PERFORMANCE STUDY OF BOOSTED DECISION TREES IN A GRAVITATIONAL WAVE BURST SEARCH.

S. Mohapatra, S. Fischetti, L. Cadonati

University of Massachusetts, Amherst, USA
email: *satya@physics.umass.edu*

We explore the implementation of multivariate classification analysis in gravitational wave burst searches, to separate signals from background. We focus on the Boosted Decision Tree algorithm[1] and the coincidence between three interferometers (two of which co-located) as applied in the LSC S5 burst analysis with the Omega pipeline. The Boosted Decision Tree algorithm, from the ROOT TMVA package [2], is applied to bandwidth, duration, H1H2-coherent energy, H1H2-correlated energy and L1-normalized energy.

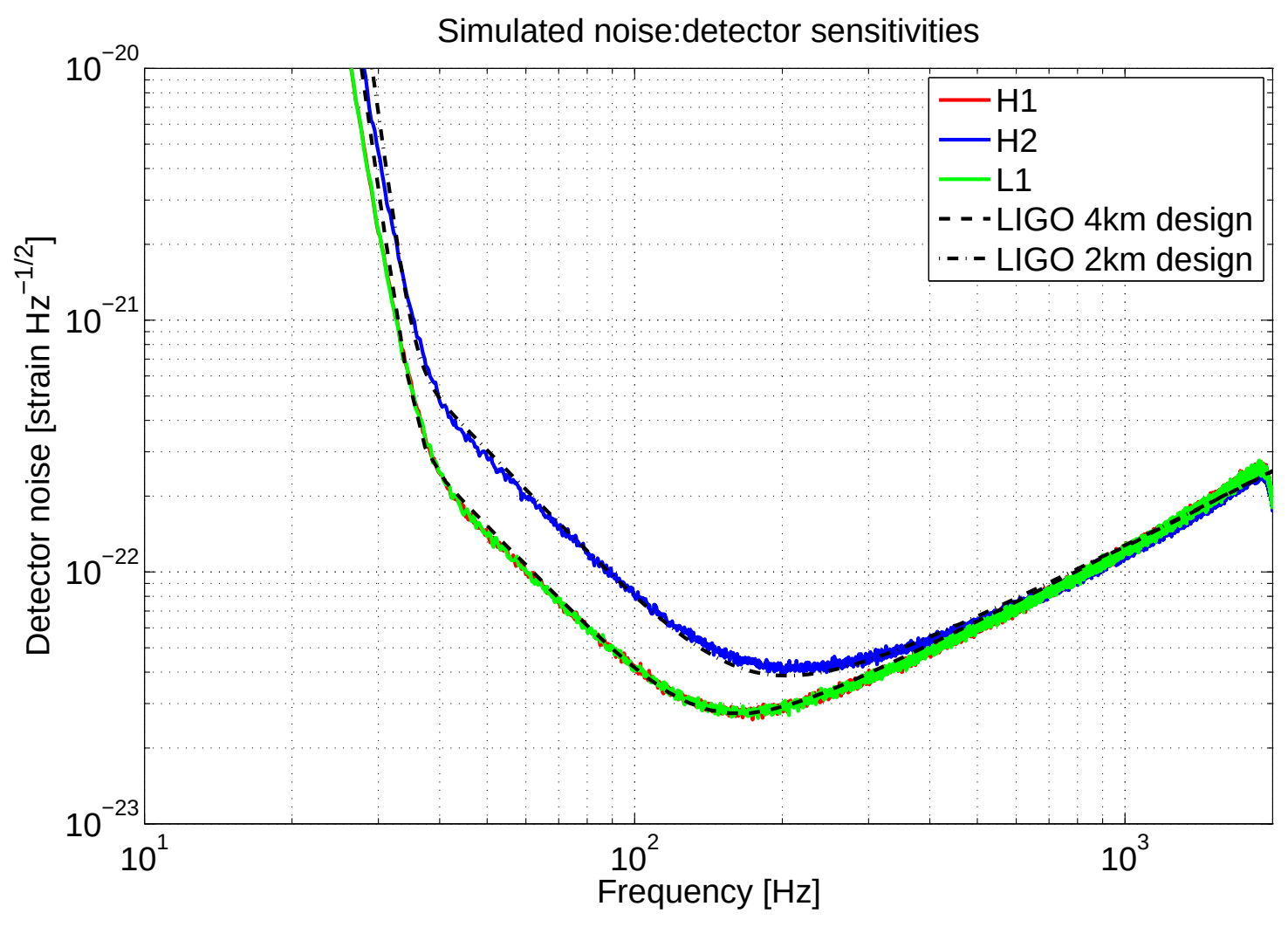
In this preliminary study, the signal sample are simulated binary black hole coalescences (EOBNR waveform [7], coded in LALApps [6], in the total mass range 100-350 $M_\odot$) and the background sample are accidental coincidences in simulated gaussian noise. We compare the signal-noise discrimination obtained by the BDT algorithm to the cuts applied in the LSC burst analysis of the first year of S5.

