LASER INTERFEROMETER GRAVITATIONAL WAVE OBSERVATORY
- LIGO -
CALIFORNIA INSTITUTE OF TECHNOLOGY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Technical Note **LIGO-T1900287–v1** 2020/01/22

Data Clustering Techniques for the Correlation of Environmental Noise to Signals in LIGO Detectors

Jacob Bernhardt, Anamaria Effler, Rana Adhikari

California Institute of Technology
LIGO Project, MS 18-34
Pasadena, CA 91125
Phone (626) 395-2129
Fax (626) 304-9834
E-mail: info@ligo.caltech.edu

Massachusetts Institute of Technology
LIGO Project, Room NW22-295
Cambridge, MA 02139
Phone (617) 253-4824
Fax (617) 253-7014
E-mail: info@ligo.mit.edu

LIGO Hanford Observatory
Route 10, Mile Marker 2
Richland, WA 99352
Phone (509) 372-8106
Fax (509) 372-8137
E-mail: info@ligo.caltech.edu

LIGO Livingston Observatory
19100 LIGO Lane
Livingston, LA 70754
Phone (225) 686-3100
Fax (225) 686-7189
E-mail: info@ligo.caltech.edu

http://www.ligo.caltech.edu/

1 Introduction

The LIGO project uses laser interferometry to measure gravitational waves (GWs). LIGO interferometers transduce their relative arm length differences caused by GWs to a signal composed of optical power, known as DARM. Due to the amplitude scales of astrophysical GWs, The LIGO detectors have to operate at a very high sensitivity; the spectral density of a measurable length difference is as low as 2×10^{-20} m/$\sqrt{\text{Hz}}$ at 100 Hz. The design of earthbound LIGO is thus heavily focused on the filtering and isolation of environmental noise.

To help identify and characterize environment-based noise, the LIGO detector has a Physical Environment Monitoring (PEM) system, a diverse array of environmental sensors positioned all over the facility[1]. This is used for a multitude of purposes, including the data quality report (DQR) used for time segment vetoing, based on direct coherence of PEM channels to DARM. Supplementing coincidence analysis between the two detectors, DQR prevents GW-like noise transients from being falsely categorized as events. Thus, detector livetime can be increased by figuring out how to decouple environmental noise from DARM. Directly coupling noise, found by basic coherence, has been already addressed, but the complexity of the detector causes many noise sources to up- or down-convert. These require some more careful statistical correlation to identify, and are sometimes not well understood.

Separating noise sources out of a signal can be considered a clustering problem in a space covering different frequency bands in which noise appears. A previous LIGO SURF student has evaluated several data clustering algorithms with respect to their ability to properly sort out frequency elements of seismometer signals caused by specific earthquake events[2]. Both the k-means algorithm, which aims to make clusters with low standard deviation, and the DBSCAN algorithm, which minimizes overall inter-point distance in clusters, were evaluated using multiple methods, including the Calinsky-Harabaz index and direct comparison to earthquake times via time labeling of points, ultimately showing poor earthquake identification. A long short-term memory (LSTM) recurrent neural network (RNN) seemed to work much better, but due to small input sample size, this solution may have been be plagued by over-fitting. Thus, it is imperative that a more robust frequency clustering mechanism be designed for the PEM system.

2 Objectives

- As a primary goal, **algorithms or clustering approaches which correctly identify known noise events need to be found**. As every algorithm has inbuilt assumptions about the dataset it is applied to, the results of an algorithm performance test on labeled data will yield information about the structure of the data. The general temporal non-stationarity of the DARM noise will need to be accounted for by varying testing time windows.
- The secondary goal is to **create a clustering approach to discover previously unknown noise correlations and possibly sources**. This is where the "detector characterization tool" that this project aims to advance will be functional—revealing new noise coupling pathways will help identify ways to improve the detector sensitivity.

