6-7-2024 10:30 AM

The Application of Elastic Distance in Astrophysical Time Series

Xiyang Zhang, *Western University*

Supervisor: Yu, Hao, *The University of Western Ontario*
Co-Supervisor: Kulperger, Reg J., *The University of Western Ontario*
Co-Supervisor: Valluri, Sree R., *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Statistics and Actuarial Sciences

Abstract

Elastic distances, e.g. dynamic time warping (DTW), evaluate the similarity between query and reference sequences by dynamic programming. The 1-Nearest-Neighbor predictor with DTW is one benchmark in time series classification. However, DTW is less efficient in astronomical time series because of ignorance of the information in time stamps and its dependence on the shape and magnitude between query and reference sequences. We apply two elastic distances which integrate the information in the time domain, time warp editing distance (TWED) and Skorohod distance, which is calculated by using Fréchet distance, to three astronomical datasets to compare with DTW and Euclidean distance. The first dataset aims to classify signals emitted from various astronomical sources with multiple bandpasses from the Large Synoptic Survey Telescope (LSST). The TWED shows the optimal 1-NN classification performance with a 0.74 loose accuracy. In the second dataset, we explore the possibility of shrinking the size of the gravitational wave (GW) template banks to reduce the computational waste of matching data with similar templates. With the threshold of 5%, similar templates can be removed without losing the effectualness of the template bank. In the final dataset from LIGO and Virgo detections, we establish an early warning process of GW by locating the coincident period between detectors. Though DTW distance outperforms others, TWED achieves a 0.72 detection ratio and a 0.47 average significant ratio. These results of three astrophysical datasets reveal the applicability of elastic distances in the astrophysical domain, especially TWED.

Keywords

Time series analysis, Dynamic Time Warping, LVK Collaboration, Gravitational Waves

Summary for Lay Audience

Elastic distances, e.g. dynamic time warping (DTW), evaluate the similarity between query and reference sequences by dynamic programming. Unlike Euclidean distance, elastic distances allow misalignment between indices. With minimal matching by misalignment, the traditional DTW distance embedded in the 1-Nearest Neighbor predictor (1-NN) is one benchmark used in time series classification. But DTW is less efficient in astrophysical time series because DTW distance focuses only on shape and magnitude between sequences and ignores the information in distortion in time stamps. To overcome this, in this thesis, we introduce two elastic distances which integrate the misalignment in the time domain: time warp editing distance (TWED) and Skorohod distance, which is calculated by using Fréchet distance. Then we apply two elastic distances in three astrophysical datasets and compare them with Euclidean and DTW distances, to explore the applicability of elastic distances in the astrophysical field. The first dataset aims to classify signals emitted from various astronomical sources with multiple bandpasses from the Large Synoptic Survey Telescope (LSST). The TWED shows the optimal 1-NN classification performance with the highest accuracy. We also have experiments on gravitational wave datasets. In 2015, LIGO and Virgo Collaborations generated multiple template banks and detected the first GW signal by matched filtering. In the second dataset, we simulate GW from the merger of binary black holes and construct a toy template bank. By evaluating the elastic distance between templates, we shrink the size of the template bank to reduce the computational waste of matching data with similar templates. With the threshold of 5%, similar templates can be removed without losing the effectualness of the template bank. In the final dataset with raw data from GW detections, we establish an early warning process of GW by locating the coincident period between detectors, where coincidence is measured by elastic distance. Though DTW distance outperforms others, TWED achieves a 0.72 detection ratio and a 0.47 average significant ratio. These results reveal the applicability of elastic distances in the astrophysical domain, especially TWED.

Acknowledgements

I would like to express my deepest gratitude to Prof. Yu, Prof. Kulperger and Prof. Valluri, for their great tolerance of my delay and patience when helping me solve the problems in my thesis. To those who helped me a lot in my academic career, Dr. Chishtie, Dr. Dergachev and Prof. Mohanty thank you for your altruistic guidance and suggestions and to a graduate student. I am also grateful to my parents who supported me financially and to Hanyu Chen who supported me mentally to finish my Master's degree.

This research has made use of data or software obtained from the Gravitational Wave Open Science Center (gwosc.org), a service of the LIGO Scientific Collaboration, the Virgo Collaboration, and KAGRA. This material is based upon work supported by NSF's LIGO Laboratory which is a major facility fully funded by the National Science Foundation, as well as the Science and Technology Facilities Council (STFC) of the United Kingdom, the Max-Planck-Society (MPS), and the State of Niedersachsen/Germany for support of the construction of Advanced LIGO and construction and operation of the GEO600 detector. Additional support for Advanced LIGO was provided by the Australian Research Council. Virgo is funded, through the European Gravitational Observatory (EGO), by the French Centre National de Recherche Scientifique (CNRS), the Italian Istituto Nazionale di Fisica Nucleare (INFN) and the Dutch Nikhef, with contributions by institutions from Belgium, Germany, Greece, Hungary, Ireland, Japan, Monaco, Poland, Portugal, Spain. KAGRA is supported by Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan Society for the Promotion of Science (JSPS) in Japan; National Research Foundation (NRF) and Ministry of Science and ICT (MSIT) in Korea; Academia Sinica (AS) and National Science and Technology Council (NSTC) in Taiwan.

Table of Contents

List of Tables

List of Figures

List of Appendix

Chapter 1

1 Introduction

To modelling the similarity between two time series is widely applied in speech recognition, neural science and other areas of research. The dynamic time warping (DTW) [50], which measures the similarity between two time series by recursively computing the nearest matching point between two time series, is well-known for its robustness in dealing with a time offset. It is applied to recognize human actions [52], optimize the retrieval of similar time series [60], and reconstruct gravitational wave core-collapse supernova signals [56]. Inspired by DTW, a few more algorithms are developed following the dynamic programming approach in comparing similarities between two series. Modifications are applied to traditional DTW, such as sparse time series [44], uniform scaling with DTW [30], EventDTW [35], FastDTW [51], GDTW [25], Time-weighted DTW [34], Soft-DTW [22], shapeDTW [61], DDTW [36] and so on. A discussion on several myths about DTW [49] can be taken into consideration when constructing a new algorithm. In general, there are several desired properties of the elastic distance in general when constructing a novel distance from the known [32].

- Ability to compare time series of different lengths.
- Being robust to spikes, dropouts, wandering baseline and missing values, and other issues that are common outside of benchmark datasets.
- The invariances in amplitude and offset offered by DTW and Euclidean distance, as well as additional invariances, including phase invariance, order invariance, liner trend invariance and stutter invariance.
- Ability to be computed very efficiently, allowing great scalability.

When dealing with astrophysical time series, we have to modify or find the optimal elastic

distance, in dealing with common issues from the datasets, which is characterized as the unevenly spaced time series with missing value, e.g. distortion in amplitude and phase of red-shift when light travels [38] and glitches contaminate the data with gravitational wave signals of interest [21]. Usually, combined with prior knowledge of physics and problem-oriented mathematical derivation, most abnormal or mismatches in data can be explained. However, with a giant amount of data collected from observatories and detectors, [1, 20], it is hard to distinguish useful information when there is no prior knowledge related.

One popular astrophysical dataset comes from the LIGO-Virgo-KAGRA (LVK) Collaboration with on-going continuous detection operations [6]. After 30 years of preparation in both theoretically and instrumentally, they made the first detection of the gravitational waves (GW) event [13] on September 14, 2015, with the event name GW150914. Predicted by Einstein's theory of General Relativity, gravitational waves are waves generated by the motion of massive accelerating objects, such as colliding black holes and neutron stars [42]. Considered as the ripples in the spacetime curvature by analogy with that of water waves, gravitational waves travel through the cosmos at the speed of light, stretching in one direction while compressing in a perpendicular direction when passing. By measuring the stretching in spacetime, multiple laser interferometer gravitational wave detectors were built during the last few decades. These detectors measure the changes in the light-travel time between photons inside two arms of detectors, led by LIGO and Virgo [28]. These detectors which are located far from each other, alleviate the false alarms of non-gravitational wave oscillation caused by earthquakes or other local vibrations by searching only for similar signals. Besides there are at least two detectors available and with proper functioning, e.g. LIGO Hanford and LIGO Livingston in USA, and Virgo near Pisa, Italy. It is possible that we can find the 'coincident' signals with similar waveforms from the same source by excluding those unlike episodes or similar signals with too large time-shift [28].

In this thesis, we study the characteristics of the astrophysical datasets and the requirements

of ideal elastic distance. We apply DTW, TWED [41] and Skorohod distance [53] in our application to several datasets to explore the potential application of elastic distance in an astrophysical dataset without the preliminary knowledge of templates. In Chapter 2, we discuss the collection mechanism of astrophysical time series data and several features of the data. After that, we introduce the DTW, TWED, Skorohod distance and Fréchet distance in detail in the following Chapter 3, with comparison between each other both theoretically and numerically. All of our experiments are described in Chapter 4 which includes three applications: classifying signals from the Large Synoptic Survey Telescope (LSST) [20] in Section 4.1.2, shrinking the size of template bank containing waveform from binary black holes simulated by Python module PyCBC [58] in Section 4.2.1, and early warning of the possible detection of event data from Gravitational Wave Transient Catalog - 1 (GWTC-1) [27] in Section 4.2.2. In each section, we compares the pros and cons of the application and the proposals to improve the performance. The conclusions are presented in Chapter 5.