LIGO-DCC number: G0900565-v7

Motivation

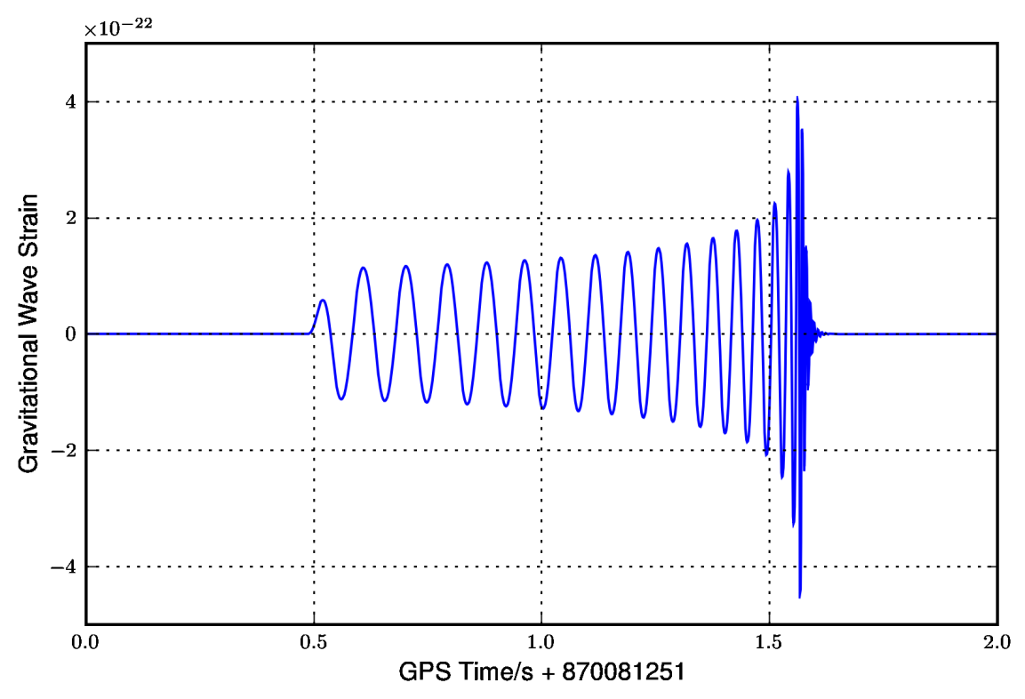
We tested alternative coincidence techniques on Gaussian noise colored to model the initial LIGO sensitivity, injected with EOBNR waveforms [7] for coalescence of BBH in mass range 100-350 $M_\odot$, which are produced with LALApps code [6].
The results are compared to the coincidence cuts imposed in the LSC S5 year 1 (S5Y1) analysis [5].
We simulated 1 week of data and aimed at a 1% false alarm probability or 10^{-8} Hz false alarm rate (FAR).

The Noise spectra



Gaussian Noise colored to model initial LIGO.

