi^{8/3}\mathcal{M}^{5/3}f^{11/3}. \qquad (6)$$

Sources which satisfy this condition will henceforth be referred to as *detached* binaries. For this data set, we have the advantage of knowing that any binaries with $\dot{f}_0 > 0$ are detached, allowing us to make a determination of D_L. When analyzing data from the Galaxy itself, we will not have this foresight. For the rare cases where $\dot{f}_0$ can also be measured, systems being driven only by the emission of gravitational radiation must satisfy the braking index condition Ref.XXX

$$n \equiv \frac{f\ddot{f}}{\dot{f}^2}. \qquad (7)$$

LISA's ability to measure $\dot{f}_0$, and how ignoring this parameter impacts the data analysis, has been preliminarily explored in [23] and [24]. We will address this important detail in a follow-on study [25].

To compute the instrument response to a particular galactic binary signal we use the fast-slow decomposition as detailed in [26].

We utilize a prior on the location of any given binary constructed from the number density of stars in the galaxy. Our model for the galactic distribution is similar to that from which the MLDC datasets were drawn Ref.XXX. The density profile has two components, one from the disk and one from the bulge:

$$\rho_{\text{bulge}} = \frac{1}{\left(\sqrt{\pi}R_b\right)^3}e^{\frac{-r_{\text{gc}}^2}{R_b^2}}$$

$$\rho_{\text{disk}} = \frac{1}{4\pi R_d^2 Z_d}\text{sech}^2\left(\frac{z_{\text{gc}}}{Z_d}\right)e^{\frac{-u_{\text{gc}}}{R_d}}$$

$$\rho_{\text{galaxy}} = A\rho_{\text{bulge}} + (1-A)\rho_{\text{disk}} \qquad (8)$$

where $r_{\text{gc}}^2 = x_{\text{gc}}^2 + y_{\text{gc}}^2 + z_{\text{gc}}^2$, $u_{\text{gc}}^2 = x_{\text{gc}}^2 + y_{\text{gc}}^2$, and $\{x_{\text{gc}}, y_{\text{gc}}, z_{\text{gc}}\}$ are the cartesian galactic coordinates of the source. For our purposes, the parameters of the galaxy distribution used are [27]

$$\{R_b, R_d, Z_d, A\} = \{690 \text{ pc}, 2520 \text{ pc}, 302 \text{ pc}, 0.24\} \qquad (9)$$

Using this distribution we build a joint prior

$$p(f_0, \dot{f}_0, \theta, \phi, \mathcal{A}_0) = C\rho_{\text{galaxy}} \qquad (10)$$

with a complicated normalization constant C that we approximate by Monte Carlo integration over the prior volume. The quantities not constrained by this prior are the orientation parameters, which we take as having uniform *a priori* distributions over $[0, 2\pi]$ for ψ and φ_0, and $[-1, 1]$ for $\cos\iota$.

For templates of detached systems we determine the distance to the binary using equations 5 and 6. For mass-transferring binaries we have no way of making such a determination without better understanding of the orbital dynamics. We construct a separate prior for such systems, marginalizing Eq. 10 over f_0, $\dot{f}_0$, and $\mathcal{A}_0$ leaving behind a prior on $\{\theta, \phi\}$ only and uniform distributions on the marginalized parameters.

2. The noise model

Given nominal levels for the shot- and acceleration-noise (S_s and S_a), the baseline noise power spectral density for the LISA A and E channels is

$$S_n(f) = \frac{4}{3}\sin^2\frac{f}{f_*}\left[\left(2+\cos\frac{f}{f_*}\right)S_s \right.$$
$$\left. + 2\left(3+2\cos\frac{f}{f_*}+\cos\frac{2f}{f_*}\right)\frac{S_a}{(2\pi f)^4}\right]. \quad (11)$$

To this we must add an estimate of the confusion noise S_c which is derived from data simulations

$$S_c(f) = \begin{cases} 10^{-44.8}f^{-2.4} & 10^{-4} < f < 4.5\times 10^{-4} \\ 10^{-47.15}f^{-3.1} & 4.5\times 10^{-4} < f < 1.1\times 10^{-3} \\ 10^{-51}f^{-4.4} & 1.1\times 10^{-3} < f < 1.7\times 10^{-3} \\ 10^{-74.7}f^{-13} & 1.7\times 10^{-3} < f < 2.5\times 10^{-3} \\ 10^{-59.15}f^{-7} & 2.5\times 10^{-3} < f < 4\times 10^{-3} \end{cases} \quad (12)$$

To allow for modeling error in the noise levels, and the vagaries of the particular noise realization in the data, we include parameters which characterize departures from this theoretical noise power spectral density as described in [15]. A separate noise level is defined for each of several narrow bandwidth segments of data, each of length N_{NB} frequency bins. The ith segment is rescaled as $S_n(f) \to \eta_i S_n(f)$. For Gaussian noise, the expectation value for η when measured over N_{NB} bins is $\sigma_\eta^2 = 1/\sqrt{N_{\mathrm{NB}}}$. We accommodate for additional ignorance with respect to the noise level by using a normal distribution $N[1, 4\sigma_\eta^2]$ as the prior on each η_i.

