A proof-of-concept analysis of the second round of the MLDC data sets to search for an isotropic stochastic signal

Emma L. Robinson(1) and Alberto Vecchio(1,2)

(1) School of Physics and Astronomy, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK
(2) Department of Physics and Astronomy, Northwestern University, 2131 Tech Drive, Evanston, IL 60208, USA

(Dated: June 18, 2007)

I. INTRODUCTION

We consider an approach to the search for a stochastic signal in the LISA data stream. Our algorithm is designed to identify an isotropic stochastic contribution by exploiting the different response to gravitational radiation of the TDI variables \(A\), \(E\) and \(T\) in the low frequency region \((\lesssim 1\,\text{mHz})\) of the observational window.

The second round data sets of the MLDC provide an opportunity to test this approach by exploiting the foreground signal produced by the population of galactic binaries. Although the signal from a galactic population of WDs is well known to be anisotropic, in this study we focus on the isotropic component of this radiation; our analysis technique can, however, be generalized to take into account anisotropies.

Here we present results from the analysis of the challenge 2.1 data set. The analysis approach and the results should be considered as a proof-of-principle, and a first step toward the development of an end-to-end analysis pipeline to study (isotropic) stochastic signals. We also hope that it can be useful in addressing some of the subtleties associated with the analysis issues that one confronts for this class of signals.

II. ANALYSIS SCHEME

We consider the TDI observables \(A\), \(E\) and \(T\) that can be constructed from the TDI-Michelson variables \(X\), \(Y\) and \(Z\), in which the MLDC data sets are released. Let us consider first a small frequency interval around a frequency \(f\) in which the instrumental noises and GW stochastic signal can be considered as constant. Then the variance of the \(A\), \(E\) and \(T\) channels can be formally written as:

\[
\langle |\hat{A}(f)|^2 \rangle = \langle |\hat{E}(f)|^2 \rangle = \sigma_n^2(1 + \epsilon(f)) \tag{1}
\]

\[
\langle |\hat{T}(f)|^2 \rangle = \sigma_T^2(f) = k(f)\sigma_n^2(f). \tag{2}
\]

where \(\sigma_n^2(f)\) is the variance of the instrumental noise in \(A\) and \(E\), \(\sigma_T^2(f)\) is the variance of the \(T\) channel (that to good approximation does not contain a contribution from the GW signal) and \(\sigma_n^2(f)\) is the variance of the GW stochastic signal; \(\hat{A}(f)\), \(\hat{E}(f)\) and \(\hat{T}(f)\) represent the Fourier transform of \(A\), \(E\) and \(T\), respectively. Over a sufficiently small frequency region \(\Delta f\) centered on \(f\), all the above quantities can be considered as constant, but we have explicitly included the dependency on \(f\) to stress their variation over the full observational window.

We thus have three unknown variables: the variance of the noise in the \(A\) and \(E\) channels, \(\sigma_n^2\), an estimate of the relation between the noise in the \(A\) and \(E\) channels and the \(T\) channel, \(k\), and an estimate of the strength of the GW signal when compared with the detector noise, \(\epsilon\). The key method by which to identify a GW contribution is to measure \(\epsilon\) from the data, once we have estimated \(k\), or marginalized over its uncertainty; if \(\epsilon > 0\) then we can conclude that a signal is present. This method follows conceptually the ideas put forward in [1, 2], although it is now cast in a rigorous formalism and applied for the first time to an actual (synthetic) data set.

The analysis scheme proceeds in the following way. We split the time-domain data into \(n\) segments of length \(\tau\) and Fourier transform each segment. We then consider a frequency band over which the noise and GW signals are approximately constant (containing \(\nu\) frequency bins), so that \(\sigma_n^2(f) \simeq \sigma_n^2\), \(\sigma_T^2(f) \simeq \sigma_T^2\), \(k(f) \simeq k\) and \(\epsilon(f) \simeq \epsilon\). We can construct a likelihood function on \(\epsilon\), \(k\) and \(\sigma_n\), given a set of \(A\), \(E\) and \(T\) data \(\{A_{JK}\}, \{E_{JK}\}\) and \(\{T_{JK}\}\) for \(J = 0, \ldots, n-1\) and \(K = 0, \ldots, \nu-1\):

\[
L(\epsilon, k, \sigma_n|\{A_{JK}\}, \{E_{JK}\}, \{T_{JK}\}) = (2\pi)^{-3\nu/2} \sigma_n^{-3\nu} (1 + \epsilon)^{-\nu} k^{-\nu/2} \times \exp \left( -\frac{1}{2\sigma_n^2} \sum_{J=0}^{n-1} \sum_{K=0}^{\nu-1} \frac{\left|{\hat{A}_{JK}}\right|^2}{1 + \epsilon} + \frac{\left|{\hat{E}_{JK}}\right|^2}{1 + \epsilon} + \frac{\left|{\hat{T}_{JK}}\right|^2}{k} \right). \tag{3}
\]