Chapter 2

2 Features of Astrophysical Data

This chapter introduces a few typical features in astronomical time series data sets based on two major surveys of detection, photometry and spectroscopy. Following the observation mechanisms of both approaches, a general time series which emphasizes the timestamp of each observed value is defined in a convenient framework for further discussion on the incorporation of temporal domain information. We formulate the typical features with visualization on both simulated and real-time series.

2.1 Astronomical Survey Mechanisms

Spectroscopy and photometry are two important techniques in measuring the radiative energy from astronomical objects, e.g., light emitted from stars. Both techniques employ charged-couple detectors to record the signal passing through the pre-processing instrument which is the major difference between them.

The astronomical survey based on spectroscopy applies the refractive properties of light by using a prism or grating to separate a beam of light into a rainbow of colours [57], e.g., the visible light of a star. The electromagnetic energy is then recorded as the spectral distribution of wavelength, which reveals the properties of astronomical objects, such as their mass, density, distance, composition and luminosity.

In contrast, the photometric approach applies the coloured band-pass filter to capture and measure the energy in specific wavelength, such as ultraviolet, visible and infrared wavelengths. After taking photos with extremely high resolution in each band-pass, the intensity of light, or so-called flux, is measured for each luminous object in one sky range. The photometry has the advantage of a broader observation range and sensitivity to a fainter object compared to spectroscopy [57]. Typically, the Large Synoptic Survey Telescope (LSST)

survey aims to explore the unknown astronomical variability with decay time from 0 to 1 second and -4 to -20 negative logarithm of the flux [39]. The LSST survey is naturally inherent with the classification modelling idea. The difficulty is to distinguish the unknown flux trajectory by the trajectories from those of known stars, in other words, discover new astronomical objects. An example of visualizing the measured light flux is shown in Figure 2.1[38]. The shaded distribution is the light flux transmitted by each band-pass filter. The measured flux in each band-pass is a sum of the photons within given wavelength range, denoted as *ugrizy* which correspond to ultraviolet, green, red, infrared (*izy*).

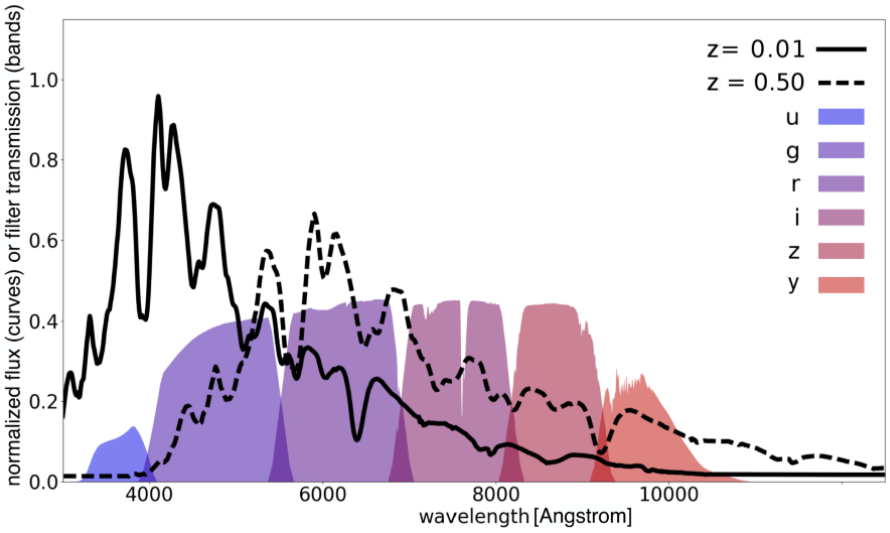


Figure 2.1: The influence of redshift on the distribution of detected wavelength. The *ugrizy* is the filter corresponding to different wavelength range on the abscissa. The ordinate is the fraction of light transmitted by the filter. The solid black curves are the spectra for a nearby Type Ia supernova at redshift $z = 0.01$, while the dashed curve is at redshift $z = 0.5$, corresponding to a longer distance [38].

Besides these, the gravitational waves, which are the ripples of space-time curvature in the universe, are another source of astronomical events, e.g., the collapse of binary black holes. The Laser Interferometer Gravitational-Wave Observatory (LIGO) has two Michelson interferometers which measure the interfered laser beam reflected by the beam splitter after passing through two long arms (4 km) which are aligned at right angles to each other. The longer arm is more sensitive to detect the length changes by the squeezing and stretch-

ing effect of the gravitational wave. When the gravitational wave passes two arms, the difference between the two arm lengths is proportional to the strength of the passing gravitational wave, referred to as the gravitational-wave strain [7]. The strain data is published in the LIGO open data center recorded with 16384 Hz or down-sampled 4096 Hz[15]. An example of strain event time series is shown in Figure 2.2 which is 1024 seconds of data with a down-sampled recording rate 4096 Hz. The timestamp is near the observation of the first detection of gravitational waves emitted from compact binary coalescence, named as GW150914 [13].

In this thesis, in order to test the time series similarity with the inclusion of temporal information in the astronomical data sets, we are focusing on two data sets, the simulated photometric light flux curve observed in LSST survey published as a Kaggle competition [8, 38], and the event strain time series from Gravitational-Wave Transient Catalog, GWTC-1[27], observed during the first and second operation run O1 and O2 from LIGO & VIRGO Gravitational Wave Open Science Center [15]. We focus on both 'extracted' data from extremely large databases: at least 85-petabytes resultant data set of LSST during its entire 10-year survey [20], and almost 780 GBs for the whole O1 run of LIGO-Hanford over nearly 6 months of operation with 16384 Hz of data [1].

2.2 Features of Astronomical Time Series

The measured values and timestamp values are two components of astronomical time series. In modeling astronomical time series data, the flux value is measured commonly in both spectroscopy and photometry. The astronomer usually obtains decades of observation for one source object to fully understand its features. The extensively large types of traditional observations are detected and classified by their flux pattern, such as periodic variation, quasi-periodic variation, outburst events and transient events [39].

Another non-negligible component of the astronomical Time Series is the timestamp value. The traditional time series modelling or classification considers only the observed value x_i, which is accurate enough to model the pattern. Also, the sampling procedure is consistent in default without any irregularity while irregularity has been discussed in the domain of time series [48]. Conversely, in modelling astronomical time series, the timestamp sequence is non-negligible: observed time for each observed value is used to estimate parameters, such as mass, rotational frequency, location and so on.

Therefore, we are introducing the definition of astronomical time series which emphasizes the timestamp. The time series is one realization, or a sample trajectory of a stochastic process $\{X_\cdot(\omega)|\omega \in \Omega\}$, denoted as $\{X_t|t \in T\}$ [16].

Definition A Stochastic Process is a family of random variables $\{X_t, t \in T\}$ defined on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$, denoted as $\{X_\cdot(\omega)|\omega \in \Omega\}$.

Each observation $X_t \in \{X_t\}$ is measured at a specified time $t \in \mathbb{T}_0$, in which $\mathbb{T}_0 \subset T \subset \mathbb{R}$ [16]. There are two types of time series, the discrete-time series and continuous-time series in which the timestamps are recorded discretely, e.g., $\mathbb{T}_0 = \{1, 2, \dots\}$, or continuously, e.g., $\mathbb{T}_0 = [0, 1]$. To incorporate the time indices t in the elastic distance of discrete-time series, we align the observation X_t by its timestamps t with increasing order. Then each observation X_t becomes a paired tuple, (x_i, t_i), in which $x_i \in \mathbb{R}^p$, $t_i \in \mathbb{T}_0$ and $t_i \leq t_{i+1}$.

Definition The time series X is defined as a set of paired tuple $(x_i, t_i) \in \mathbb{R}^p \times \mathbb{T}$, which is

$$X = \{X_i = (x_i, t_i) | x_i \in \mathbb{R}^p, t_i \in \mathbb{T}, t_i \leq t_{i+1}, i = 1, \ldots, m\}. \tag{2.1}$$

The emphasis of ordered timestamps t_i in the above definition promotes the introduction on the features of astronomical data sets. The following classification scheme is data-oriented based on simulated data from the LSST survey and gravitational wave strain data from LIGO. For both simulated and real astrophysics data, there are a few issues that require investigation in the classification model:

- Distorted observed value:

  $x_o = f(x_e, p_{o,e})$, which means observed value x_o is transformed from emitted value x_e by a function $f(\cdot)$ with parameters $p_{o,e}$. The subscript o stands for observation and e stands for emission.

  The Doppler shift influences the observed value; e.g., for a distant star outside of the Milky Way, the cosmological redshift z influences the rate of arrival of the photons compared to its emission rate. This phenomenon is called cosmological time-dilation, resulting in a fainter and redder observed light flux when the star is more distant from the observatory. The relationship between observed wavelength λ_o and emitted wavelength λ_e is $\lambda_o = (1 + z)\lambda_e$. From this formula, we observe that the higher redshift shifts the spectrum to longer wavelength, which means redder. Also, higher redshift reduces the light flux value, which means it is much fainter when observed. The Doppler shift leads to the fact that the further the object is, the fewer and more redshifts are observed, leading to less light flux and redshifted, wavelength distribution. As a result, the light flux data received from the observatory requires a transform to counteract the redshift effect. A simulation example is shown in the curves of Figure 2.1[38].