3. The likelihood function

With the noise and signal models now declared, the likelihood $p(s|\vec{\theta})$ is computed over N Fourier bins, built from the assumption of colored Gaussian noise where the noise power spectral density is being fit over $N_{\mathrm{seg}} = N/N_{\mathrm{NB}}$ narrowband segments of data, via:

$$\ln p(s|\vec{\theta}) = -\frac{1}{2}\sum_\kappa^{\mathrm{A,E}}\left[(r_\kappa|r_\kappa) + N_{\mathrm{NB}}\sum_n^{N_{\mathrm{seg}}}\ln\eta_\kappa^n\right]. \quad (13)$$

The residual

$$r_\kappa = s_\kappa - \sum_i^M h_\kappa^i(\vec{\lambda}^i) \quad (14)$$

appears in equation 13 inside of the noise-weighted inner product defined as

$$(a|b) = \frac{2}{T_{\mathrm{obs}}}\sum_f \frac{\tilde{a}^*(f)\tilde{b}(f) + \tilde{a}(f)\tilde{b}^*(f)}{\eta(f)S_n(f)} \quad (15)$$

where, as described above, η takes the same value over N_{NB} Fourier bins.

IV. THE ALGORITHM

A. Overview

In this section we will describe each phase in the pipeline. Following a coarse overview of the entire procedure, a detailed, step-by-step description of the algorithm can be found in the subsections.

The data are first divided into small bandwidth subsets, each of which is independently analyzed. Each data window is studied iteratively. At each iteration, we include an additional template in the model until we reach the maximum evidence. For each iteration, we apply a variety of tricks to increase the efficiency of the search, many of which spoil the statistical properties of the chain. After the "illegal" chains have located the modes of the target posterior distribution function they are used to produce proposal distributions for the subsequent steps in the search where we take care to satisfy detailed balance. The evidence is determined by including the number of templates in the model as a parameter. For example, during iteration I the RJMCMC is allowed to move between models containing $0 \le i \le I$ templates. The model in which the RJMCMC spends the most iterations, $\mathcal{M}_{\mathrm{max}}^i$, is the one with the highest evidence. If $i_{\mathrm{max}} < I$ then that window is finished being analyzed, and the MAP parameters from model $\mathcal{M}_{\mathrm{max}}^i < I$ are stored in the master list for the full dataset.

B. Preparing the Data

The bandwidth of a typical galactic binary waveform is sufficiently small that the data can be divided into small subsets, or windows, each of which spans a relatively small range in frequency, and each window can be analyzed independently. For practical purposes, this makes it simple to scale the analysis code from a testing platform to running on the full dataset. While many CPUs are required to process all of the data, they do not need to talk to one another while doing so.

The galactic binary search will be hindered at low frequency by power from high SNR black hole binaries in the data, as well as bursts from cosmic strings. Both the black hole binaries and the cosmic strings have very unique time-frequency characteristics. This means the brightest sources can be cleanly removed from the data without marring the signal from the galaxy. The accurate detection of black holes has been a main theme of LISA science studies and is definitely a manageable task, while the high-accuracy detection of the cosmic strings has been successfully demonstrated in the previous round of the MLDC [?]. We do not see the removal of black hole or cosmic string sources as a substantial technical challenge, and so have focused the efforts in this paper on training data with only instrument noise and the signal from the galaxy. For the blind data analysis to follow this work, a collaborative effort will be needed to perform

this first cleaning step.

In each window, care needs to be taken at the edges as templates will try to fit signals from adjacent data segments which have power "leaking" into that which is being analyzed. Thankfully this was addressed in the BAM algorithm by having a smaller acceptance region within each window where the initial frequency must fall if the source is going to be taken as a true detection. Bordering this acceptance region to the edges of the window are the "wings" of the data segment where templates are included in the model but the sources they recover are not stored as detections. Adjacent windows are tiled so that the end of one acceptance region is the beginning of another. Thus we have full coverage of the data without any overlap between acceptance regions, and without the risk of double-counting sources that happen to lie at the interface of two data segments. Figure 1 shows a cartoon depiction of a data window.

The size of the windows is not something that can be fixed for all signals. The amplitude of a galactic binary waveform scales as $f^{2/3}$, meaning the discrete Fourier transform of a signal typically has significant power across a larger bandwidth as we move to higher frequency data. Furthermore, for detached binaries (which make up the bulk of the galactic population) $\dot{f}$ scales as $f^{11/3}$ so signals with high initial-frequency typically spread their power over more bins during the course of the observation.

Because of this frequency-dependent bandwidth, the size of the wings and, for efficiency's sake, the size of the windows, is frequency dependent.

C. The Search Phase

The purpose of the search phase is to rapidly locate the sources in the data (i.e., the modes of the posterior distribution function). Here we are not concerned with satisfying detailed balance, or producing samples which represent the posterior, but instead are focused on efficiency. The two most substantial cost saving enhancements come from reducing the dimension of the search by maximizing the likelihood over "extrinsic parameters" (orientation and distance) using the F-statistic [28, 29], and by making the signals a bigger target in the search space through simulated annealing.