Sample simulated waveform

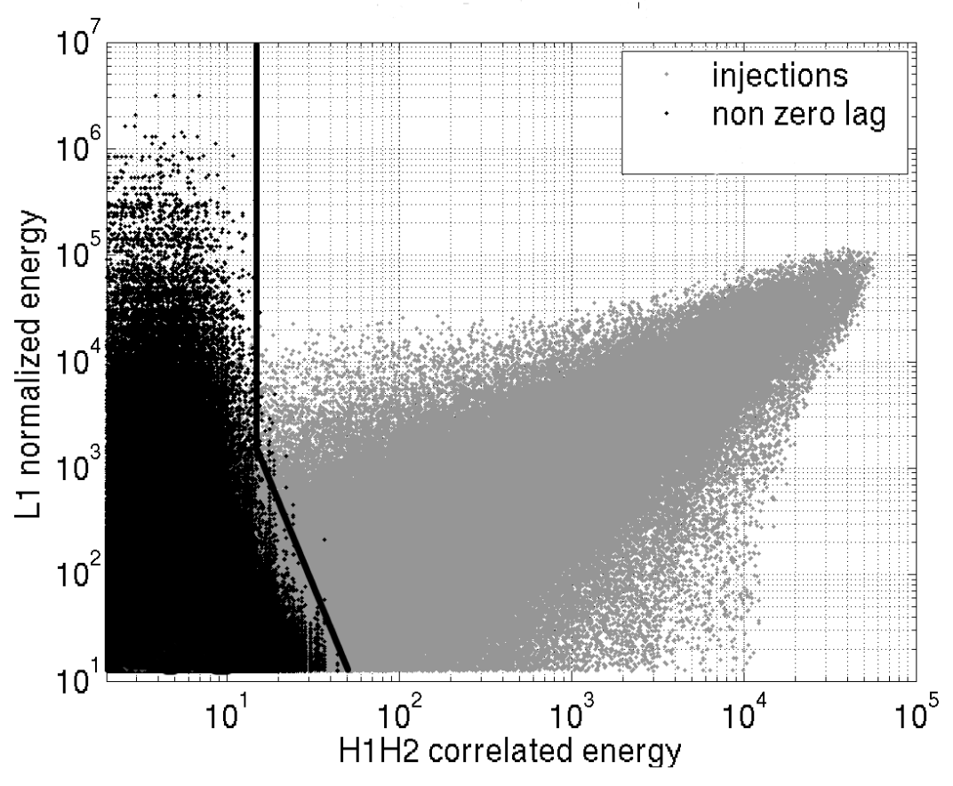


Sample of simulated waveform from the coalescence of two black holes with Mass1=82 $M_\odot$ and Mass2=71 $M_\odot$.

Notation:
LLO: LIGO Louisiana Observatory with the 4km interferometer **L1**.
LHO: LIGO Hanford Observatory with two interferometers: **H1** (4km) and **H2** (2km).

Signal/background separation in the S5 burst analysis with the Omega pipeline

- The *Omega pipeline* [4] is a multi-resolution time-frequency search for statistically significant excess signal energy, equivalent to a templated matched filter search for sinusoidal Gaussians in whitened data.
- The resulting clustered triggers indicate a time-frequency tile with excess power, whose significance is converted into a SNR, equivalent to ρ for matched filter with sine-gaussians.
- The **normalized energy Z** of a single interferometer is related to the SNR ρ by $Z = \rho^2/2$.
- The **H1H2-coherent energy** is the square of the frequency-dependent weighted sum of data from the two Hanford detectors, which maximizes the effective SNR.
- In the absence of any correlation between the Hanford detectors, the coherent data stream is characterized by the **H1H2-incoherent terms**.
- The **H1H2-correlated energy** is the difference between the coherent and the incoherent terms.
- The S5 cut on L1 energy and H1H2-correlated energy is tuned for a preset FAR, as shown in the plot from the S5Y1 burst paper [5].