- Uneven sampling:

  $\Delta t_i \neq \Delta t_j$, in which $\Delta t_i = t_i - t_{i-1}$.

  Generally, in most data collection procedures, the time series is usually observed at a synchronized frequency. For example, the LIGO group publishes the gravitational wave strain data which has the 16384 Hz sample rate, which means the beam detector receives signal emitted from every 6.1×10^{-5} seconds in the interferometer. The observed strain time series is then approximated as a continuous trajectory sampled from the overall gravitational wave stochastic process. In other cases, it is hard to guarantee the synchronized sampling rate; e.g., the simulated data set, photometric LSST time series, have various observation timestamps because of its survey mechanism. The observatory has a 3 billion pixel camera which collects photos each 30 seconds in different grid regions of the sky. Two objects from different types in separate searching regions have intermittent timestamps of observation as shown in Figure 2.3. The object 29252 from class 4 has a denser observation timestamp compared to the object 11742403 from class 8.

- Missing value:

  $\Delta t_i = t_i - t_{i-1} \gg C$, where C is a normal sampling time rate.

  It's difficult to guarantee complete data collected during the scheduled observation range in the observatory. As an example, in the simulated data set, photometric LSST time series, large gaps are shown because the light flux from the object is not visible during night in the observatory. Instead of using the missing data mechanism, such as missing at random (MAR) or missing at completely random (MACR), we interpret the missing value by the large gap between two consecutive observing timestamps, which exceeds some threshold, denoted as a constant C. Applying this simplified version of modelling for missing data is due to the property of the observatory. In Figure 2.3, there are visible gaps between consecutive observations, which means

that the camera is taking pictures in another region of sky to complete a periodic task cycle. However, the variation of the light flux of one star is a stochastic process determined by its own mass, age, location, chemical composition and other significant feature parameters. These physical parameters are usually fixed except for an event, such as collapse and implosion, e.g., the stars usually locate at fixed point, which means the observed flux has the same influence from the Doppler shift. Thus, the observing time range is independent of the star's evolution process. As a result, the missing pattern can be modeled by a threshold C. Modelling this missing pattern is indispensable. The simplified situation takes the difference between consecutive timestamps as the feature of missing pattern.

- Noise:

  $x_{o,i} = x_{e,i} + \epsilon_{o,e,i}$, in which $\epsilon_{o,e,i} \gg x_{e,i}$ and $\epsilon_{o,e,i}$ is independent of $x_{e,i}$.

  Apart from the intrinsic physical mechanism, the observations usually include large amount of background noise during the decades of observation because the observational experiments in astronomy are mostly triggered by rare events. Taking Advanced LIGO strain data as an example, the background GW noise is dominant. The GW signal from the binary black holes merger is extremely small, hidden among background noise with large amplitude. The sources of the noise include quantum sensing noise, seismic noise, suspension thermal noise, mirror coating thermal noise, gravity gradient noise, transient noise from anthropogenic sources or weather or equipment malfunctions, and noise with spectral lines in certain high frequencies which is generated from electrical and mechanical devices or resonances [28]. Take the noise in the stadium as an example, the difficulty of detecting a GW signal is that of distinguishing one voice among the 100,000 people in the middle of the stadium during a football game. To interpret it, a basic additive noise model is given in terms of the observed measurement $x_{o,i}$, the emitted value from the target source $x_{e,i}$ and

error term $\epsilon_{o,e,i}$. An example is shown in Figure 2.2. The signal from binary black holes merger is revealed after filtering with a 35–350 Hz Butterworth band-pass filter to suppress large fluctuations outside the detectors' most sensitive frequency band [13].

- Discontinuity:

  $x_{t-}(\omega) = \lim_{s\uparrow t} x_s(\omega) \neq x_t(\omega)$, $x_{t+}(\omega) = \lim_{s\downarrow t} x_s(\omega) = x_t(\omega)$.

  The stellar evolution is not completely observed by human sight due to its extremely long lifetime. Though the stars are evolving at a slow pace, the vast universe contains numerous stars. This trade-off allows humans to observe multiple transient events of the stars, such as explosions and collapses before the stars turn into black holes. So, it is possible to see the abnormal jump and then the decaying duration in the data, e.g., in Figure 2.3, as a type II supernova, the object 29252 experiences core collapse at the beginning of the observation, emitting a large amount of energy due to the fusion of its core. Another object 11742403 is the kilonova, which is modelled by an explosive event from the merger of two neutron stars [37]. A kilonova event is rare, dim and with short decay duration compared to the supernova. The horizontal line in both time series indicates the discontinuity, and we model it by the right-continuous with left limits from the decaying light intensity.

To deal with the above features in an astronomical dataset, the idea of warping time will help us transform the time series. As we are focusing on the time-domain distance-based classifiers, we are introducing several distances in the following chapters while embedding the solution of the above issues.

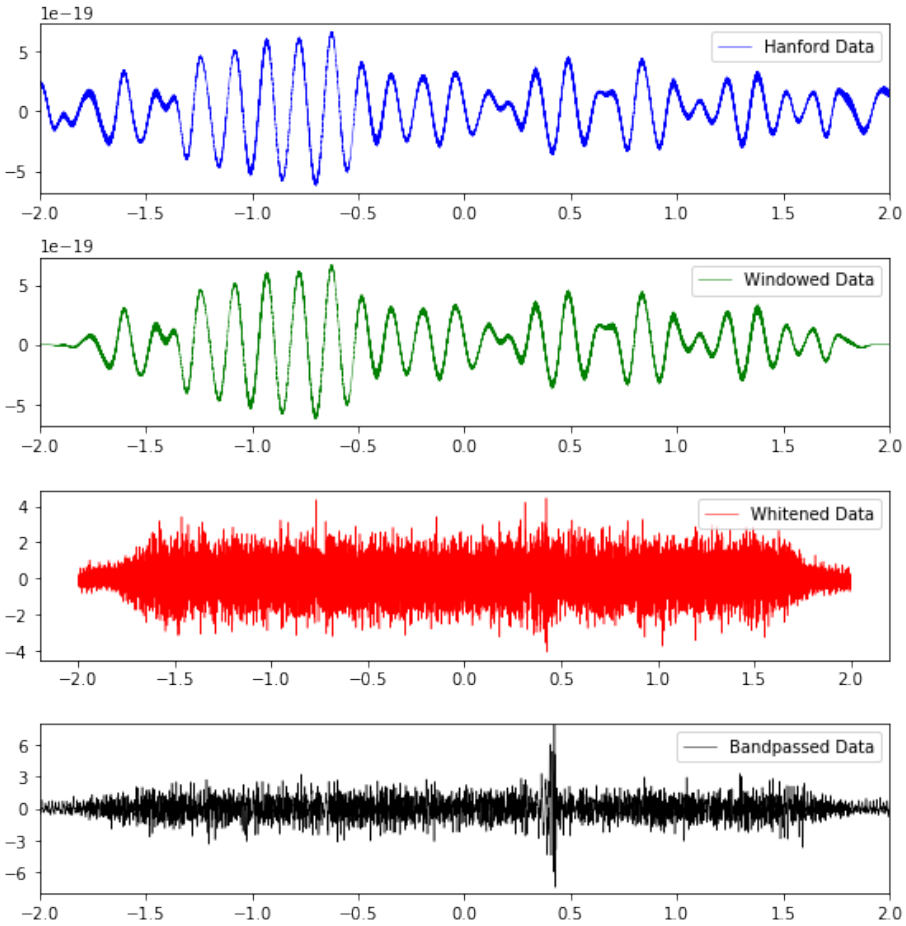


Figure 2.2: The GW150914 event data starts from timestamp 1126257414. The first panel labelled with 'Hanford Data' is the original strain data. The second panel labelled with 'Windowed Data' applies a tapered cosine window, or so-called Tukey window, at the beginning and end of the original data. The third panel labelled with 'Whitened Data' is the whitened strain data given the power spectral density calculated from full samples. The last panel labelled with 'Bandpass Data' is applying a Butterworth filter to the original strain data, with maximally flat frequency response in the band-pass, revealing the event hidden in the noise.