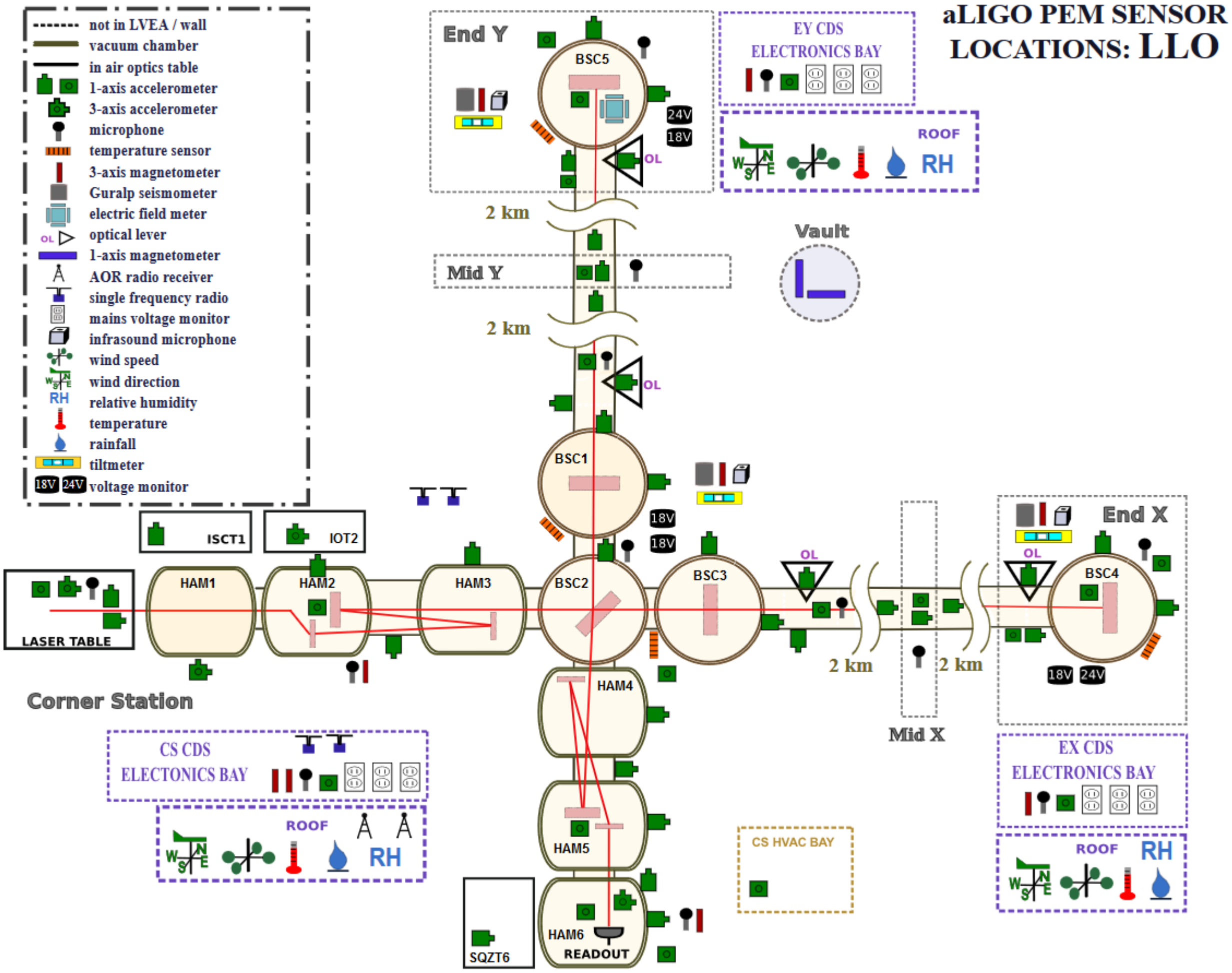

Figure 1: Schematic PEM map at the LIGO Livingston Observatory (L1). Shaded areas are in vacuum.

