**TITLE**

**A thesis submitted in partial fulfillment of the requirements for the degree**

**MASTER OF SCHENCE**

**in**

**DATA SCIENCE AND ANALYTICS**

**by**

**THOMAS CANNON**

**AUGUST 2020**

**at**

**THE GRADUATE SCHOOL OF THE UNIVERSITY OF CHARLESTON, SOUTH CAROLINA AT THE COLLEGE OF CHARLESTON**

**Approved by:**

**ABSTRACT**

**TITLE**

**A thesis submitted in partial fulfillment of the requirements for the degree**

**MASTER OF SCHENCE**

**in**

**DATA SCIENCE AND ANALYTICS**

**by**

**THOMAS CANNON**

**AUGUST 2020**

**at**

**THE GRADUATE SCHOOL OF THE UNIVERSITY OF CHARLESTON, SOUTH CAROLINA AT THE COLLEGE OF CHARLESTON**

We present a new method for categorizing Gamma-Ray Burst (GRB’s) emission episodes with similar light curves from the Burst and Transient Source Experiment (BATSE), onboard NASA’s Compton Gamma-Ray Observatory (CGRO). We compare normalized time-series data from any two respective GRBs’ 64ms light curves using several statistical tests. The comparisons are used in the construction of similarity matrices as input in a hierarchical clustering algorithm. With the new application of this data mining tool, we begin to see similar GRB light curves cluster together by emission properties that exist independent of their amplitude and time scales, leading to a unique understanding of GRB physics.

**ACKNOWLEDGEMENTS**

Test

**CONTENTS**

[1. Introduction 7](#_Toc45703314)

[1.1 Historical Context 7](#_Toc45703315)

[1.2 GRB Emissions 8](#_Toc45703316)

[1.3 Objectives 9](#_Toc45703317)

[1.4 Data 10](#_Toc45703318)

[1.5 Thesis Summary 11](#_Toc45703319)

[2. Similarity Matrices Descriptions 13](#_Toc45703320)

[2.1 Euclidian 13](#_Toc45703321)

[2.2 L1 Norm 13](#_Toc45703322)

[2.3 Dynamic Time Warping 13](#_Toc45703323)

[2.4 Correlation 13](#_Toc45703324)

[3. Agglomerative Hierarchical Clustering Description 14](#_Toc45703325)

[4. Methods 15](#_Toc45703326)

[5. Results 17](#_Toc45703327)

[6. Conclusion 18](#_Toc45703328)

[7. Future Work 19](#_Toc45703329)

**LIST OF FIGURES**

Test

1. Introduction

## 1.1 Historical Context

In 1963, the Partial Test Ban Treaty prompted the United States to launch the Vela satellites in order to monitor and enforce a ban on nuclear testing. The Vela satellites carried the ability to detect gamma ray radiation from nuclear explosions originating on earth, and in doing so, accidentally discovered energetic flashes of gamma radiation out in space. (Klebesadel et al., 1973). Over the next few decades, the debate over the origins of this phenomena had still not been settled, when in 1991, the Burst and Transient Source Experiment (BATSE) on board the Compton Gamma-Ray Observatory (CGRO) launched. The new data from BATSE showed an isotropic distribution of these bursts occurring anywhere around the sky at a frequency of about one event per day. This distribution indicated that the GRB events originated beyond the Milky Way (Paczynski, 1991; Meegan et al., 1992). Years later, the cosmological origin of these events were confirmed when a redshift was obtained on an event named GRB 970228 (van Paradijs et al., 1997), meaning that GRBs occur outside of our galaxy and far back in time. Models to describe the GRB prompt emissions have been proposed, but over almost half a century later, we still do not have a grasp on the mechanisms responsible for causing the prompt emission.