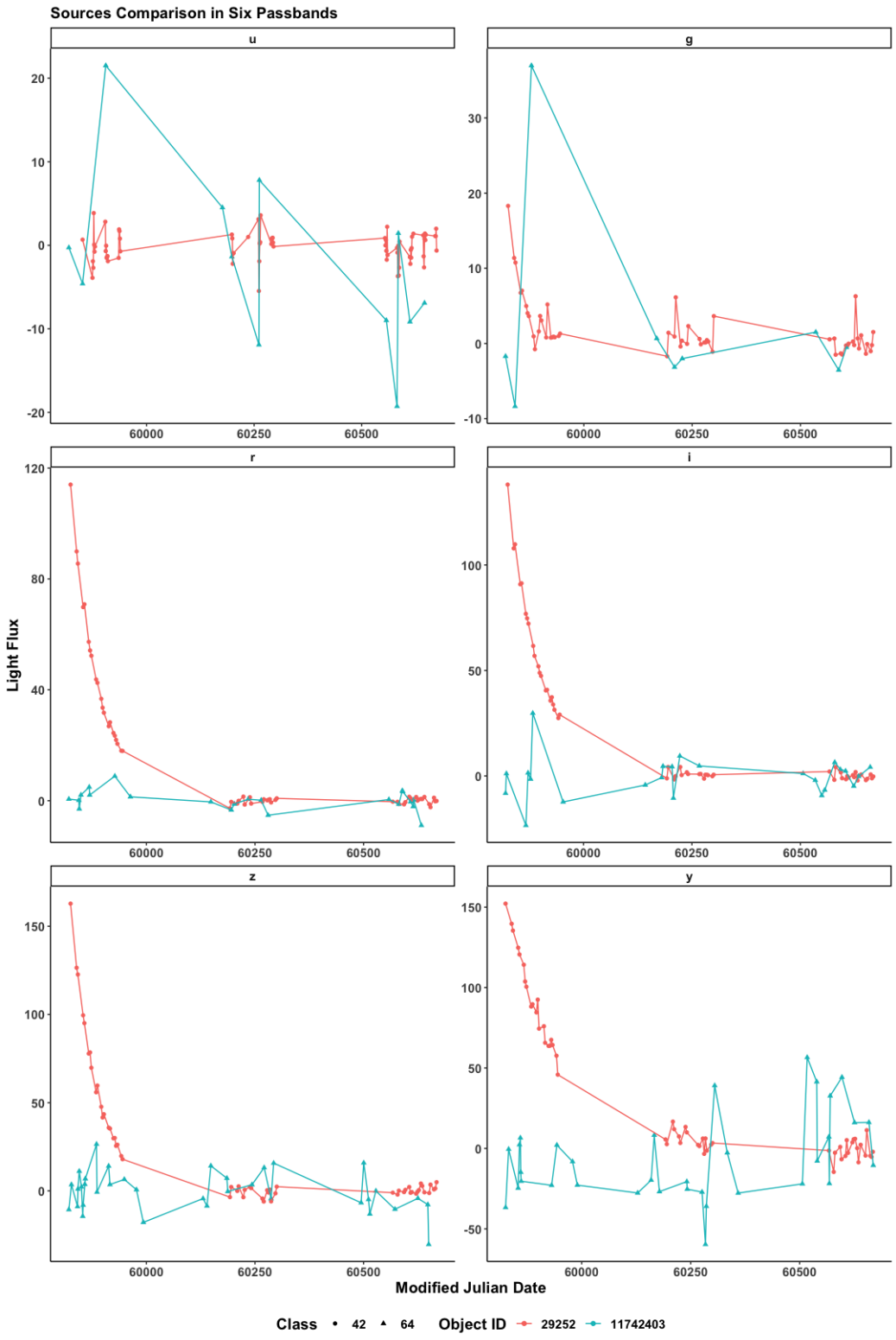


Figure 2.3: Light flux of two objects from different classes in six different astronomical filters.

Chapter 3

3 Quantify Similarity between Time Series

To estimate the similarity between object is a common task in the area of data mining, especially crucial for computer vision and pattern recognition. A few measures and algorithms have been developed and upgraded during the last few decades in dealing with finding patterns from meteorological observation, stock prices, astronomical signals and so on. To deal with the astronomical time series with the five issues we analyzed in the previous chapter, we are introducing four approaches to quantifying similarity between time series.

An algorithm-based approach is dynamic time warping (DTW) [50], which is the most reliable and practical method of all situations where estimating similarity is required. The DTW distance is usually considered as the most fundamental method, such as applying k-nearest neighbour using DTW as distance. The classic DTW is embedded with dynamic programming. The objective function of DTW is the overall minimum cumulative distance D, defined as the summation over the least distance cost among all matched pairs under certain warping criteria $\phi(\cdot, \cdot)$.

Inspired by DTW, Time Warp Edit Distance (TWED) includes the timestamp information in its loss function[41]. TWED reaches the balance between the so-called 'infinite-stiff' L_p distance and 'no-stiffness' DTW measure by defining a stiffness parameter λ. It borrows the ideas of Editing distance by defining the warping criteria ϕ by delete, match and add edit operations. Also, it adds the penalty to the timestamp difference between paired observations. So, the penalty loss function is defined by $l(X_i, Y_j) = |t_i - s_j| + \lambda$. The distance function $d(\cdot, \cdot)$ is defined in L_p space.

In the last, to complete the whole story, a measure is included, Skorohod distance [53]. It was first invented by A.V. Skorokhod to study the convergence in distribution of stochas-

tic processes with jumps. It might be useful in dealing with discontinuous astronomical signals. Meanwhile, Fréchet distance is applied to calculate the abstract Skorohod distance.

3.1 Dynamic Time Warping

Dynamic time warping (DTW) is a recursive method used to compare pattern similarity between two time series by dynamic programming. A warping curve is defined by selecting matching points with minimum distance cost between pairs of observed values from query and reference series. Besides the cumulative minimum distance between the query and reference time series, the warping curve permits the important property of DTW: the distortion in timestamps or series indices in pattern similarity matching.

The DTW distance was first introduced to deal with the nonlinear fluctuation in a speech pattern time axis [50]. In comparison to the Euclidean distance, the DTW distance has the property of elasticity, which allows a 'mismatch' in the time axis. So, it deals with a few common issues in real-time series problems:

- The sampling frequency is not synchronized.
- The sampling procedure is not consistent. A few intervals are missing.
- The intrinsic generation mechanism varies between observed series.

Two time series, query and reference, are denoted as follows

$$X = \{X_i = (x_i, t_i) | x_i \in \mathbb{R}^p, t_i \in \mathbb{T}, i = 1, \ldots, m\} \tag{3.1}$$

$$Y = \{Y_j = (y_j, s_j) | y_j \in \mathbb{R}^p, s_i \in \mathbb{T}, j = 1, \ldots, n\} \tag{3.2}$$

in which, t_i and s_j are timestamps correspondingly. The subsequence up to $k-$th observation

is denoted as

$$X(k) = \{X_i = (x_i, t_i) | i \leq k\},$$

$$Y(k) = \{Y_j = (y_j, s_j) | j \leq k\}.$$

For each pair of the observations, the cost of matching (X_i, Y_j) is defined by the distance between two observations, denoted as $d(\cdot, \cdot)$. To match two observations, a warping function $\phi(\cdot)$ is defined in the timestamp domain, from query to reference timestamp series, $\{t_i | i = 1, \ldots, m\} \rightarrow \{s_j | j = 1, \ldots, n\}$. For example, $\phi(t_i) = s_j$ represents the query observation time t_i is matched with the reference observation time point s_j. The observed value $X_i = (x_i, t_i)$ is then shifted to s_j and then evaluated the distance to Y_j rather than Y_i.

Compared to the Euclidean distance, which is defined to measure the distance between two sequences, X and Y, with the same length ($m = n$), the warping path of Euclidean distance is the identical function $\phi(i) = i$ defined on the index of the observation $i \leq m = n$. This identical warping path is also the diagonal of the distance matrix $M(x_i, y_j)_{n \times m} = d(x_i, y_j)$ when $n = m$. The introduction of the warping function allows the warping path deviation from the diagonal of the distance matrix M. The warping path is the series of matched timestamps pair of the warping function, which is defined as

$$\Phi(X, Y) = \left\{(t_i, s_j) | \phi(t_i) = s_j, i = 1, \ldots, m, j = 1, \ldots, n\right\} \tag{3.3}$$

$$= \left\{(t_1, s_1^*), (t_2, s_2^*), \ldots, (t_K, s_K^*), | \phi(t_k) = s_k^*, 1 \leq k \leq K = \max(m, n)\right\}. \tag{3.4}$$

From the symmetry of the DTW distance, $\mathrm{DTW}(X, Y) = \mathrm{DTW}(Y, X)$, which will be proved later, the warping path is based on the timestamp t_k in the query sequence without loss of generality, which is denoted as $\phi(t_k) = s_k^*$ in Eq.(3.4). The warping index is defined by the length of the longer time series, $K = \max(m, n)$. In other words, the key feature of DTW is permitting the wiggling in reference time series Y in order to achieve the best

alignment with query reference X. The DTW distance is then defined as the summation of the local distance between the matched pair $\left(X_k = (x_k, t_k), Y_k^* = (y_k^*, s_k^*)\right)$, in the warping path $\Phi(X, Y)$:

$$\mathrm{DTW}(X, Y) = \min_{\Phi(X,Y)} \sum_{k=1}^{K} d\left(X_k, Y_k^*\right) = \min_{\Phi(X,Y)} \sum_{k=1}^{K} d\left((x_k, t_k), (y_k^*, s_k^*)\right), \quad \text{in which } \phi(t_k) = s_k^* \tag{3.5}$$

Here the summation index k is based on the query sequence $X = \{(x_i, t_i)\}$ by the warping relation $\phi(t_k) = s_k^*$. The local distance $d(\cdot, \cdot)$ is usually chosen as the Euclidean distance. Compared to the identical warping function in Euclidean distance, the warping function of DTW is an injective function when the query sequence is shorter than the reference sequence ($m < n$), and vice versa, a surjective function when $m > n$.