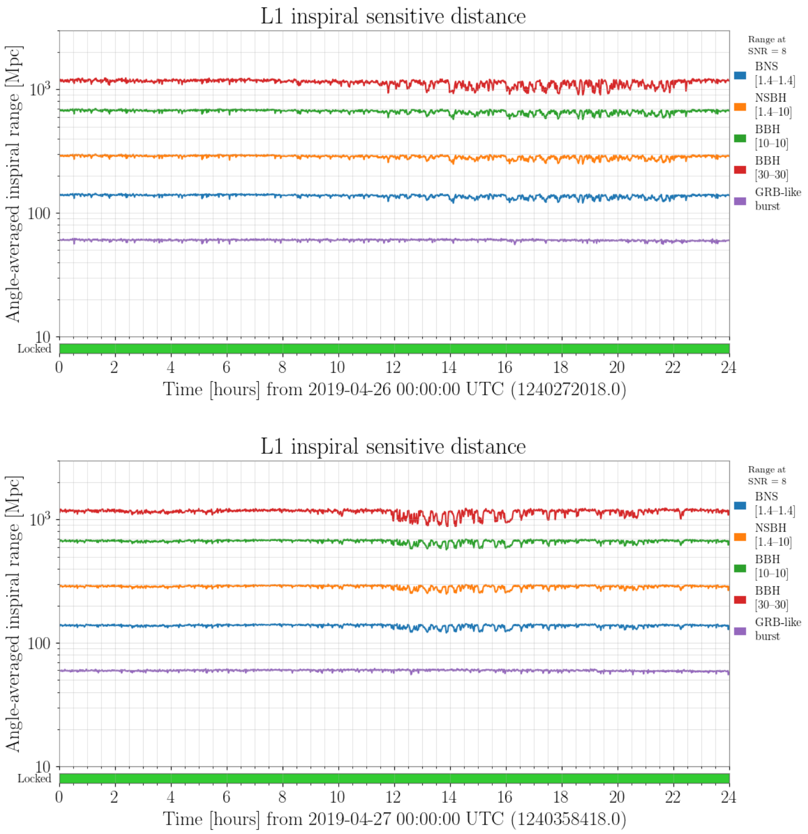

Figure 2: Detector range at L1 seems to consistently reduce during the day (∼6am-5pm CST). For large BBH in these plots, the reduction is about 300 Mpc. The source of this has been pinpointed to the Y end station, but the mechanism isn't fully clear.

3 Approach

Initially, a program will be written to take the spectral power of any PEM channel, in the form of band-limited RMS (BLRMS), likely using established methods like looping through a smoothed spectogram of the channel[3].

To reach the first objective, a modular `python` testing suite will be written to probe the structure of the multidimensional frequency-domain sensor data. This will strategically implement `scikit-learn` clustering algorithms and classifiers with different optimal regimes of function or working assumptions and evaluate them using point labeling. This will require, additionally to researching clustering or unsupervised classification algorithms, thinking of as many variables which may affect the data structure (such as looking at different time windows) and intelligently testing them. Optimizations will need to be considered so that run times are reasonable.

The program tackling the second objective will use working clustering approaches identified in the first objective to find new noise correlations. In the event that no individual algorithm or technique outperforms the rest for all types of sensory data, the final program will use the modular programming environment created for the testing suite to match techniques to the regimes that they work in. The structure of the input data as determined by the first objective, including the dimensionality probed by the extra variables, may lend itself to additional algorithms that can be used to combine the target regimes. To this end, extra algorithm research will be conducted with specific consideration of the solved structure.

4 Interim Report 2

Due to the unsupervised nature of clustering, some postprocessing of clustered data is required for the clusters to be meaningful. This was the primary focus of Weeks 5, 6, and 7. During Week 6, I took an interesting trip to CIT.

4.1 Labeling and attribute extraction

A third python script was created to characterize generated clusters. The main points of this are to find the following for each cluster:

- average length of event
- average periodicity
- channels/bands which it dominates
- channels/bands which dominate it

and make power spectra for the clustered data representative of each cluster. This is done by taking the median of the power spectra for many time intervals clustered together.

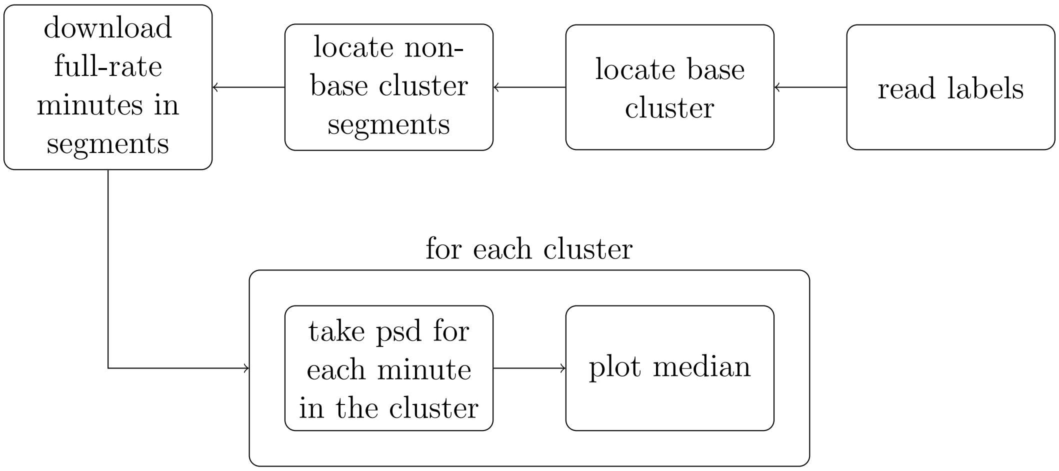
download full-rate minutes in segments
locate non-base cluster segments
locate base cluster
read labels
for each cluster
take psd for each minute in the cluster
plot median

Figure 3: Roughly the states of the "representative spectra" script. The most complicated overlooked detail in this figure is the caching of downloads. The stream writing functions used in the BLRMS-generation script have been moved and are now included from a more general location.

Taking the spectra of the clusters provides a signature for each cluster that can be programmatically validated, allowing new states to be detected without re-clustering, and also a way to easily identify frequency conversion that is happening during coupling. The script can extract other attributes that are helpful for chasing down the source and eliminating it, like the periodicity or dominating channels.