## 1.2 GRB Emissions

GRB emissions, while their complexities range, have a defined, non-random structure, the cause of which is still highly debated. **GET A 2X2 PIC OF BURSTS AS EXAMPLES** GRBs are made of pulsed radiation, and in recent years, have been studied thoroughly to give a greater understanding of the physics behind a GRB event. The basic units of a GRB are its pulses (Hakkila & Preece (2011); Hakkila et al. (2015, 2018)). The properties of a single pulse have been thoroughly measured (Golenetskii et al.(1983); Liang & Kargatis(1996); Norris et al.(1996); Norris(2002); Ramirez-Ruiz & Fenimore(2000)) and can be fitted by a four-parameter empirical model (Norris et al.(1996)). The Norris model is a monotonic function used for extracting the shape of a single pulse to several overlapping pulses in a time-series GRB light curve (Hakkila & Cumbee(2009); Hakkila et al. (2008); Norris et al. (2005)). Unfortunately, in a highly structured GRB emission episode, when a burst is comprised of what seems to be a large number of overlapping pulses, it becomes difficult to understand the emission structure and accurately understand the processes of the GRB event (Hakkila & Cumbee(2009)). This is because GRB pulses are actually non-monotonic (Hakkila & Preece 2014; Hakkila et al. 2015, 2018). On top of the monotonic Norris model, GBR pulses exhibit residual fluctuations in phase with the pulse structure above the background noise. The residuals most commonly appear on top of a pulse as a triple-peaked structure approximately centered around the pulse peak but can propagate more than three peaks and also exist out of phase with the pulse peak. This structure is important to note because it is difficult to understand the evolution of these peaks with respect to signal to noise (S/N). This residual structure more often occurs in bursts with a high S/N, and bursts with lower S/N will typically have this structure washed out. The washed out pulses fit the monotonic model well, while bursts with less noise have structure that requires more explanation.

Many GRBs are classified into two different categories: Long and Short **GET A HISTOGRAM OF LONG/SHORT** (Kouveliotou et al.(1993); Mukherjee et al.(1998); Hakkila et al.(2003)). These categories were selected primarily based on duration. However, recent work has shown that similar correlative pulse properties not only exist in both Long and Short bursts (Hakkila & Cumbee(2009)), but that the Long and Short bursts share common trends of these different property correlations such as duration, lag, peak flux, hardness ratio, asymmetry, and fluence (Hakkila & Preece(2011)). This suggests that the Long and Short bursts originate from similar physical processes.

## 1.3 Objectives

Astronomy has a long history of observing complex objects and events with no way to initially build a unified model to objectively explain what subjectively looks to the observer as the same phenomena. One of the most well-known examples that comes to mind is the life cycle of stars – with super giants, Sun-like stars, red dwarfs, neutron stars, etc. We began with an intuition that all of these objects are somehow related and governed by the same physics, but it was not until after centuries of classification attempts and piecemeal understanding of the parts that we developed a more cohesive model to explain all of the avenues and evolution of the stellar lifecycle – which can be generally represented in the Hertzsprung–Russell diagram. Now, with modern computing and an adequate dataset, we have means and motivation to help sort through single GRB events and attempt to explain the relationships between each GRB in order to see the population as a whole.

# Data

BATSE (Fishman(1992)) **GET A PIC OF BATSE** was an experiment on board CGRO that launched in April 1991 and operated for over 9 years. It contained 8 detectors on each of the satellite’s corners, creating an isotropic view of the gamma ray sky. When a significant change in the gamma ray background occurred in the detectors, it would begin recording an observation, counting and binning the number of photons detected from the interaction of gamma rays with the detector’s sodium-iodide based crystals. The data was collected in four different energy channels, ranging from highly energetic X-rays to gamma rays. A GRB can vary in its emission throughout each channel, and in some cases not emit above the background enough in one channel to even be noticeable. For the scope of this analysis, we are going to sum the four channels into a single time-series array.