Simulated annealing is another common trick performed when using Markov Chain Monte Carlo-like methods to rapidly locate the modes of the distribution. It works by initially suppressing the influence of the likelihood terms in the Hasting's ratio (Eq 2) by "heating" the distribution being searched

$$\frac{p(s|\vec{\theta}_y)}{p(s|\vec{\theta}_x)} \rightarrow \left(\frac{p(s|\vec{\theta}_y)}{p(s|\vec{\theta}_x)}\right)^{\beta} \qquad (16)$$

with the exponent playing the role of an inverse "temperature" $\beta = 1/T$ with $0 \leq \beta \leq 1$. Early iterations of
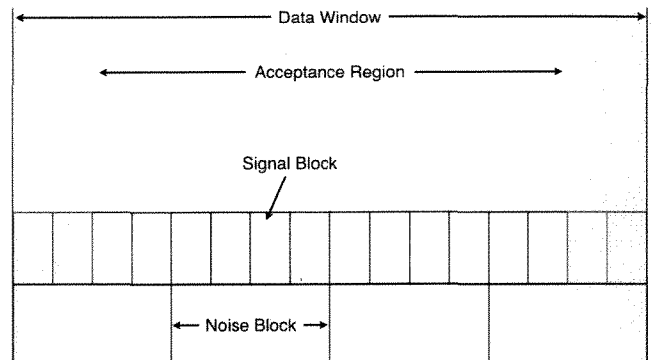

FIG. 1: A cartoon depiction of a data window to be independently searched. The shaded portions are where detected binaries will be excluded from the master list. All templates within the same Signal Block are updated simultaneously. One parameter η is used to measure the noise PSD over each Noise Block. This concept was derived for the BAM algorithm and adapted from Ref. [10]

the chain are run at high temperature (small β) forcing the likelihood ratio term to ~ 1. The influence of the likelihood is gradually increased as the search space is gradually "cooled" at each iteration i via

$$\beta = \begin{cases} \left(\frac{1}{T_{\max}}\right) T_{\max}^{i/\tau} & 0 \leq i < \tau \\ 1 & i \geq \tau \end{cases} \qquad (17)$$

until β goes to 1 and the chains are sampling the target distribution.

Simulated annealing requires some tuning in order for it to actually improve the search's efficiency. The adjustable parameters to the annealing scheme, the maximum temperature $T_{\max}$ and the cooling time τ, need to be custom suited for each problem. Conceptually, what we want is for $T_{\max}$ to be high enough that chains are able to freely explore the full prior range, while not being so absurdly hot that we spend many iterations with no hope of locking on to any of the signal. A reasonable rule of thumb is that the effective SNR of the signals as seen by the tempered chain is reduced from the true SNR by a factor of $\sim 1/\sqrt{T}$. To this end, we want $T_{\max}$ to be $\sim \text{SNR}^2$, and can reasonably determine it based on the excess Fourier power in the data:

$$T_{\max} = (s|s) - 4N. \qquad (18)$$

Setting the cooling rate τ is an exercise in trial and error, and depends on the bandwidth of the data window, with higher frequencies warranting longer cooling times.

In addition to simulated annealing, and in all other phases of the analysis, we use the now commonplace method of "parallel tempering" where multiple chains are run simultaneously at different temperatures, with exchanges of parameters between chains subject to the detailed balance condition.

To further increase the efficiency of the search, we wish to reduce the volume of the search space. For this purpose, the F-statistic is a tool which has proven extremely useful in LIGO/Virgo searches [30], black hole searches for LISA [31], and galactic binaries [10, 29?]. Because of it's frequent use in the GW data analysis literature, we leave the details to the aforementioned references.

While the speed-up in the search time when using the F-statistic is substantial, once the (approximately) best-fit frequency and sky location for each source in the model have been located, it becomes a liability. For a model containing N_S sources, a single F-statistic evaluation involves $4N_S$ calls to the waveform generator as well as a $4N_S \times 4N_S$ matrix inversion, ultimately costing more than $4N_S$ likelihood evaluations. Therefore, once the F-statistic has done its job and found the modes, the chains are much more efficient reverting to the Gaussian likelihood as described in §II. The points of the chain from the F-statistic search are discarded as the "burn in" samples.

At this point we are interested in producing an ensemble of samples that approximate the target posterior distribution function. However, we are not yet ready to abandon some of our cost saving measures in favor of detailed balance. The posteriors for these signals are multimodal and in a few percent of trial runs, the burn-in phase ends on a secondary maximum of the distribution. The nature of these near-degeneracies was explored with detail in [10]. To summarize, the orbital motion of the LISA constellation, as well as the finite number of data points, imparts a harmonic structure to the waveforms on either side of the initial frequency of the binary. The templates can fix themselves to a sub-dominant harmonic while still achieving overlaps with the injected signals of $\gtrsim 70\%$.

Such features in the posterior expose the weaknesses of an "out of the box" MCMC. While the chain is guaranteed to eventually converge, in most cases we are not willing to wait long enough. While there are certainly more elegant ways of overcoming these types of challenges, we offer a brute-force approach inspired by delayed rejection [32, 33]. We propose some position in parameter space $\vec{\theta}_y$ and, without asking the Hastings ratio for any input, temporarily adopt this position and evolve the chains from there. After some number of updates the chain arrives at $\vec{\theta}_y'$. We then look back and calculate the transition probability $\alpha_{\vec{\theta}_y' \leftrightarrow \vec{\theta}_x}$. This transition probability does not satisfy the detailed balance condition, and so the samples from the chain will be biased in some way. For a lot of effort, delayed rejection performs this type of exploration while preserving detailed balance as described in detail within a gravitational wave data analysis context by Trias et al in Ref [13]. For our purposes, we are not yet concerned with the statistical properties of our chain and accept the fact that our recovered distributions will be biased.