We use a Jeffrey’s prior on \(\sigma_n\) and flat priors on \(\epsilon\) and \(k\) to produce a posterior PDF on all three variables. We then...
marginalize over $\sigma_n$ to obtain a posterior on $\epsilon$ and $k$:

$$p(\epsilon, k|\{A_{jk}\}, \{E_{jk}\}, \{T_{jk}\}) \propto \left(\frac{1 + \epsilon}{k}\right)^{n_\nu/2} \left(\sum_{j=0}^{n-1}\sum_{K=0}^{\nu-1}|\tilde{T}_{jk}|^2\right)^{-3n_\nu/2} \times \left(\frac{\sum_{j=0}^{n-1}\sum_{K=0}^{\nu-1}|\tilde{A}_{jk}|^2 + \sum_{j=0}^{n-1}\sum_{K=0}^{\nu-1}|\tilde{E}_{jk}|^2}{\sum_{j=0}^{n-1}\sum_{K=0}^{\nu-1}|\tilde{T}_{jk}|^2 + \sum_{j=0}^{n-1}\sum_{K=0}^{\nu-1}|\tilde{T}_{jk}|^2 + \frac{1 + \epsilon}{k}}\right)^{-3n_\nu/2}.$$  \label{eq:4}

Unsurprisingly, the result depends on the ratio $(1 + \epsilon)/k$ of the two unknown quantities. Starting from Eq. (4) we can then perform a numerical integration to marginalize the PDF over $k$, to obtain a posterior PDF on $\epsilon$. However, the results on $\epsilon$ critically depend on the prior knowledge on $k$ - this lies at the very heart of the challenge of identifying an isotropic GW stochastic contribution. One would hope that it is possible to derive a deterministic relation between $k$ and other (independent) observable variables, but to our knowledge there are no results in the published literature that shed light on this issue. It has been argued that one could obtain some prior knowledge on the range of $k$ through modeling of the instrument and monitoring of subsystems [2]; however it would be preferable to base the analysis on the data only. We still do not have a “clean” solution for this problem; here we take a pragmatic approach and show through the analysis of the challenge 2.1 data sets the strength and limitations of the current analysis strategy and the impact of different assumptions on $k$, and hope to address some subtleties associated with this analysis challenge.

### III. RESULTS

Our results are based on the analysis of the challenge 2.1 data set, where we consider only the first year of data. This choice is purely out of convenience and has no impact on the results of the analysis (except for a small reduction of SNR). For completeness we also show the analysis of the training 2.1 data set, where we consider only the first year of data. Our results are based on the analysis of the challenge 2.1 data set, where we consider only the first year of data. We then consider the challenge 1.1.1a data set, from which we estimate $k$ in the frequency range $0.1 \text{ mHz} \leq f \leq 1 \text{ mHz}$, to obtain a posterior PDF on $k$, which is essential to derive some of our results. The 1.1.1a data set can be considered a “noise-only” data set as the only signal it contains is radiation from a single WD binary system at a frequency $>1$ mHz; we can therefore use it to independently estimate $k$. This is clearly a luxury we will not have during the real mission, but it is highly valuable to highlight both the strengths and the limitations of our analysis approach. Additionally, we present the results of the analysis of the challenge 1.1.1b data set, which also contains no signal at the frequency range we are studying; this allows us to demonstrate that our analysis finds results consistent with no signal being present.

The parameters of our analysis are as follows. We break up the first year of the data set 2.1 into overlapping Hann-windowed segments of equal length $\tau = 61440$ seconds. We then carry out the analysis over three representative frequency bands:

- band 1: $0.1 \text{ mHz} \leq f \leq 0.2 \text{ mHz}$
- band 2: $0.6 \text{ mHz} \leq f \leq 0.7 \text{ mHz}$
- band 3: $0.8 \text{ mHz} \leq f \leq 0.9 \text{ mHz}$

Each band contains $\nu = 6$ or 7 frequency bins - in the analysis we assume (which is clearly not strictly true) that $k$, $\epsilon$, $\sigma_n$ and $\sigma_k$ are exactly constant and we consider the three bands independently and in turn.