The satellite orbited earth in an elliptical orbit, which plummeted the experiment in an out of earth’s radiation belts. The radiation belts contributed to the background noise in the detectors, the level of which can be easily seen between any two bursts. There were also other sources of background radiation that muddy the data such as solar flares, X-ray binary systems (eg. Vela X-1), and gamma ray producing black holes (eg. Cyg X-1). Besides background noise sources and a weak detection in one of the four energy channels, S/N can change for other reasons such as an occultation of the source or a failure in one of the detectors or energy bands.

The data used for this study is archived from the life of the experiment. The summed four channel data has a time resolution of both 256ms and 64ms per burst. The BATSE experiment collected data at 256ms until a specified trigger criterion was met. Once met, BATSE would then record the remainder of the burst in 64ms time resolution. Because of this, some GRB samples in our working dataset will have both of the time resolutions along their light curves.

The data is freely available to download as ASCII files. The files contain a few lines of meta-data with information such as the count of bins of the burst and the total number bins since the trigger time followed by four tall columns of count data with each column representing an energy channel and each row representing a time bin. We sum each column together across the rows in order to produce the combined-four channel data, represented as one column of photon counts per 64ms time bin.

We will also be using the duration table from the BATSE 4B Catalog (Paciesas et al. 1999). This table contains the T90 times, which is defined as the time in which the middle 90% of the flux if the burst is observed. We use these time frames to help put boundaries on the emission episodes for use in preprocessing and normalization.

## 1.5 Thesis Summary

As mentioned above, since GRB emission episodes – which are comprised of pulses – have correlated properties to their duration, fluence, and spectral properties, we have an argument supporting a normalized comparison of the time-series emission data of every emission to every other emission with the intent of uncovering classifications of bursts that are clustered to one another. Despite the data being normalized, there are still biases we will potentially be carrying over into the analysis from the raw data. These biases will be mitigated through the steps of the clustering process. Three of the largest steps are data preprocessing, building an adequate similarity matrix, and choosing the proper clustering routine. Several different methods between preprocessing and building matrices were attempted, leading to multiple pipelines to draw results from. Building similarity matrices from time-series data of different lengths is an area of active research where novel ideas are being tested. Therefore, the definitions of several different ways to build a similarity matrix will be given special attention in section 2. In section 3, we define agglomerative clustering. Section 4 will describe how the steps of preprocessing the data and the application of the defined methods for producing the similarity matrices and clustering. Section 5 will discuss the biases, strengths, and weaknesses for select permutations of the pipelines from preprocessing to cluster results. It will also discuss the results themselves and what it means for GRB physics.

1. Similarity Measures for Time-Series Data

Binned Time-series data is a sequence of real numbers representing the total counts of an event per a given increment of time. We consider a similarity measure as a resemblance value that is calculated between any two vectors and exists outside the influence of any other vectors. A good survey for similarity measures of time-series data was written by (Liao (2005)), and an application and comments on some of these techniques described by (Igleisas (2013)).

In order to construct a similarity matrix, we need to generate a value of resemblance between two vectors of every GRB to every other GRB, creating an upper triangular matrix of values that are meant to represent how similar any two emission episodes are to each other. Given any two time-series sets of GRB data, and we present the following methods of calculating a similarity value.

## 2.1 Euclidean Distance

Euclidean distance has been used in time-series matching and similarity distances for years (Faloutsos et al. (1994)). For our purposes, we need to assess the Euclidean distance measure between the differences in the and vectors to create our Euclidean similarity measure . Thus, it is represented by

It is important to note that this measure will only work on vectors of equal length, which can be solved by resampling the binned GRB data. Euclidean distance is also invariant to time dependent features of a vector. As in, a novel structure in one emission episode can also appear in a different emission episode, but if these two structures are out of phase, the Euclidean distance will not be able to see it. It is blind to feature correlation unless the two features are in phase. Furthermore, while this metric is a good representation of how similar any two vectors are, it is not normalized between every pair of vectors, meaning that unless the values in the vectors are all on the same scale and the vectors themselves are all of a similar length, then the similarity values between each pair – even if the pairs are normalized to each other in time and scale – are not comparable.