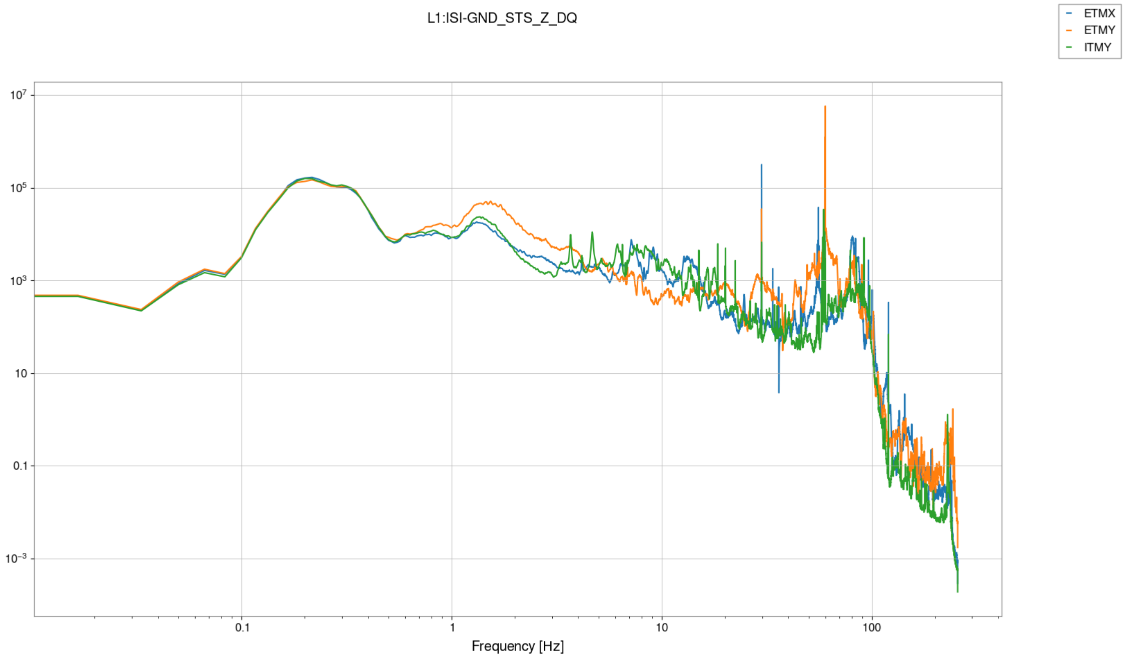

Figure 4: The representative spectrum of a cluster corresponding to train-dominated times. Notice that between 1 and 10 Hz, the seismic motion at ETMY (orange) is greater than at the other VEAs by a factor of about 10.

4.2 A technical note

Downloading and saving full-rate data takes an exorbitant amount of disk space, especially when only a portion of the frequency content is going to be used. This calls for a decimation procedure to be applied to raw downloads before they are saved.

At CIT, Rana mentioned that the default low-pass filtering options in scipy's resampling function produce significant aliasing noise (> 1%) when downsampling by a large factor. According to a test[1] done by Eric Quintero, this issue can be remedied without sacrificing runtime by using (1) a number of FIR taps proportional to the downsampling factor, rather than the default fixed value, and (2) a non-default window (`blackmanharris`).

For a full-rate time series `raw: gwpy.timeseries.TimeSeries`, the fastest and best procedure for resampling to `rate: int [Hz]` would be something like

```
raw.resample(n=20*raw.sample_rate.value/rate, window='blackmanharris')
```

4.3 BLRMS generation and future clustering

Anamaria has identified some interesting bands for the microphones and accelerometers. The BLRMS-generation script has been chugging away in a multi-threaded mode on the LDAS grid with these parameters for approximately a week.

[1] see https://git.ligo.org/NoiseCancellation/GWcleaning/issues/2

Next, these BLRMS channels will be clustered with `GDS-CALIB-STRAIN` as well as the SenseMon BNS range and the absolute value of temperature sensors in each VEA.

4.4 Other algorithm testing

I mentioned in the first interim report an incomplete test to compare the efficacy of different clustering algorithms implemented by `sklearn`. This completed, and from a first glance, the Spectral Clustering and Gaussian Mixture algorithms seemed to generalize better than k-means over different feature timescales. However, algorithm upgrades will be helpful only after the full analysis using k-means is completed, so the topic has not been revisited.

References

[1] A. Effler, R. M. S. Schofield, V. V. Frolov, G. González, K. Kawabe, J. R. Smith, J. Birch, and R. McCarthy, Classical and Quantum Gravity **32**, 035017 (2015).

[2] LIGO Document T1700198-v1

[3] aLIGO LLO Logbook entry 45374 by Gabriele Vajente