To find the optimal warping function Φ, the dynamic programming is applied by evaluating the distance cost $d(X_i, Y_j)$ for each observation in the query sequence t_i. The sequential optimization framework is available for dynamic programming. To achieve the overall minimum of the objective function $D(X, Y)$, for each query sequence index, s_k, $k = 1, \ldots, K(= \max\{m, n\})$, we find the minimum of the DTW distance of the previous subsequence. In general, the DTW distance $\mathrm{DTW}(X, Y)$ is calculated recursively by the following function.

$$D(X(k), Y^*(k)) = d(X_k, Y_k) + \min \begin{cases} D(X(k), Y^*(k-1)) \\ D(X(k-1), Y^*(k)) \\ D(X(k-1), Y^*(k-1)). \end{cases} \tag{3.6}$$

The optimal value of the global warped distance, $\mathrm{DTW}(X, Y)$, is dependent on the shape of the warping function. In the original DTW, there are a few conditions [50] for a valid

warping path Φ. Taking one warped pair $\phi(t_k) = s_k^*$ from the warping path Φ as an example, we have $t_k = t_i$ and $s_k^* = s_j$ and the following conditions:

- Monotonocity:

  $t_{k-1} \leq t_k$ and $s_{k-1}^* \leq s_k^*$;

- Continuity:

  $t_k - t_{k-1} \leq t_i - t_{i-1}$ and $s_k^* - s_{k-1}^* \leq s_j - s_{j-1}$;

- Boundary:

  $\phi(t_1) = s_1, \quad \phi(t_m) = s_n.$

The monotonic warping path matches the order of the observation in the time-warped query sequence, indicating the monotonic trend in the time domain. A continuous warping path eliminates the jumping between matched points in the time axis. The allowable warping path is then restricted to the adjacent cells in the distance matrix M. The boundary condition exploits all information contained in both query and reference sequences by matching the head and tail of sequences. It also matched the range of the warping path, $K = \max(m, n)$ in Eq.(3.4).

The algorithm to calculate the DTW distance is in **Algorithm 1**. It has the time complexity $O(m \times n)$. The other generalized DTW distances follow a similar dynamic programming scheme.

An example of DTW is shown in Fig 3.1. Here we use the following two time series, with the Uniform distribution Unif$(-0.5, 0.5)$ as the fluctuation,

$$X = \{(x_i, t_i) | t_i = \{1, \ldots, 10\},\ x_i = 2t_i + \text{Unif}(-0.5, 0.5),\ x_5 = 2t_5 + \text{Unif}(-0.5, 0.5) - 2\}, \tag{3.7}$$

$$Y = \left\{(y_j, s_j) | s_j = \{1, \ldots, 10\},\ y_j = 2s_j + 1\right\}. \tag{3.8}$$

The DTW distance is 8.978 based on the Euclidean distance as the local distance $d(\cdot, \cdot)$. After the calculation of the local distance matrix M, the warping path is found by the step pattern denoted in Eq.(3.6). An artificial noise is added to the reference time series, $x_5 = 2t_5 + \text{Unif}(0, 1) - 4$, which leads to a pathetic alignment in both time series: Y_3 and X_6 match two points each. This phenomenon is the so-called *singularity* [36], which leads to a further discussion on the modification of DTW.

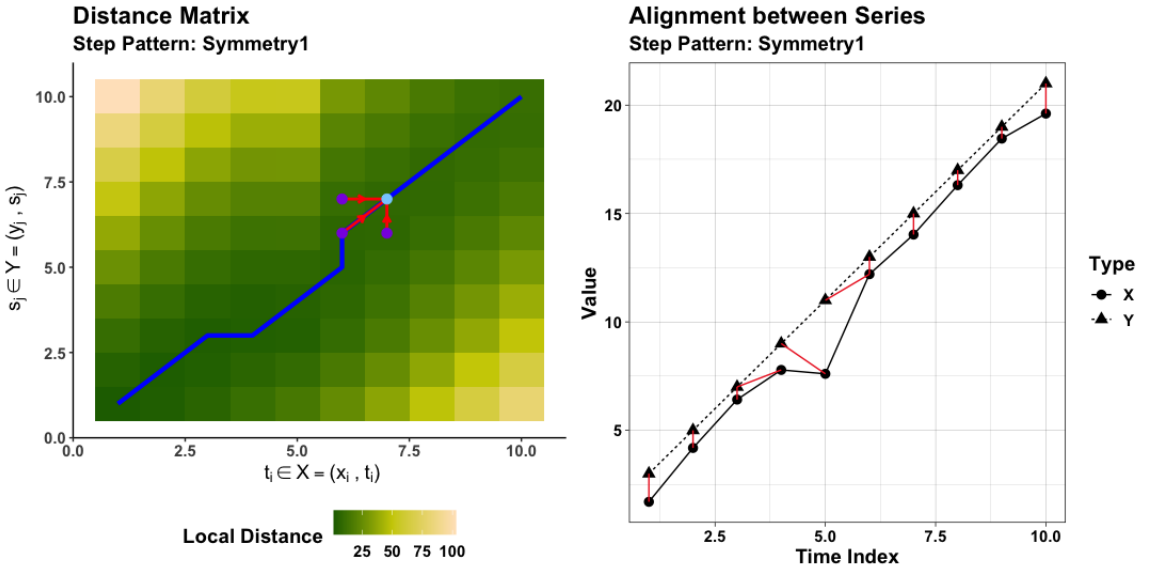


Figure 3.1: The example of DTW local distance matrix M, warping path $\Phi(X, Y)$ and its alignment $\phi(t_i) = s_j^*$. We use the 'symmetric1' step pattern which is in the middle of the left plot.

Algorithm 1: Original Dynamic Time Warping

Data: Query $X = \{X_i = (x_i, t_i) | i = 1, \ldots, m\}$, Reference $Y = \{Y_j = (y_i, s_j) | i = 1, \ldots, n\}$
Input: Warping function, $\phi(\cdot, \cdot)$, Local distance metric, $d(\cdot, \cdot)$
Output: Distance, $D(X, Y) = M_{m,n}$.

```
// Initialization
```

Generate cumulative distance matrix $M_{(n+1,m+1)}$ with $M_{i,j} = 0$ for $i, j \leq 1$;
$M_{0,j} = M_{i,0} = \infty$, for $i \in \{1, \ldots, m\}$, $j \in \{1, \ldots, n\}$;
for $i = 1$ **to** n **do**
    for $j = 1$ **to** m **do**
        cost $= d(x_i, y_j)$;

```
        // Update the cumulative distance matrix M.
```

        $M_{i,j} = \text{cost} + \min\left\{M_{i,j-1}, M_{i-1,j}, M_{i-1,k-1}\right\}$;

3.1.1 Window Size and Slope Constraint

In practice, the original DTW is having the problem of singularity, especially when reference series X has different length compared to query series Y. When X has fewer observations than Y ($m < n$), there is at least one observed value pair, $X_i = (x_i, t_i)$, distorted to multiple observations in query series by warping function, and vice versa when $m > n$. Another case is when $m = n$, it is possible to have a noisy data point which leads to a similar pathetic matching shown in Fig 3.1. One approach to solve it is by introducing the constraint of the warping window and the slope constraint which restricts the path of the warping function:

- Adjustment window with window size w:
  $\left|t_k^* - s_k^*\right| \leq w$
- Slope constraint:
  The path Φ should not be too steep or too shallow.

The adjustment window condition was first introduced in [50] to avoid a too excessive matched timing difference. When $w = 0$, the DTW degrades to the Euclidean distance when choosing $w = 0$ and the boundary condition. An algorithm which implements the Sakoe-Chiba band [50], i.e. the warping window is a band around the main diagonal (see Fig 3.3), is shown as follows in **Algorithm 2**. The time complexity of **Algorithm 2** is $O(mw)$.

An example is given to show the advantages and disadvantages of the warping window [33]. The reference time series is a cosine function, $Y = cos(s_j)$ in $[0, 2\pi]$ while the query series is $X = sin(t_i) + \epsilon_i$, $\epsilon_i \sim$ Unif(0,0.1). The global DTW result is shown in Fig 3.2. Most of the query sequence have an ideal strictly increasing alignment with the reference sequence, except for the beginning and the end of the sequence. Both the beginning sample of the reference sequence (red, dashed line) and the last sample of the query sequence (black,

Algorithm 2: DTW with Window w.