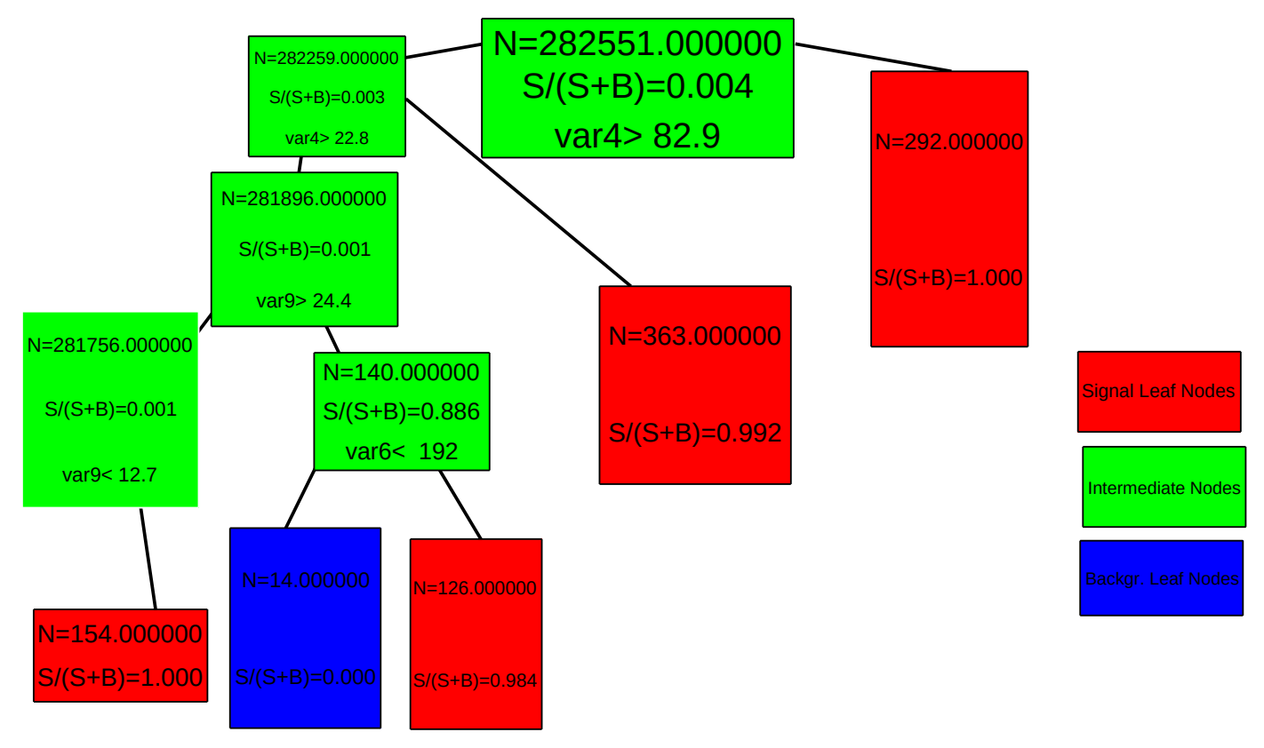


Black: accidental coincidences (background).
Grey: sine-Gaussian injections (signal).
From arXiv:0905.0020