We then consider the challenge 1.1.1a data set, from which we estimate $k$ in the frequency range $0.1 \text{ mHz} \leq f \leq 1 \text{ mHz}$, to obtain a guideline for the choice of prior for the analysis, as well as a “known” value, $\tilde{k}$. We split the $A$ and $T$ data, as before, into Hann-windowed, overlapping segments of length $\tau = 61440$ seconds and Fourier transformed each segment; we then combined the segments to find an estimate of the variances over the whole data set. The expected variance of the $T$ channel is the same as above (Equation 2) but, as there is no signal present, the variance of the $A$ channel is now simply

$$\langle|\tilde{A}(f)|^2\rangle = \sigma_A^2(f).$$  \label{eq:5}

We can therefore use the obtained variances to find an estimate $\hat{k}(f) = |\tilde{T}(f)|^2/|\tilde{A}(f)|^2$ in each frequency bin. We then find the mean value $\bar{k} = \frac{1}{T} \sum_{t=1}^{T} \hat{k}(f_k)$ over the band of interest. The values are reported in Table I, along with the prior ranges used in the following analyses.

Some words of caution are necessary before we present the results. As we have stressed, we consider the signal constant over the chosen frequency band, whereas a weak dependency on the frequency is actually present. This can
be easily rectified in future analyses. We do not consider the fact that the signal is actually anisotropic and its average level changes during the observation time. There is no fundamental obstacle to including this in the analysis, but it does become much more complex and we have not had the time to implement it. We do not apply any preprocessing to the data that would make sure that the analysis provides an estimate of “only the stochastic contribution” of the GW sky. In reality, individually resolvable strong sources contaminate our observational band, and strictly speaking we estimate the average spectral contribution in a given frequency band, which is given by the sum of the stochastic signal and average power over the year from sources that are (in principle) resolvable. Thus our results overestimate the true contribution from the stochastic signal. In a real analysis one would proceed in a more careful way, using tools that are already becoming available to first analyze the relevant frequency band for resolvable systems, and then apply the analysis presented here to the “remaining contribution” (either by using the estimate of the effective stochastic contribution which is given by several MCMC approaches now in use, or by cleaning the identified sources from the data set). There is no fundamental reason that prevented us from carrying out this more accurate analysis; however the scope of the work would have increased tremendously, and we are not yet set up to carry out such a large scale analysis.

### A. Challenge 2.1 data set

In the following we summarize the results obtained from the analysis of the first year of the challenge 2.1 data set. We present in turn the results for the three frequency bands for each choice of prior on $k$.

Figure 1 shows the posterior joint PDF on $\epsilon$ and $k$ for first year of the challenge 2.1 data set, using an unconstrained prior on $k$. Note that there is no single maximum in the PDF - this represents a good knowledge of the ratio $(1 + \epsilon)/k$, but does not allow us to constrain the individual parameters.

![Figure 1: Posterior PDFs on $\epsilon$ and $k$ for the challenge 2.1 data set, using an unconstrained prior on $k$.](image)

(a)0.1-0.2mHz.  
(b)0.6-0.7mHz.  
(c)0.8-0.9mHz.

We present in turn the results for the three frequency bands for each choice of prior on $k$.

<table>
<thead>
<tr>
<th>band</th>
<th>frequency range (mHz)</th>
<th>$k$</th>
<th>unconstrained</th>
<th>“moderately” constrained</th>
<th>“tightly” constrained</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1 - 0.2</td>
<td>$4.79 \times 10^{-5}$</td>
<td>$U[0, \infty]$</td>
<td>$U[3 \times 10^{-9}, 1 \times 10^{-8}]$</td>
<td>$U[4.5 \times 10^{-9}, 5.5 \times 10^{-9}]$</td>
</tr>
<tr>
<td>2</td>
<td>0.6 - 0.7</td>
<td>$5.88 \times 10^{-6}$</td>
<td>$U[0, \infty]$</td>
<td>$U[4 \times 10^{-6}, 8 \times 10^{-6}]$</td>
<td>$U[5.4 \times 10^{-6}, 6.6 \times 10^{-6}]$</td>
</tr>
<tr>
<td>3</td>
<td>0.8 - 0.9</td>
<td>$2.56 \times 10^{-5}$</td>
<td>$U[0, \infty]$</td>
<td>$U[1.9 \times 10^{-5}, 3.1 \times 10^{-5}]$</td>
<td>$U[2.25 \times 10^{-5}, 2.75 \times 10^{-5}]$</td>
</tr>
</tbody>
</table>

FIG. 1: Posterior PDFs on $\epsilon$ and $k$ for the challenge 2.1 data set, using an unconstrained prior on $k$. Note that there is no single maximum in the PDF - this represents a good knowledge of the ratio $(1 + \epsilon)/k$, but does not allow us to constrain the individual parameters.