Data: Query $X = \{X_i = (x_i, t_i) | i = 1, \ldots, m\}$, Reference $Y = \{Y_j = (y_i, s_j) | i = 1, \ldots, n\}$
Input: Local distance metric, $d(\cdot, \cdot)$, window size, w
Output: Distance, $D(X, Y) = M_{m,n}$.
// Initialization
Generate cumulative distance matrix $M_{(n+1,m+1)}$ with $M_{i,j} = 0$ for $i, j \leq 1$;
$M_{0,j} = M_{i,0} = \infty$, for $i \in \{1, \ldots, m\}$, $j \in \{1, \ldots, n\}$;
$w = \max\{w, |m - n|\}$. // When $m \neq n$, $w \geq |m - n|$.
for $i = 1$ **to** m **do**
    for $j = \max\{1, i - w\}$ **to** $\min\{n, i + w\}$ **do**
        // Constrained distance matrix.
        cost = $d(x_i, y_j)$;
        // Update the cumulative distance matrix M.
        $M_{i,j} = \text{cost} + \min\left\{M_{i,j-1}, M_{i-1,j}, M_{i-1,k-1}\right\}$;

solid line) have multiple matches correspondingly. When changing the window size w from 0 to 35, the number of samples with multiple matches increases. When $w = 35$, the DTW distance with window w is the same as the global DTW in Fig 3.2. In DTW with window size w, the increasing window size promotes the higher conformity of the warping path of the global DTW. By this approach, there is no need to update the whole cumulative distance matrix M, which improves the computation efficiency dramatically.

The slope constraint restricts the slope of the warping path by the step pattern. The classic symmetic1 pattern is shown in Fig 3.1 and Eq.(3.6). This constraint prevents short sequences from matching too long ones [50]. The condition is expressed as a ratio p/q where p is the number of steps allowed in either the horizontal or vertical direction, q is the number of steps allowed in a diagonal direction. Given p/q of a warping path, after p steps in the same direction (horizontal or vertical), one is not allowed to step further in the same direction without stepping at least q steps in the diagonal direction. One extreme case is that $p/q = \infty$, $q = 0$ indicates no diagonal steps are allowed, which means the warping path is the identical function in the diagonal. In general, if two sequences are aligned away from the diagonal too much, which means only a small amount of points are matched, the overall slope will be large in order to achieve the global minimum of the distance. In prac-

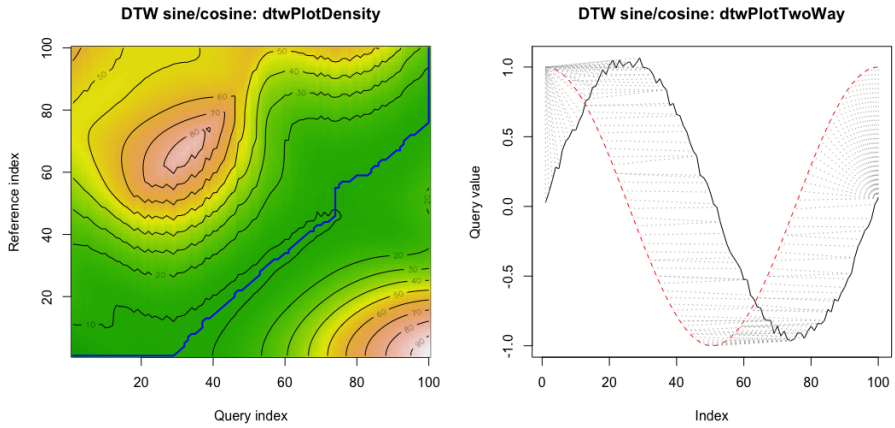


Figure 3.2: The local distance density plot and two-way time series matching plot. X and Y have the same sampling frequency and timestamp.

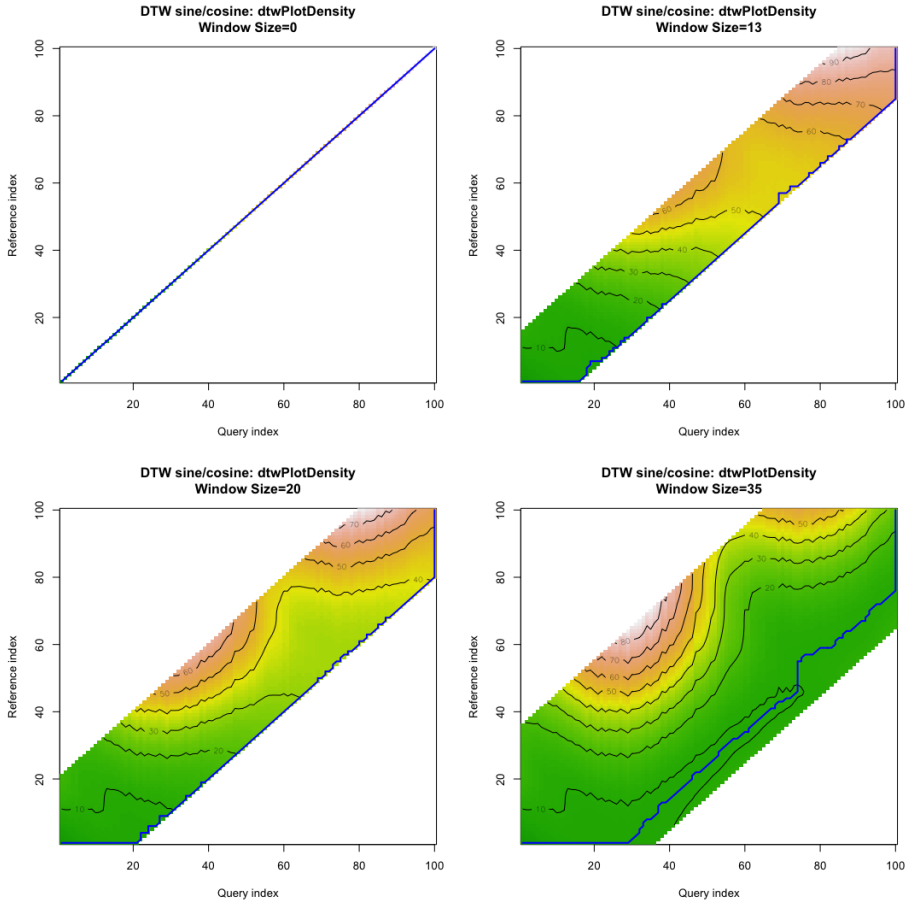


Figure 3.3: Window size restricts the number of singularities. The warping path is wiggled in the warping window boundary before reaching the global dtw optimal alignment.

tice, the extreme off-diagonal warping path represents the visible discrepancy between the query and reference sequence. So, the slope constraint makes little difference in the error of classification.

3.2 Time Warp Edit Distance

Before introducing Time Warp Edit Distance (TWED), we need to have a brief introduction of the Editing Distance with Real Penalty (ERP) distance first. The ERP distance aims to fix the non-metric property of elastic distance measures, such as DTW. Before introducing the Editing distance with Real Penalty(ERP), we need to introduce the properties of the metric. The metric $d(\cdot, \cdot)$ is a non-negative function which requires three properties:

- Symmetry:

  $d(x, y) = d(y, x)$;

- Identity of indiscernibles:

  $d(x, y) = 0 \iff x = y$;

- Triangular inequality:

  $d(x, z) \leq d(x, y) + d(y, z)$.

The DTW distance satisfies the first property due to the warping function constraints. However, it violates the identity of indiscernibles and the triangular inequality with the following examples [41]:

1. $X = \{1, 2, 2\}$, $Y = \{1, 2\}$, $DTW(X, Y) = 0$ but $X \neq Y$.

2. $X = \{1, 2, 2\}$, $Y = \{1, 2\}$, $Z = \{1\}$, $\text{DTW}(X, Z) = 2 > \text{DTW}(X, Y) + \text{DTW}(Y, Z) = 1$.

The edit distance was proposed to evaluate the dissimilarity between two strings according to the minimum number of transforms required to change into another string. The string edit distance is a metric which satisfies the triangular inequality [59]. The dynamic

programming ideas are embedded according to different types of edit distance. The Levenshtein distance is the prototype of ERP distance, which allows three transforms, insertion, deletion and substitution. Inspired by the idea of editing pattern, ERP distance is a generalization of Levenshtein distance to compare the dissimilarity between time series by defining the local distance for editing transform, d_{ERP} [19].

$$d_{\text{ERP}}(X_k, Y_k^*) = \begin{cases} |x_k - y_k^*|, & \text{if } x_k, y_k^* \text{ are not gaps} \\ |x_k - g|, & \text{if } y_k^* \text{ is a gap} \\ |y_k^* - g|, & \text{if } x_k \text{ is a gap,} \end{cases} \tag{3.9}$$

where g is a constant, which is usually fixed as 0. The 'gap' in Eq.(3.9) corresponds to the insertion and deletion transforms defined in Levenshtein distance. Specifically, to generalize the string editing transform, when we delete the X_k in reference sequence $X = \{\dots, X_{k-1}, X_k, X_{k+1}, \dots\}$, there is a gap between X_k and X_{k+1}, which means $X = \{\dots, X_{k-1}, X_{k+1}, \dots\}$ then. In other words, if the query sequence is denoted as $Y = \{\dots, Y_{k-1}, Y_k, \dots\}$, we are adding X_k to query sequence Y, which means $Y = \{\dots, Y_{k-1}, X_k, Y_k, \dots\}$ in order to match $X = \{\dots, X_{k-1}, X_k, X_{k+1}, \dots\}$. So the deletion in the reference sequence X is equivalent to the addition in the query sequence Y, and vice versa. Denote that the addition element is regarded as the gap element. A penalty of filling the gap is defined by the L1-norm from x_k/y_k^* to the constant g. The ERP distance is the L1-norm in $\mathbb{R}^p$ which allows local time shifting in essence [19]. Following the notation of subsequence, $X(k)$ and $Y(k)$, the dynamic programming updating function is shown in Eq. (3.10).