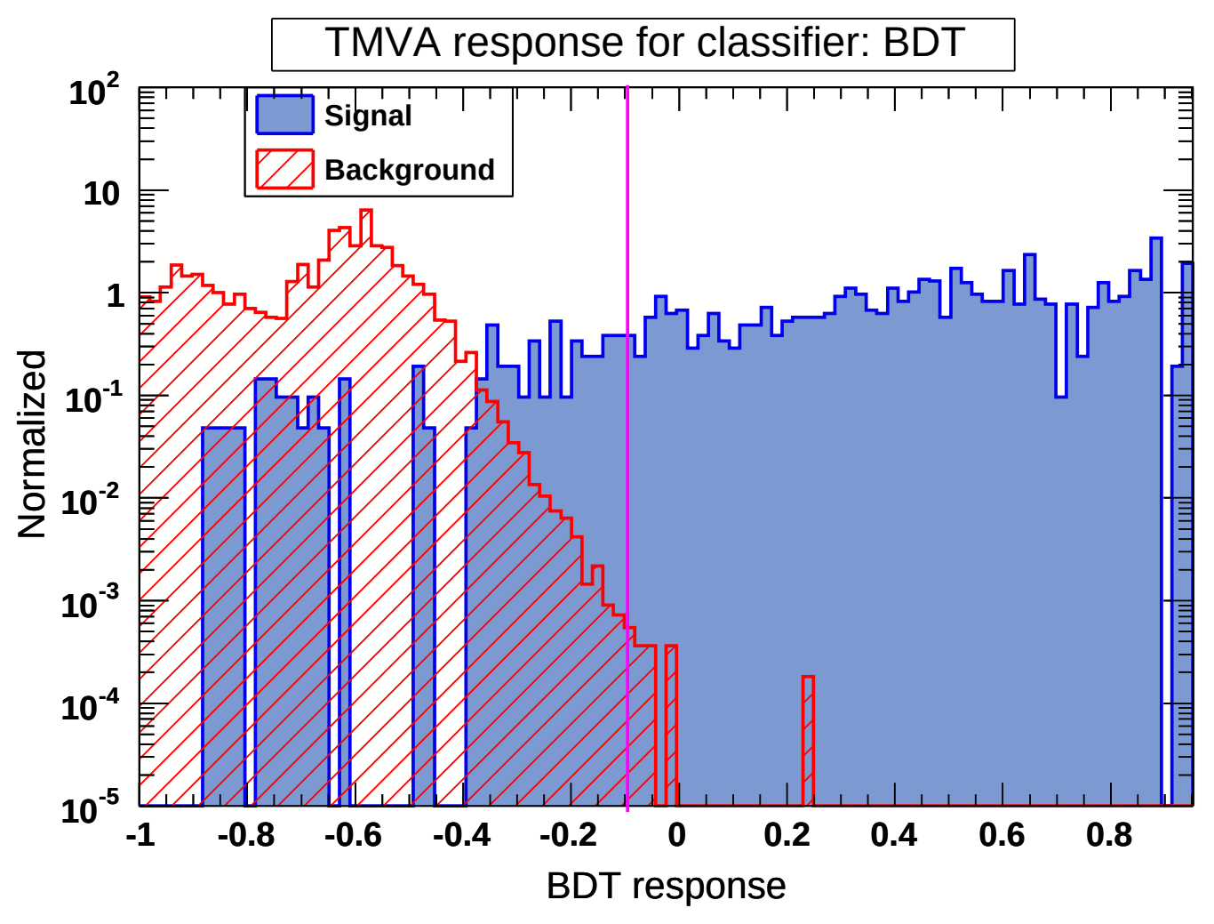
Note: The excess of background events at large L1 energy is due to non-Gaussian noise transients in the data, which are absent from our simulated, Gaussian data set.

Boosted Decision Trees (BDT)

- BDT [1] are a method for classifying an event as "signal" or "background", based on a series of "node-like" decisions (cuts) for a number of physical variables.
- The BDT response is the combined vote of many individual decision trees, derived from the same training sample by boosting (re-weighting) events.
- The BDT response can be used as a univariate discriminant to distinguish signal from background.
- A different set of signals and backgrounds are used to train and then test the BDT classifier.



Sample Decision Tree.