TABLE I: Choice of values of $k$ obtained from the 1.1.1a data set. The data set was split into overlapping, Hann windowed segments of length $\tau = 61440$ seconds, then Fourier transformed, and an estimate of the variances of the $A$ and $T$ channels found by combining these segments. We then found the mean, $\bar{k}$, of the ratio $\tilde{k}(f) = |\tilde{T}(f)|^2/|\tilde{A}(f)|^2$ over each frequency band. The prior ranges are chosen to demonstrate how different knowledge of the value of $k$ affects the posterior on $\epsilon$; the first column is the case where we have no knowledge of $k$, the second column contains a “moderately” constrained prior on $k$ and the third column represents very good knowledge of $k$. 

In order to obtain an estimate of $\epsilon$, there are two ways one can proceed. First, if one can somehow determine the
FIG. 2: Posterior PDFs on $\epsilon$ for the challenge 2.1 data set, using a “known” value of $k$ and two different priors on $k$ (see Table I).

FIG. 3: Marginalized PDFs on $\epsilon$ at 0.6-0.7mHz, using a “known” value of $k$ and two different priors on $k$ (see Table I).

value of $k$, it can be treated as a known parameter, and the posterior PDF in Equation 4 simply becomes a function on $\epsilon$. As a demonstration, we used the estimates of $\hat{k}$ that were obtained from the challenge 1.1.1a data as our “known” values; they are shown in Table I. Figure 2 shows the posterior PDFs on $\epsilon$ in this regime for each frequency band (red line). Again, comparing the different frequency bands, we can see that the expected increase of the value of $\epsilon$ with increasing frequency.

A more realistic situation, however, would be the case in which one can estimate $k$, but with some uncertainty. If we use these values to constrain the prior on $k$, we obtain a much better estimate of $\epsilon$ when we marginalize. We used the 1.1.1a data set to choose two different sets of priors - one set is “moderately” constrained, and one is “tightly” constrained; the choices of priors are shown in Table I. The range of the latter was chosen to be $\hat{k}(1 \pm 0.01)$, while the range of the former was chosen to be the extrema of the values of $\hat{k}(f)$ in the relevant frequency band. Figure 2 also shows the marginalized posterior PDFs on epsilon for these two different priors on $k$ (the joint posterior PDFs on $\epsilon$ and $k$ are shown in Appendix A). We can see that as the knowledge of $k$ improves, the posterior PDF on $\epsilon$ approaches the PDF obtained from a “known” value of $k$.

B. Consistency checks: other data sets

We also applied these analyses to the training 2.1 data set and the challenge 1.1.1b data set, in the 0.6-0.7mHz band, for the same three different priors and “known” value of $k$ as above. Figures 3(a) and 3(b) show the marginalized posterior PDFs on $\epsilon$ for the training 2.1 and challenge 1.1.1b data sets respectively (the joint posterior PDFs are shown in Appendix A). It can be seen that, for the training 2.1 data set, the posterior PDFs on $\epsilon$ are consistent with the challenge 2.1 data set, while the challenge 1.1.1b analysis does not show any strong evidence of the presence of a signal.
IV. ACKNOWLEDGEMENTS

The authors would like to thank Joe Romano for many extensive and fruitful discussions, upon which this work is based.

APPENDIX A: JOINT POSTERIOR PDFS

FIG. 4: Posterior PDF on $\epsilon$ and $k$ for the challenge 2.1 data set, using a “moderately” constrained prior on $k$ (see Table I).

FIG. 5: Posterior PDF on $\epsilon$ and $k$ for the challenge 2.1 data set, using a “tightly” constrained prior on $k$ (see Table I).

FIG. 6: Joint posterior PDFs on $\epsilon$ and $k$ for the training 2.1 data set at 0.6-0.7mHz, using different choices of prior on $k$ (see Table I).
(a) Unconstrained prior on $k$.

(b) "Moderately" constrained prior on $k$.

(c) "Tightly" constrained prior on $k$.

FIG. 7: Joint posterior PDFs on $\epsilon$ and $k$ for the challenge 1.1.1b data set at 0.6-0.7mHz, using different choices of prior on $k$ (see Table I).