Our lazy implementation of delayed rejection is perfectly suited to prevent us from sticking on a subdominant harmonic of the waveform. The secondary modes appear at integer multiples of the LISA modulation frequency $f_m = 1/\text{year}$. We propose a shift in frequency by some $n f_m$, where n is a random integer drawn from U[-6,6], adopt that solution, and then allow the chains a few iterations to refine the remaining parameters (in particular, the sky location) before comparing back to the current solution. This addition to the search procedure mitigates the instances of recovered sources having the central frequency on a harmonic induced by the orbital motion.

Figure 2 is a scatter plot of chain samples in the $p(s|\vec{\theta}) - f$ plane. The plot is centered at the true frequency of the injected source. The sub-dominant modes of the distribution are clearly visible as peaks in the likelihood surface occurring at even integer multiples of $f_m T = 2$ Fourier bins. In this example, the search chain approached from higher frequency. The multi-modal structure of the likelihood surface due to orbital motion is roughly symmetric about the injected value. A single Markov chain without the benefit of delayed rejection or parallel tempering would be severely challenged by these local maxima.
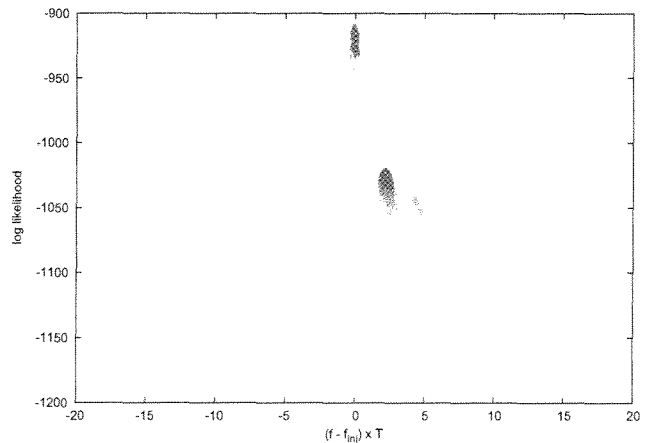

FIG. 2: Delayed Rejection at work: This scatter plot shows samples from a search chain in the $p(s|\vec{\theta}) - f$ plane. The plot is centered at the true frequency of the injected source. Delayed rejection enables the chain to efficiently leave a secondary maximum in favor of the true source parameters.

D. Parameter Estimation & Model Selection

Upon the completion of the search phase, the chains have produced samples of some biased distribution function, the modes of which closely correspond to that of the target posterior. To accurately refine the characterization of the model and to evaluate wether or not it is the one favored by the data, we run a RJMCMC following all of the rules to ensure the chain samples are representative of the target posterior. RJMCMCs are notoriously difficult to work with in high-dimension problems. The proposal distributions have to be informative enough that drawing a random point out of (for this example) an 8 dimensional parameter space produces a reasonable fit to the data, but are not so constraining that the improved fit to the data is overly penalized by how strongly you forced the trial solution to that point. For these techniques to work efficiently, the chains can not tolerate a proposal that is anything but a good approximation to the posterior.

Fortunately for us, we spent the search phase producing such an approximation. While these chains could not be used (ethically) for model selection or parameter estimation, there are no rules against employing them to build suitable proposal distributions. This concept was originally suggested by Green [20] and we have described in detail the procedure applied to this work in Ref. [15]. In short, we bin the chains into an 8D grid using fisher matrix estimates of the parameter variances to set the cell size in the grid. The number of samples from the chain that land in a given cell is proportional to the probability density in that cell. The proposal distribution randomly chooses a cell weighted by its probability, and then uniformly draws within the cell to come up with the parameters. Recently, Farr and Mandel [34] introduced an improvement on how to bin the chain samples by using a kD-tree data structure instead of our Fisher-scaled grid. It's clear that this will further improve the efficiency of these proposals by eliminating the dependence on the grid size, something which we had to carefully tune. We will transition to the kD-tree decomposition in future work.

As mentioned previously, this analysis is performed iteratively, including an additional signal in the model at each iteration. During the characterization phase of the procedure on iteration I the RJMCMC is exploring models containing anywhere between 0 and I signals. The model with the strongest support is the one in which the RJMCMC spends the most iterations. When two models are similarly supported by the data a more subtle selection has to be made. Between any two models we can calculate the Bayes factor

$$B_{ij} = \frac{p(s|\mathcal{M}_j)}{p(s|\mathcal{M}_i)} = \frac{\text{\# of iterations in } \mathcal{M}_j}{\text{\# of iterations in } \mathcal{M}_i} \qquad (19)$$

which is easily interpreted as the odds (ignoring prior preferences for one model over another) that $\mathcal{M}_j$) is preferred over $\mathcal{M}_i$. Thus, $B_{ij} \sim 1$ means that both models

are similarly supported by the data. While that is a nice interpretation, when assembling a catalogue of detectable galactic binaries and producing a residual fit for subsequent searches, decisions need to be made about which model to pick in these marginally distinguishable cases. This, sadly, somewhat reduces the Bayes factor to a statistic used for model selection. Nevertheless, we have to draw a line in the sand, and have chosen for this study a Bayes factor "threshold" of 12:1 needed to prefer a higher dimensional model, above which the support for $\mathcal{M}_j$ is canonically considered "strong" [35].