## 2.2 Zero-Normalized Cross-Correlation

As stated above, Euclidean distance is blind to the correlation between features. One potential way to mitigate that would be to line up two GRB vectors on their most prominent features using a standard cross correlation. While this method works well describing the correlation between any to vectors, it has the same problem as the Euclidean distance measure where, when working with a population of similarity measures between many vectors, the measures are not on a standard scale to make the pairs comparable. A Zero-Normalized Cross-Correlation (ZNCC) (Lewis (1994); Yoo (2009)) does not have this problem. ZNCC is widely used in image processing and is used to normalize and measure the similarities between two images of different exposures. We can retool this method to work for one-dimensional vectors as well.

Assuming vectors of equal length, we select the max value from ZNCC as the distance measure, giving

ZNCC attempts to fix the problem of blindness to feature correlation that the Euclidean method has as well as the problem or normalization of similarity measures between multiple pairs of vectors. While the normalization is fixed, the shifting around of vectors in search of the max cross-correlation could potentially yield chaotic results.

## 2.3 Dynamic Time Warping

Dynamic Time Warping (DTW) (Berndt & Clifford (1994)) has been used to generate similarity metrics (Keogh (2002)), which have been used in clustering and classification (Łuczak M (2016)). It does not need to be given vectors of the same length and is created to spot feature correlation. It is widely used in voice recognition to help machines match voice to words where two individuals are speaking the same phrase at different cadences. It finds the best alignment between two sets of the peaks and valleys of speech data to determine their similarity.

DTW allows a non-linear mapping of two vectors by minimizing the distance between them for vectors of lengths that are the same or different, where and . **INSERT DTW GRAPHIC** DTW builds an -by- cost matrix that contains the distances between two points and . Then a warping path is formed by a set of matrix components, where . In addition, the warping path should satisfy three local constraints:

1. Endpoint constraint: and
2. Monotonicity constraint: if and

then and ;

1. Continuity constraint: if and

then and .

There are many warping paths that satisfy these conditions. The warping path that minimizes the warping cost is considered the DTW distance and is what we use as our similarity measure

## 2.4 Regularized Manhattan

Stuff about norm

1. Agglomerative Hierarchical Clustering Description

An Agglomerative Hierarchical Clustering algorithm works from the ``bottom up"; meaning it begins with every element in its own cluster. As described by Jain, et. al. \cite[]{jai99}, and modified to represent GRB light curves, the algorithm proceeds as follows:

\item With a matrix of comparative 'similarity distances' between every GRB to every other GRB, each distance is initially treated as its own cluster.

\item The two clusters considered the most similar out of the entire matrix are then merged into one cluster. The matrix is then updated to reflect the merger by considering new position values that represent the combined cluster.

\item If every GRB is within the same cluster, then stop. Otherwise, return to step $2$.

Fortunately, there are several programming languages that will take matrix inputs and return clustering results. In our case, we used an IDL routine called CLUSTER\\_TREE.\footnote{ For the IDL clustering routine used see \url{http://www.exelisvis.com/docs/cluster\\_tree.html}}\footnote{For the analogous MATLAB routine see \url{http://www.mathworks.com/help/stats/linkage.html}}\footnote{For the analogous Python Scipy routine see \url{http://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html}}

1. Methods

Before the 64ms light curve data could be evaluated with the two-sampled $\chi^2$ test, it needed to appropriately prepared. Under every BATSE GRB detection there is considerable background of gamma-ray flux. This background is often changing due to the CGRO's orbit in and out of areas of higher and lower gamma-ray radiation around earth. This sometimes gives the background a slope, which needed to be subtracted out. In order to subtract out the background, we begin by defining the area of the background as everywhere outside the middle $90\%$ of the GRBs' fluence. The times representing this bounded fluence are called t90 times, and were selected for convenience from a table of published values\footnote{see \url{http://www.batse.msfc.nasa.gov/batse/grb/catalog/current/tables/duration\\_table.txt}}. An additional $10\%$ of the t90 times were added to either end to ensure that the total fluence of the GRB was accounted for. This area, which represents the burst, was then removed from the GRB light curve to leave only recorded background data of the times surrounding the GRB event. This background was then linearly fit, and the slope of the fit was then subtracted from the light curve, leaving a flattened background.