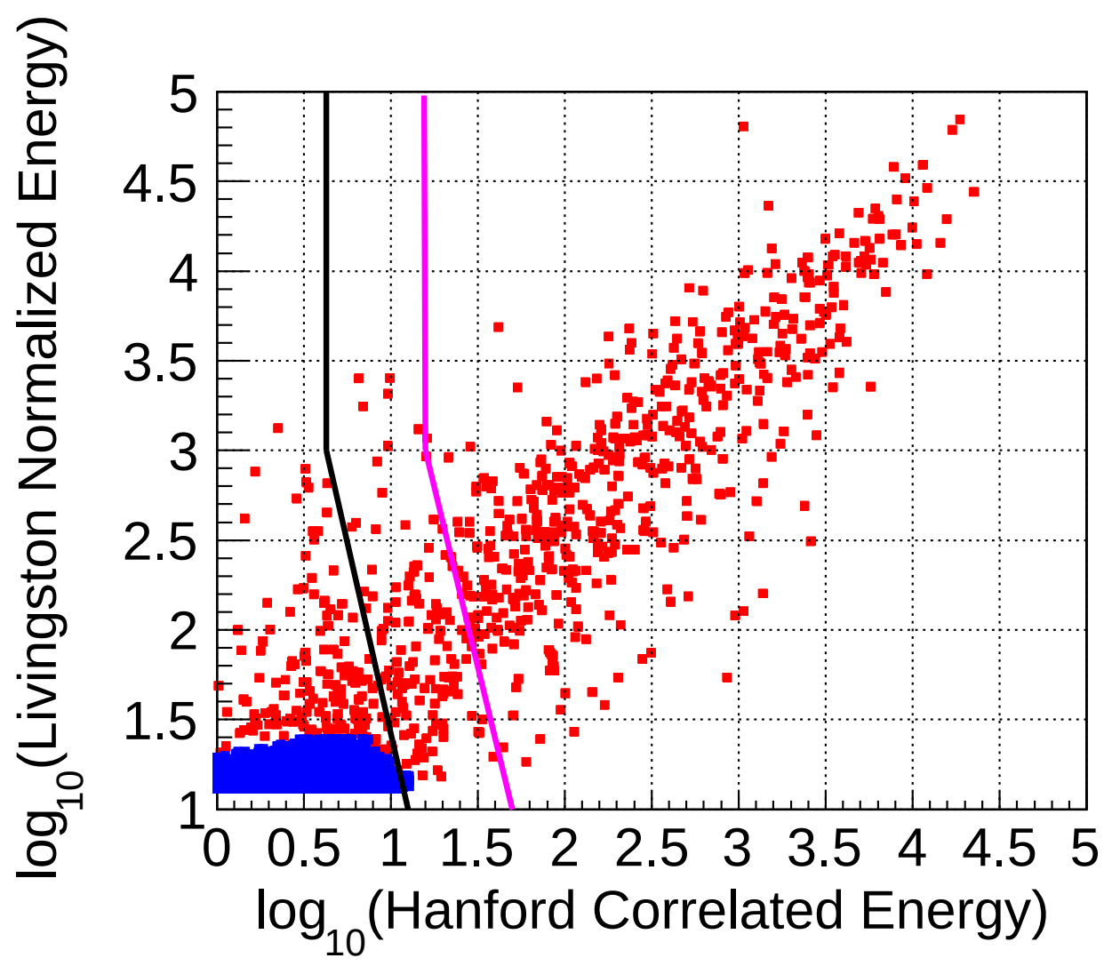


The magenta line is the BDT cut for a 1% false alarm probability (10^{-8} Hz false alarm rate).