A more satisfying thing to do would be to repeat the analysis for several different B_{ij} cutoffs, producing appendices to the final source catalogue of more speculative detections, but the time for such an exercise was not warranted at this juncture.

If the highest dimension model is supported by the data, we store the map parameters and begin another iteration. If not, the MAP parameters from the winning model are stored and no more iterations on that window are performed. Only sources with initial frequency inside the acceptance region are added to the "master" list of detections. Figure 3 shows ~ 5 windows' worth of training data, with the best fit waveforms over-plotted. A higher frequency window was chosen so that the different signals in the fit could be distinguished.
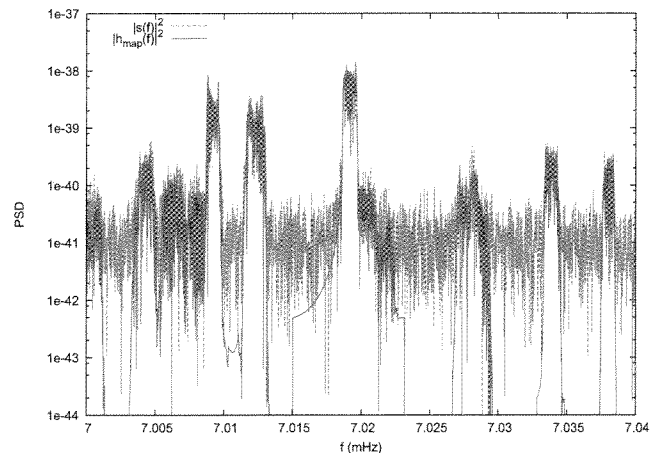

FIG. 3: Power spectral density of a small bandwidth segment of the training data. This figure spans ~ 5 search windows. The green [dashed] line is the original data, while the red [solid] line shows the best fit signal model.

E. Source Subtraction

The biggest cost to performing these MCMC runs is the waveform calculation. While our waveform model is extremely efficient, there is not much that can be done about the colossal number of templates that need to be computed to resolve 10000+ sources. To help mitigate

this expense, we want to hold fixed the waveform parameters for sources which are not significantly overlapping other sources in the window.

After each iteration, if the new source is within some pre-defined number of frequency bins (depending on the characteristic bandwidth of sources in that window), than on the following iteration this template, along with any others that satisfy the closeness condition, and the new source included in the model, are allowed to vary. Otherwise, the signals located in previous iterations are held fixed. This keeps the number of "active" sources per window, per iteration, much lower than the total number of detectable binaries in that segment of data. We are essentially searching on the residual of the data from previous iterations, but with the caveat that if new sources get too close to existing detections than all nearby signals need to be simultaneously re-characterized. FIgure 4 shows the same data segment as in Figure 3, but with the best fit model regressed from the original data. The residual is consistent with stationary Gaussian noise at the level of the instrument noise.
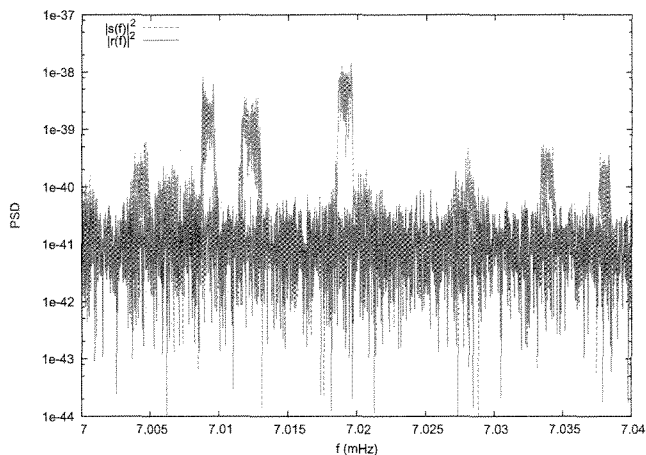

FIG. 4: Same as Figure 3, but with the red [solid] line replaced by the residual power spectral density after the best fit model has been regressed from the data.

V. THE RECOVERED SOURCE CATALOGUE

We performed a comprehensive test run of the algorithm on the MLDC challenge 4 training data set, containing only instrument noise and the signals from the simulated galaxy. In the interest of saving time for the blind data, the search was not carried to completion in all frequency windows. In this section we quantify the performance of the algorithm by comparing the recovered catalogue to the source list supplied with the training data.

1. Evaluating the recovered catalogue

. The total number of detected galactic binaries that we located in the data before moving on to the blind data set was ~ 9000. The merit of our recovered catalogue was judged using the MLDC challenge 3 evaluation software available as part of the lisatools software package [?].