The burst event regions were then also selected by their t90 times. This time, there was no additional time added because we operated under the assumption that similar bursts would have similar t90 times and therefore still evaluate well under the $\chi^2$ test. These regions were then scaled to a common fiducial time scale.

It is from here that we proceed with gathering 'similarity distances' in the form of $\chi^2$ values. With these values in a single document, we run them through the clustering algorithm to produce the relationships of every GRB to every other GRB.

1. Results

With the clustering completed, we move the data into a dendrogram for easy visualization. With our large dataset, the dendrogram becomes too large to visualize on a single page. Therefore, as seen in Fig. 1, we sample a part of the dendrogram for preliminary results discussion. The dendrogram visualizes different GRB light curves as leaves in branches. Each branch represents a cluster of GRBs that have similar characteristics as seen in Fig. 2. The vertical length of a branch is a measure of how similar a GRB or cluster of GRBs is to its connecting GRB or cluster of GRBs. Sometimes, the tree will display several adjacent GRBs in a tight cluster that show evidence for an evolving continuum of GRB properties as seen in Fig. 3. Also, There are some GRBs that are almost completely unique. These are displayed as branches on the tree whose nodes, common to the rest of the tree, break off very high up. In the algorithm, these GRBs would have been selected last as a comparable relative to any other GRB or cluster of GRBs. Curiously, these GRBs also have an incredibly high signal to noise.

1. Conclusion

Hakkila et al. (2018) asks, “As signal-to-noise decreases, does structure disappear before or after the smoothly-varying remainder of the pulse disappears?”

We are still only beginning to discover what this data mining technique is teaching us about GRB light curves. Currently, this technique allows us to compare GRB light curves in a new and interesting way. By using a two-sampled $\chi^2$ test results as `similarity distances' in a Agglomerative Hierarchical Clustering Algorithm, we are able to achieve our comparisons. This technique seems to be good at finding GRBs that evolve similarly and possibly have similar progenitors. However, with a large amount of new data to investigate, we are still determining the meaning of the results of the analysis. We are also working on ways to improve and speed up the computationally expensive process.

1. Future Work

While the method is sound the data preparation could be improved upon. A better definition of the burst start and end times would work nicely. Also, a more rigorous analysis needs to be done with the results. We have developed a new tool with which to explore GRB physics. However, in the larger scheme of things, this method does not have to be used simply for GRB light curves. It is easily adaptable to any transient or two dimensional data.

**REFERENCES**

BATSE data <http://www.batse.msfc.nasa.gov/batse/grb/catalog/current/>

Berndt, D. & Clifford,J., Using dynamic time warping to find patterns in time series, in: KDD Workshop, Seattle, vol. 10, 1994, pp. 359–370

Faloutsos et al., Fast subsequence matching in time-series databases, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, 1994, pp. 419–429.

[Fishman(1992)]{1992grbo.book..265F} Fishman, G.~J.\ 1992, Gamma-ray Bursts - Observations, Analyses and Theories, 265

[Golenetskii et al.(1983)]{gol83} Golenetskii, S. V., Mazets, E. P., Aptekar, R. L., \& Ilinskii, V. N. 1983, Nature, 306, 451

[Hakkila et al.(2003)]{hak03}Hakkila, J., Giblin, T. W., Roiger, R. J., Haglin, D. J., Paciesas, W. S., \& Meegan, C. A. 2003, ApJ, 582, 320