$$D_{\text{ERP}}(X(k), Y^*(k)) = d_{\text{ERP}}(X_k, Y_k^*) + \min \begin{cases} D_{\text{ERP}}(X(k-1), Y^*(k-1)), & \text{if } x_k, y_k \text{ are not gaps} \\ D_{\text{ERP}}(X(k-1), Y^*(k)), & \text{if } y_k \text{ is a gap} \\ D_{\text{ERP}}(X(k-1), Y^*(k-1)), & \text{if } x_k \text{ is a gap.} \end{cases} \tag{3.10}$$

The ERP distance defined from the above equation satisfies the triangle inequality from the definition of the above local distance in Eq. (3.9) [59]. Notice that the ERP distance is not defined to include timestamp space $\mathbb{T}$ in Eq. (3.10). The Time Warp Edit Distance (TWED) exploits timestamp in order to control the elasticity of the measure [41]. TWED uses the maximum timestamp difference threshold λ to decide whether a pair of samples, (X_i, Y_j) is matched. If the time indices difference between X_i and Y_j, $|i - j|$, is larger than λ, then X_i and Y_j are not 'matched', which requires deletion transform in either X or Y and a penalty λ. The local cost distance of the TWED d_{TWED} is then defined in Eq. (3.11)

$$d_{\text{TWED}}(X_k, Y_k^*) = \begin{cases} d(x_k, x_{k-1}) + \nu d(t_k, t_{k-1}) + \lambda, & \text{delete } X_k \\ d(x_k, y_k^*) + \nu d(t_k, s_k^*) + d(x_{k-1}, y_{k-1}^*) + \nu d(t_{k-1}, s_{k-1}^*), & \text{match } X_k \text{ and } Y_k^* \\ d(y_k^*, y_{k-1}^*) + \nu d(s_k^*, s_{k-1}^*) + \lambda, & \text{delete } Y_k^*. \end{cases} \tag{3.11}$$

Following the notation of subsequences, $X(k)$ and $Y(k)$, the dynamic programming updating function of TWED is

$$D_{\text{TWED}}(X(k), Y^*(k)) = d_{\text{TWED}}(X_k, Y_k^*) + \min \begin{cases} D_{\text{TWED}}(X(k-1), Y^*(k)), & \text{delete } X_k \\ D_{\text{TWED}}(X(k-1), Y^*(k-1)), & \text{match } X_k \text{ and } Y_k^* \\ D_{\text{TWED}}(X(k), Y^*(k-1)), & \text{delete } Y_k^*. \end{cases} \tag{3.12}$$

ν in Eq.(3.11) is the stiffness parameter which is a non-negative constant to control the penalty of the time difference between the pair of elements from query and reference time series. It is common to use the absolute difference, L_1 norm, in timestamp space T.

Algorithm 3: Time Warp Edit Distance

Data: Query $X = \{X_i = (x_i, t_i) | i = 1, \ldots, m\}$, Reference $Y = \{Y_j = (y_i, s_j) | i = 1, \ldots, n\}$

Input: Warping function, $\phi(\cdot, \cdot)$, Local distance metric, $d(\cdot, \cdot)$

Output: Distance, $D(X, Y) = M_{m,n}$.

```
// Initialization
```

Generate cumulative distance matrix $M_{(n+1,m+1)}$ with $M_{i,j} = 0$ for $i, j \leq 1$;

$M_{0,j} = M_{i,0} = \infty$, for $i \in \{1, \ldots, m\}$, $j \in \{1, \ldots, n\}$;

for $i = 1$ **to** n **do**

  for $j = 1$ **to** m **do**

    cost = $d_{\text{TWED}}(X_i, Y_j)$;

    `// Update the cumulative distance matrix` M.

    $M_{i,j} = \text{cost} + \min\left\{M_{i,j-1}, M_{i-1,j}, M_{i-1,k-1}\right\}$;

Similar to DTW, two toy time series defined in Eq.(3.8) are tested by TWED with parameter $\lambda = \nu = 1$. Calculated from Eq.(3.11) and Eq.(3.12), the TWED is 22.02. From the local distance matrix in Figure 3.4, we recognize that there is no singularity shown in the warping path, which is exactly the Euclidean distance. In this approach, the TWED reveals its potential to detect the abnormality by choosing suitable parameters. A simulation of TWED performance with different sets of parameters is included in the following section.

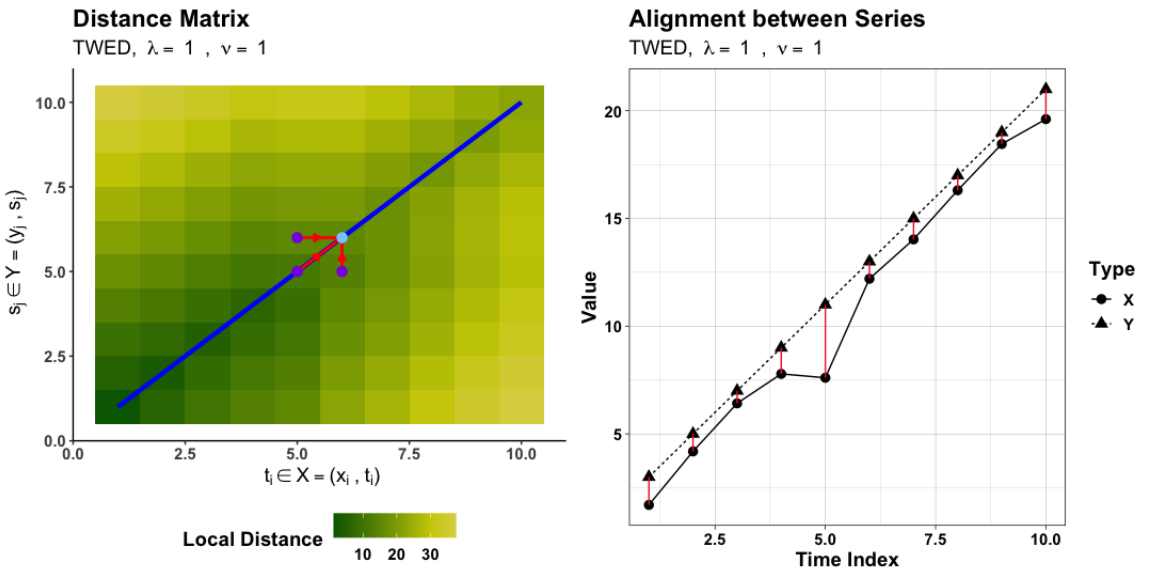


Figure 3.4: The example of TWED local distance matrix M, warping path $\Phi(X, Y)$ and its alignment $\phi(t_i) = s_j^*$. We use the 'symmetric1' step pattern which is in the middle of the left plot.

3.3 Simulation Test

Before applying the DTW and TWED to the astrophysics dataset, a simulation pretest is performed. The first part is about the distance distribution of two Gaussian white noises. The theoretical distributions are derived for both Euclidean distance and DTW distance. Then, a few experiments are conducted to test how TWED distance alleviates the impact of distortion in amplitude, uneven sampling, missing value, noise and discontinuity compared to DTW distance.

3.3.1 The Distribution of Distance

We list a preliminary test on the performance of Euclidean distance and DTW on Gaussian white noise, $\mathcal{N}(0, 1)$. We denote that both query and reference time series, X and Y in Eq.(3.2), follow the Gaussian white noise independently, $X \sim \mathcal{N}(0, 1)$, $Y \sim \mathcal{N}(0, 1)$.

From the definition of the Euclidean distance, the Euclidean distance is calculating the

square root of the summation of the squared difference between X and Y. Denote that $X, Y \sim \mathcal{N}(0, 1)$ independently, then we have $(X - Y)^2/2 \sim \chi^2(1)$. Notice that $\chi^2(1)$ is also a Gamma distribution with shape $k = \frac{1}{2}$ and scale $\theta = 2$, denoted as $\Gamma(\frac{1}{2}, 2)$. The Euclidean distance between two Gaussian white noises is then distributed as a chi distribution.

$$\frac{1}{\sqrt{2}} d_{\text{Euclidean}}(X, Y) = \sqrt{\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{2}} \sim \chi(n), \tag{3.13}$$

in which $n = m$ are the lengths of X and Y correspondingly. A simulated result is shown in the left plot of Figure 3.5, indicating a perfect fit of Eq.(3.13) and the distance distribution under the Euclidean distance.

The situation for DTW distance is complicated due to the property of the warping path and step pattern shape. We use the squared Euclidean distance $d(x, y) = (x - y)^2$ as the local distance metric for simplicity and generality. The step pattern of DTW is 'symmetric1' as shown in Figure 3.1.