For each recovered source $h_{\rm rec}$, the evaluation software searches through the list of (SNR $\gtrsim 1$) sources in the simulated galaxy, and determines which injected signal $h_{\rm inj}$ gives the lowest noise-weighted residual $(h_{\rm inj} - h_{\rm rec}|h_{\rm inj} - h_{\rm rec})$. Once $h_{\rm rec}$ is paired with the corresponding $h_{\rm inj}$ the correlation between the two waveforms

$$\text{Corr}_{\rm inj,rec} \equiv \frac{(h_{\rm inj}|h_{\rm rec})}{\sqrt{(h_{\rm inj}|h_{\rm inj})(h_{\rm rec}|h_{\rm rec})}} \qquad (20)$$

is computed and stored in a "report card" for the recovered catalogue. A histogram of the correlations for our recovered population of sources is depicted in Figure 5.
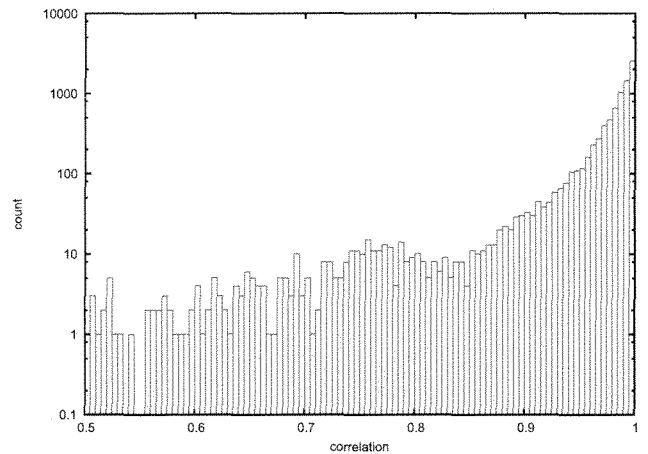

FIG. 5: Correlations between recovered sources and their corresponding signal in the training data. Of the ~ 9000 detections over 90% had a correlation above 0.9 with an injected signal.

Along with the correlation files, the MLDC evaluation software saves the parameter error between each recovered signals and its corresponding source in the data. The distribution of errors for the entire catalogue will reveal any systematic biases in the parameter recovery. Figure 6 shows the distribution of recovered parameter biases. We are pleased to report that all parameters show a strong peak at zero bias.

The galaxy search serves the dual purpose of recovering a trove of information about the galaxy, and removing a substantial amount of foreground signal-power, facilitating the search for signals at cosmological distances. The residual power spectral density of the training data after having removed the recovered signals is shown in Figure 7. The dashed [green] trace shows the galaxy-only training data (without BHBs, EMRIs, etc.), while the
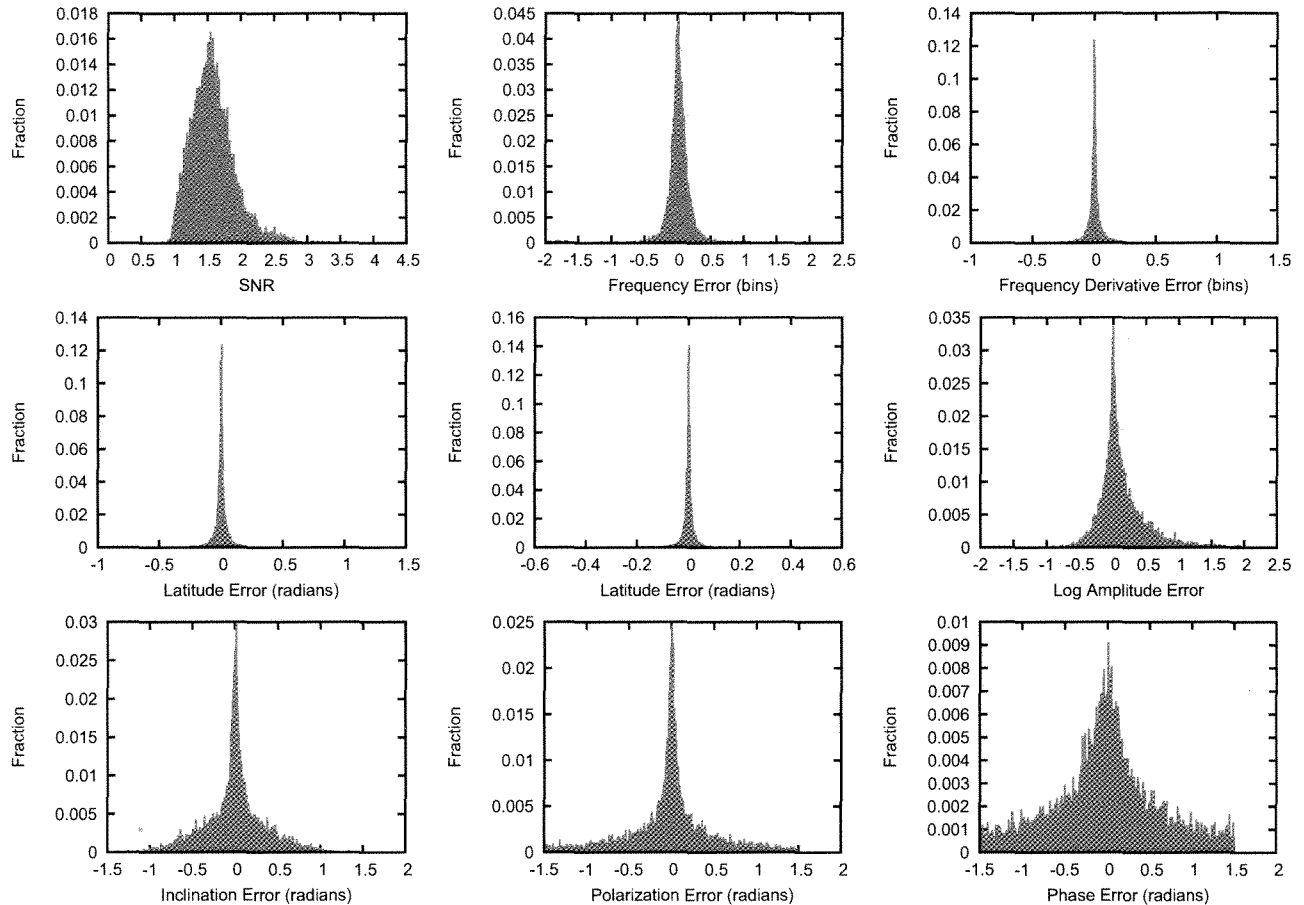
FIG. 6: [Top left] The distribution of SNRs for the recovered signals. The x-axis is in $\log_{10}$ SNR.. [All others] Distributions of the difference between the recovered signal parameters and the associated source in the data. We see no significant systematic biases between our best fit parameters and true values for the recovered sources. The orientation parameters of the binary are typically not well measured, hence the larger tails on the error distributions.