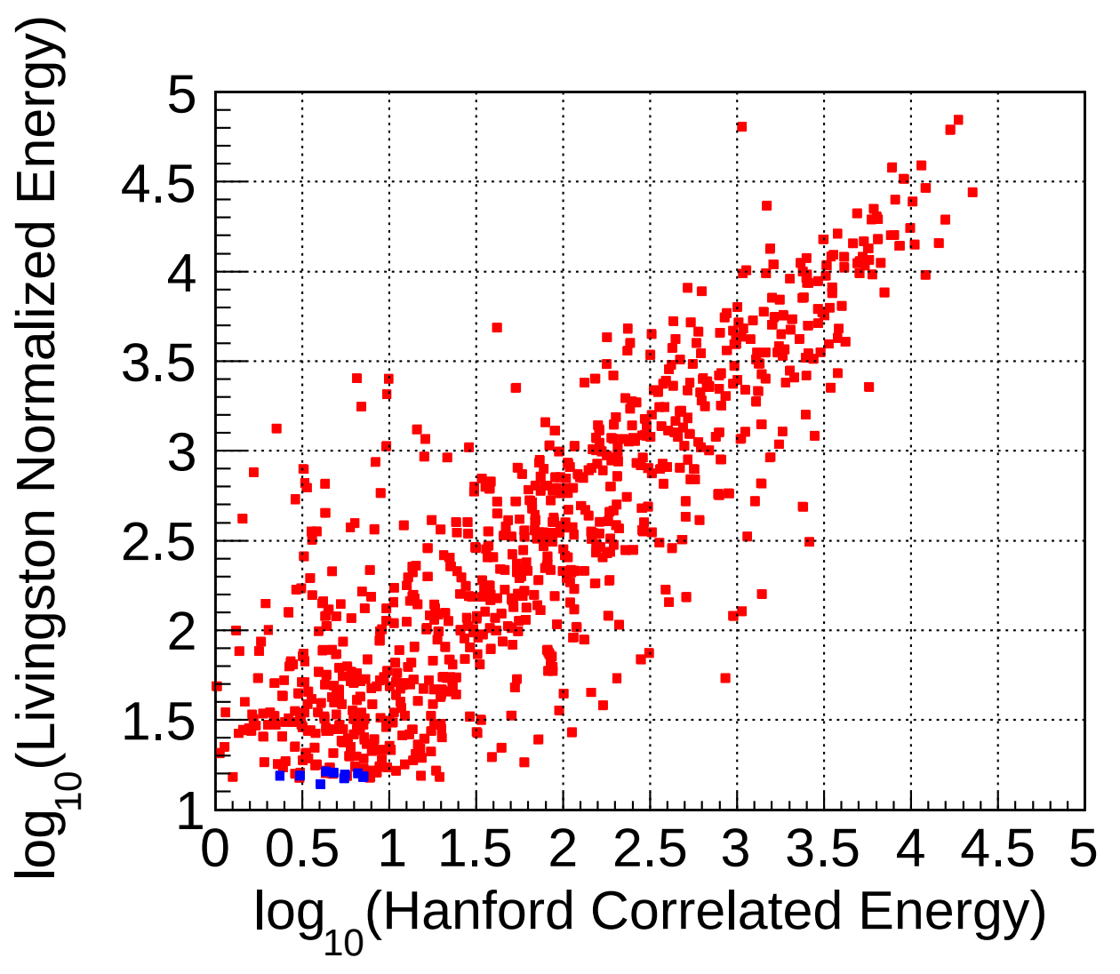
Variables used in the BDT tuning:
H1H2: frequency, duration, bandwidth , coherent energy, correlated energy
L1: frequency, duration, bandwidth, normalized energy.

Comparison of BDT cut and cut used in S5Y1

- A cut motivated from S5Y1 burst paper[5] is shown in the figure left with two variables: H1H2-correlated energy and L1-normalized energy.
- Background: 1000 time slides with the step of 0.5 seconds; noise only data.
- Signal: injections detected by the *Omega pipeline* with SNR > 5.5.
- The figure on the right shows events surviving the BDT cut.
- Both cuts leave 10 background events from 1000 time lags, for a 1% False Alarm Probability, or a 10^{-8} Hz FAR on one week of simulated data.



Cut used in S5: the magenta line is the S5 cut and the black line is similar cut tuned for a desired FAR.