[Hakkila et al.(2008)]{hak08}Hakkila, J., et al. 2008, ApJ, 677, L81

[Hakkila \& Cumbee(2009)]{hak09}Hakkila, J., \& Cumbee, R. S. 2009, in AIP Proc. 1133 (ed. Meegan, Gehrels, \& Kouveliotou), 379

[Hakkila \& Preece(2011)]{hak11} Hakkila, J., \& Preece, R. 2011, ApJ, 740, 104

Igleisas, F., Energies 2013, 6, 579-597; doi:10.3390/en6020579

[Jain et. al (1999)]{jai99}Jain, A.K. Murty, M.N. and Flynn, P.J. 1999, ``Data Clustering: A Survey", ACM Computing Surveys, 31, 3

Kate, R., Using dynamic time warping distances as features for improved time series classification, Data Min. Knowl. Discov. 2 (2016) 283–312.

Keogh, E. Exact Indexing of Dynamic Time Warping. In Proceedings of the 28th International Conference on Very Large Data Bases, Hong Kong, China, 20–23 August 2002; pp. 406–417.

Klebesadel, R. W., Strong, I. B., & Olson, R. A. 1973, ApJ, 182, L85

[Kouveliotou et al.(1993)]{kou93}Kouveliotou, C., Meegan, C. A., Fishman, G. J., Bhat, N. P., Briggs, M. S., Koshut, T. M., Paciesas, W. S., \& Pendleton, G. N. 1993, ApJ, 413, L101

Lewis J. Fast normalized cross-correlation. Vision Interface. 1995;10:120–123.

[Liang \& Kargatis(1996)]{lai96}Liang, E., \& Kargatis, V. 1996, Nature, 381, 49

Liao, T.W. Clustering of time series data—A survey. Pattern Recognit. 2005, 38, 1857–1874.

Łuczak M (2016) Hierarchical clustering of time series data with parametric derivative dynamic time warping. Expert Syst Appl 62:116–130

Meegan, C. A., Fishman, G. J., Wilson, R. B., Horack, J. M., Brock, M. N., Paciesas, W. S., Pendleton, G. N., & Kouveliotou, C. 1992, Nature, 355, 143

[Mukherjee et al.(1998)]{muk98}Mukherjee S., et al. 1998, ApJ, 508, 314

[Norris et al.(1996)]{nor96}Norris, J. P., Nemiroff, R. J., Bonnell, J. T., Scargle, J. D., Kouveliotou, C., Paciesas, W. S., Meegan, C. A., \& Fishman, G. J. 1996, ApJ, 459, 393

[Norris(2002)]{nor02}Norris, J. P. 2002, ApJ, 579, 386

[Norris et al.(2005)]{nor05}Norris, J. P., Bonnell, J. T., Kazanas, D., Scargle, J. D., Hakkila, J., \& Giblin, T. W. 2005, ApJ, 627, 324

Paciesas et al. THE ASTROPHYSICAL JOURNAL SUPPLEMENT SERIES, 122:465–495, 1999 June

Paczynski, B. 1991, Acta Astron., 41, 257

[Ramirez-Ruiz \& Fenimore(2000)]{ram00}Ramirez-Ruiz, E., \& Fenimore, E. E. 2000, ApJ, 539, 712

Rodgers, J.L.; Nicewander, W.A. Thirteen ways to look at the correlation coefficient. Am. Stat. 1988, 42, 59–66.

van Paradijs, J., et al. 1997, Nature, 386, 686

Wang, Xiao & Yu, Fusheng & Pedrycz, Witold & Wang, Jiayin. (2019). Hierarchical clustering of unequal-length time series with area-based shape distance. Soft Computing. 23. 6331- 6343. 10.1007/s00500-018-3287-6.

Yoo J-C, Han T. Fast Normalized Cross-Correlation. Circuits, Systems, and Signal Processing. 2009;28(6):819–843.

**FIGURES**

Test

**APPENDICES**

Test