solid [red] line is the residual after our recovered catalogue has been regressed from the data. Notice we only performed the search out to 0.01 Hz. At higher frequencies the signals are typically bright and isolated. This makes them easy to find, but computationally expensive to characterize, as the bandwidth of the signal increases with frequency and amplitude. Individual high frequency windows have been tested and will be included in our blind search, but we omit them from the results here. It is clear from the residual that there are more detectable sources than the 9000 on which we are reporting. However, as a test of the algorithm we take this as a satisfying result.

precedented astronomical tool. Our follow on work to this paper will address some specifics, but in the mean time there is one piece of "low hanging fruit" that is too good to pass up. As mentioned in §??, the inclusion of $\dot{f}_0$ in the waveform model presents us with an opportunity to disentangle the luminosity distance from the overall GW amplitude. To demonstrate this capability, we select from the catalogue any binary with SNR>15 and $\dot{f}_0 T^2 \geq 1$ and compute the D_L using Eqs ??. The resulting galactic map is shown in Figure 8. As promoted in the LISA science case studies [?], we recover a three-dimensional map of the entire Galaxy.

2. Galactic binary astronomy from the recovered catalogue

. While the focus of this work has been the algorithm and its performance on simulated data, the real fun begins when we brandish the source catalogue as an un-

Admittedly, we artificially benefit from knowing that any binary with $\dot{f}_0 > 0$ is dynamically evolving under radiation reaction forces only. How this assumption biases the spatial distribution of the detected population is a top priority of our subsequent studies.
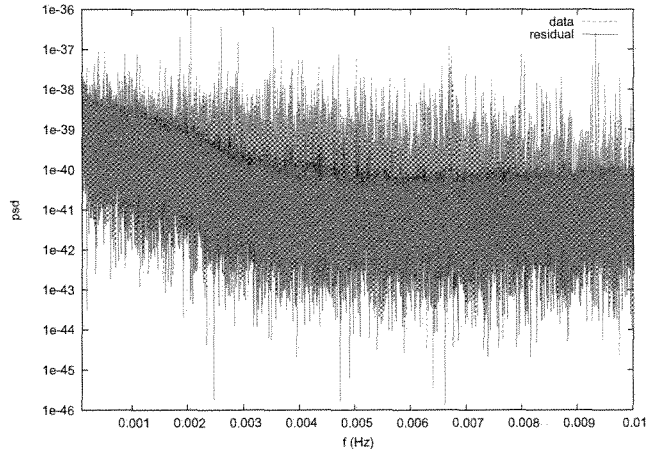
FIG. 7: Power spectral density of the training data (dashed/green) and residual (solid/red) after removal of the recovered sources. The remaining spikes in the residual are detectable signals left behind after we prematurely halted the search in favor of analyzing the blind data.

VI. DISCUSSION

Using [10] as a foundation, we have developed our own galactic binary detection code and tested it on the round 4.0 training data supplied by the Mock LISA Data Challenge Task Force. New features to this search algorithm are the inclusion of parallel tempering and delayed rejection to increase the efficiency of the search, and using a trans-dimensional MCMC to simultaneously characterize the detected binaries, and determine the most likely number of binaries in each small-bandwidth segment of data being analyzed. We also include the $\dot{f}_0$ parameter in the waveform modeling which was not part of the orig-

inal work by Crowder and Cornish, but was included in the MLDC round 3 data and subsequent entries.

The codes were written for the purpose of participating in the blind challenge, so the analysis of the training data was halted before completion. Before moving on to the challenge data, we recovered $\sim$ 9000 sources, 90% of which matched one of the injected sources with correlation greater than 0.90. It is worth noting that a smaller-scale space borne gravitational wave detector is more likely to fly in the near future than the full 5 Gm LISA mission concept. If anything, that makes our current search software overqualified to handle whatever data we are presented.

Of course, detecting the sources is the first step towards using the data to its full potential, as a unique probe of the Galaxy.