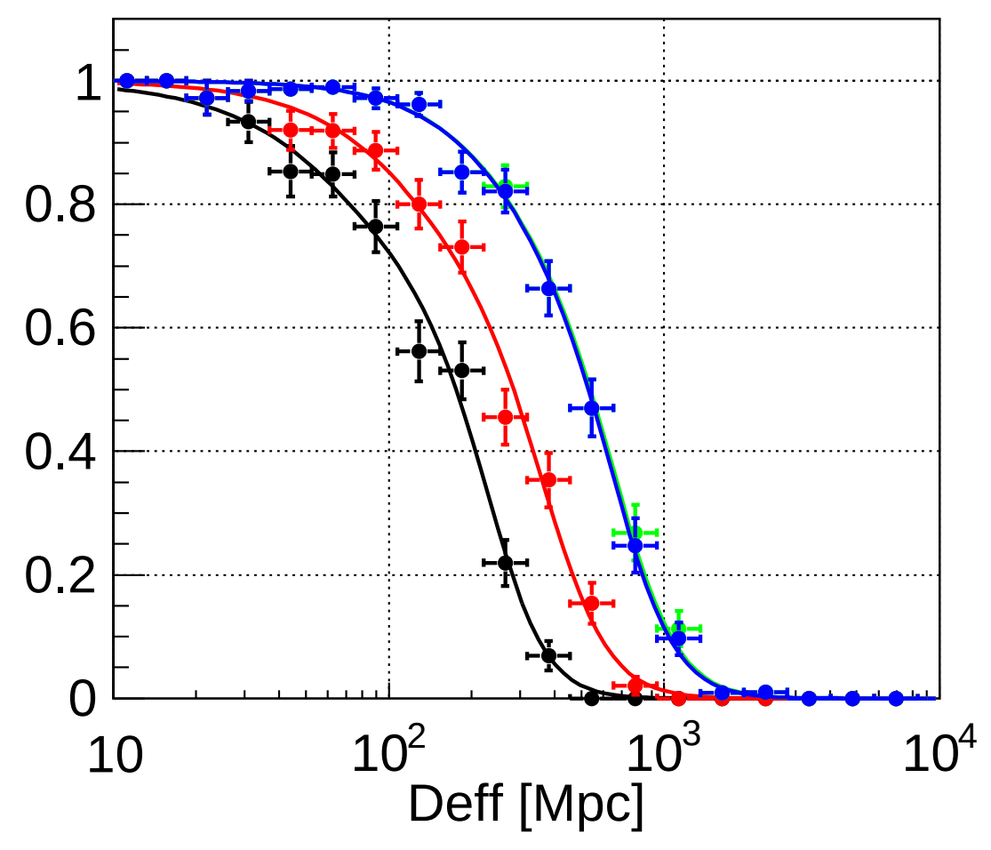
BDT cut: signal and background events surviving the cut.

Detection efficiency and FAR comparison

Comparison of detection efficiency vs effective distance:

- S5 cuts.
- Cut shaped as in S5 but shifted for a False Alarm Probability of 1%.
- BDT cut
- SNR > 5.5 for both LHO and LLO

The 50% and 90% efficiency are the distant in Mpc at which 50% and 90% of the injected signals are detected by the pipeline.



cut Methods	Efficiency 50%	Efficiency 90%	FAR in Gaussian Noise (Hz)
S5 cut + SNR > 5.5	176 ± 6	41 ± 2	0
S5 like cut tuned + SNR > 5.5	283 ± 11	76 ± 3	10^{-8}
BDT cut + SNR > 5.5	530 ± 16	178 ± 5	10^{-8}
SNR > 5.5	539 ± 19	179 ± 6	10^{-5}

The distance values are specific of the simulated noise.

The BDT cut effectively suppresses the false alarm rate while preserving detection efficiency.

Conclusion

- A multivariate classification scheme such as Boosted Decision Trees can optimize background rejection while preserving the signal detection efficiency of gravitational wave burst searches.
- Preliminary studies on actual data (inclusive of noise transients) suggest comparable performance; we are exploring a potential implementation of the BDT technique in future.

This work is supported by NSF grant PHY-0653550

References

[1] Boosted Decision Trees as an Alternative to Artificial Neural Networks for Particle Identification, arXiv physics/0408124

[2] TMVA - Toolkit for Multivariate Data Analysis, arXiv physics/0703039

[3] https://geco.phys.columbia.edu/omega/

[4] Chatterji S K 2005 *The search for gravitational-wave bursts in data from the second LIGO science run.* Ph.D. thesis MIT

[5] B. Abbott et al. *Search for gravitational-wave bursts in the first year of the fifth LIGO science run.*, arXiv:0905.0020

[6] LSC Algorithm Library, URL:http://www.lsc-group.phys.uwm.edu/lal

[7] Buonanno A. et al. 2007 *Phys. Rev.D* **76** 104049