Including the first time derivative of the frequency in the waveform model allows us to compute the distance to binaries whose dynamics are not affected by mass transfer or tidal effects. From the recovered source catalogue, we measured the sky-location and $\dot{f}$ well enough to constrain D_L of $\sim$ 1000 binaries, sampled from the entire volume of the Galaxy. This result hints at the potential for low-frequency gravitational wave astronomy to offer an unprecedented view of compact stellar remnants are distributed. Implicit in the calculation of the distance is the supposition that all mass-transferring systems have $\dot{f} < 0$. How our ability to map the distribution of the Galaxy after relaxing this assumption is impacted will be the focus of follow-on work.

VII. ACKNOWLEDGMENTS

[1] LSC, LIGO Scientific Collaboration.
[2] VIRGO, European Gravitational Observatory.
[3] A. Stroeer and A. Vecchio, Class. Quant. Grav. **23**, S809 (2006).
[4] D. Hils, P. L. Bender, and R. F. Webbink, Astrophys. J. **360**, 75 (1990).
[5] G. Nelemans, L. R. Yungelson, and S. F. Portegies Zwart, Astron. Astrophys. **375**, 890 (2001).
[6] G. Nelemans, L. R. Yungelson, and S. F. Portegies Zwart, Mon. Not. Roy. Astron. Soc. **349**, 181 (2004).
[7] P. Bender et al., (1998).
[8] MLDC, Mock LISA Data Challenge.
[9] S. Babak et al., Class. Quant. Grav. **25**, 184026 (2008).
[10] J. Crowder and N. Cornish, Phys. Rev. **D75**, 043008 (2007).
[11] A. Blaut, S. Babak, and A. Krolak, Phys. Rev. **D81**, 063008 (2010).
[12] J. T. Whelan, R. Prix, and D. Khurana, Class. Quant.

Grav. **27**, 055010 (2010).
[13] M. Trias, A. Vecchio, and J. Veitch, Class. Quant. Grav. **26**, 204024 (2009).
[14] R. Umstatter et al., Phys. Rev. **D72**, 022001 (2005).
[15] T. B. Littenberg and N. J. Cornish, Phys. Rev. **D80**, 063007 (2009).
[16] A. Stroeer and J. Veitch, (2009).
[17] P. C. Gregory, Bayesian Logical Data Analysis for the Physical Sciences (Cambridge University Press, The Edinburgh Building, Cambridge CB2 2RU, UK, 2006), No. ISBN 0 521 84150 X.
[18] R. Swendsen and J. Wang, Phys. Rev. Lett. **57**, 2607 (1986).
[19] P. Green, Biometrika **82**, 711 (1995).
[20] P. Green, in Highly Structured Stochastic Systems, edited by P. Green, N. Hjort, and S. Richardson (Oxford University Press, ADDRESS, 2003).
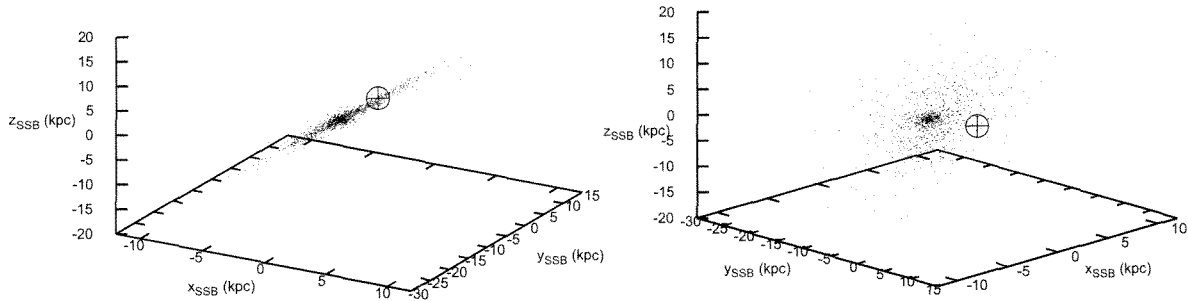[21] R. Umstatter et al., Class. Quant. Grav. **22**, S901 (2005).

FIG. 8: Scatter plots showing the spatial distribution of recovered signals for which D_L could be measured in ecliptic coordinates from different viewing angles. The $\oplus$ symbol is located at the solar system barycenter.

[22] M. Tinto, F. B. Estabrook, and J. W. Armstrong, Phys. Rev. **D65**, 082003 (2002).
[23] A. Stroeer, A. Vecchio, and G. Nelemans, Astrophys. J. **633**, L33 (2005).
[24] A. Stroeer and G. Nelemans, (2009).
[25] S. L. Larson and T. B. Littenberg, in preperation (2011).
[26] N. J. Cornish and T. B. Littenberg, Phys. Rev. **D76**, 083006 (2007).
[27] M. R. Adams, private communication .
[28] P. Jaranowski, A. Krolak, and B. F. Schutz, Phys. Rev. **D58**, 063001 (1998).
[29] A. Krolak, M. Tinto, and M. Vallisneri, Phys. Rev. **D70**,

022003 (2004).
[30] C. Cutler and B. F. Schutz, Phys. Rev. **D72**, 063006 (2005).
[31] N. J. Cornish and E. K. Porter, Class. Quant. Grav. **24**, 5729 (2007).
[32] L. Tierney and A. Mira, Stat. Med. **18**, (1999).
[33] A. Mira, Metron **LIX**, 231 (2001).
[34] W. M. Farr and I. Mandel, (2011).
[35] A. Raftery, *Practical Markov Chain Monte Carlo* (Chapman and Hall, ADDRESS, 1996).