From **Algorithm 1** and Eq.(3.6), each matched pair of observations in the warping path is the pair with the minimal distance among three pairs according to the step pattern, e.g., $\min\{d(x_i, y_{j-1}), d(x_{i-1}, y_j), d(x_{i-1}, y_{j-1})\}$ for 'symmetric1' step pattern. Notice that each distance value is the squared Euclidean distance between two independently distributed normal random variables, which are Chi-square random variables with coefficient 2, e.g., $(x_i - y_j)^2 \sim 2\chi^2(1)$, which also means $(x_i - y_j)^2 \sim \text{Gamma}(\alpha = \frac{1}{2}, \theta = 4)$. Then for each increment in the global DTW distance, the distance between matched pairs (x_i, y_i^*) is the minimum of three correlated Chi-square (Gamma) random variables. Take $d(x_{i-1}, y_{j-1})$ and $d(x_i, y_{j-1})$ as an example, the Pearson correlation coefficient between them is

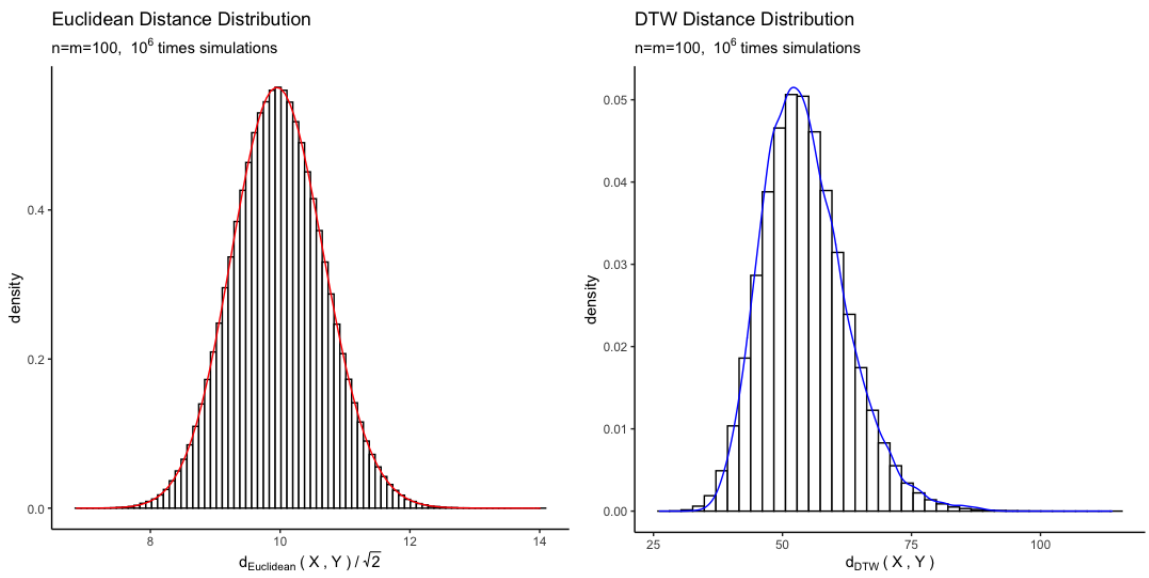


Figure 3.5: The histogram of the Euclidean distance and DTW distance between two sequences of independent Gaussian white noise. There are 10^6 times trials. The query and reference sequences have the same length ($m = n = 100$). The histogram in both plots are the simulated distances. The left plot has the histogram of the Euclidean distance distribution. The red curve is the probability density of $\chi(n)$ distribution derived from Eq.(3.13). The right plot has the histogram of the DTW distance distribution. The blue curve is the probability density curve of Eq.(3.15). It is estimated from the simulated sample by the Monte Carlo method with a Gaussian kernel.

$$
\begin{aligned}
\mathrm{Cov}\left(d(x_{i-1}, y_{j-1}), d(x_i, y_{j-1})\right) &= \mathrm{Cov}\left((x_{i-1} - y_{j-1})^2, (x_i - y_{j-1})^2\right) \\
&= \mathrm{E}\left[(x_{i-1} - y_{j-1})^2 (x_i - y_{j-1})^2\right] - \mathrm{E}\left[(x_{i-1} - y_{j-1})^2\right] \mathrm{E}\left[(x_i - y_{j-1})^2\right] \\
&= \mathrm{E}\left[y_{j-1}^4\right] + \mathrm{E}\left[y_{j-1}^2\right] \mathrm{E}\left[x_i^2\right] + \mathrm{E}\left[y_{j-1}^2\right] \mathrm{E}\left[x_{i-1}^2\right] + \mathrm{E}\left[x_i^2\right] \mathrm{E}\left[x_{i-1}^2\right] \\
&\quad - \mathrm{E}\left[(x_{i-1} - y_{j-1})^2\right] \mathrm{E}\left[(x_i - y_{j-1})^2\right] \\
&= 3 + 1 + 1 + 1 - 2 \times 2 = 2
\end{aligned}
$$

$$
\rho = \frac{\mathrm{Cov}\left(d(x_{i-1}, y_{j-1}), d(x_i, y_{j-1})\right)}{\sqrt{\mathrm{Var}(d(x_{i-1}, y_{j-1}))\mathrm{Var}(d(x_i, y_{j-1}))}} = \frac{2}{\sqrt{8} \times \sqrt{8}} = \frac{1}{4}.
$$

Similarly, the Pearson correlation coefficient of $d(x_{i-1}, y_j)$ and $d(x_{i-1}, y_{j-1})$ is also $\frac{1}{4}$. $d(x_i, y_{j-1})$ and $d(x_{i-1}, y_j)$ are independently distributed. The covariance matrix will be useful in later

simulation of the probability density function of DTW distance. Due to the fact that the theoretical distribution of the minimum of three correlated Chi-square random variables has no explicit expression, we use the Monte Carlo method to simulate it by Multivariate Normal random vector as shown in the following equation. Denote D_1, D_2, D_3 as the three distance components, we have the minimum among these three, $D_{(1)} = \min\{D_1, D_2, D_3\}$, simulated as follows

$$\begin{pmatrix} (x_{i-1} - y_{i-1})^2 \\ (x_{i-1} - y_i)^2 \\ (x_i - y_{i-1})^2 \end{pmatrix} = \begin{pmatrix} D_1^2 \\ D_2^2 \\ D_3^2 \end{pmatrix}, \quad \text{in which} \begin{pmatrix} D_1 \\ D_2 \\ D_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{pmatrix} \right). \tag{3.14}$$

Then we analyze the composition of the DTW distance by its local distance increment. Recall the definition of singularity in the warping path, which means there are the repeated matched observations in either query or reference time series, e.g., the warping path with the pattern $\{\ldots, (s, j), (s+1, j), \ldots, (s+k_s, j), \ldots\}$ has a consecutive singularity chunk of length k_s with repeated Y_j from the reference series Y, starting from X_s and ending at X_{s+k_s}. The starting point of the singularity chunk is defined as *canonical singularity* with index s. The length of a consecutive singularity chunk is denoted k_s corresponding to each canonical singularity index s. Then the warping path of the DTW distance consists of uniquely matched pairs, $\{(i, i^*)|i = 1, \ldots, N_I\}$, and singularity chunks with the canonical singularities $\{(s, s^*)|s = 1, \ldots, N_S\}$. Furthermore, for the singularity chunks with the same length k_s, e.g., $\{\ldots, (s, j), \ldots, (s+k_s, j), \ldots, (l, i), \ldots, (l+k_l, i), \}$ in which $k_s = k_l$, the canonical singularity chunks are independently distributed between each other. Then the dependent distance part can be rearranged by the length of singularity chunks, k_s. The set of canonical singularities is then simplified as $\{(s_k, s_k^*)|k = 1, \ldots, K_{\mathrm{D}}\}$, in which the s_k is the time index of the canonical singularities with a chunk of length k and K_{D} is the maximum length among all singularity chunks. We denote the number of singularity chunks of length k as n_k. Considering the boundary of DTW distance, the beginning and end of the warping

k	1	2	3	4	5	6
$\hat{n}_k$	66	18	7	3	1	1

Table 3.1: Parameters used for Monte Carlo simulation of distribution defined in Eq.(3.15). It is estimated from 10^6 times simulation of DTW distance of two Gaussian white noise series of length 100.

path derives the local distance is $D_0 \sim 2\chi^2(1)$, the DTW distance is then decomposed as

$$\begin{aligned} d_{\mathrm{DTW}}(X,Y) &= \sum_{i=1}^{K} d(x_i, y_i^*) \\ &= d(x_1,y_1) + d(x_n,y_n) + \sum_{i=1}^{N_{\mathrm{I}}} d(x_i,y_i^*) + \sum_{s=1}^{N_{\mathrm{D}}}\sum_{j=1}^{k_s} d(x_j,y_s^*) \\ &= (x_1-y_1)^2 + (x_n-y_n)^2 + \sum_{i=1}^{N_{\mathrm{I}}}(x_i-y_i^*)^2 + \sum_{s=1}^{N_{\mathrm{D}}}\sum_{j=1}^{k_s}(x_j-y_s^*)^2 \\ &= (x_1-y_1)^2 + (x_n-y_n)^2 + \sum_{i=1}^{N_{\mathrm{I}}}(x_i-y_i^*)^2 + \sum_{k=1}^{K_{\mathrm{D}}} n_k \sum_{j=1}^{k}(x_{s_k+j}-y_{s_k}^*)^2 \\ &= 2D_{0,1} + 2D_{0,n} + 2\sum_{i=1}^{N_{\mathrm{I}}} D_{(1),i} + \sum_{k=1}^{K_{\mathrm{D}}} k \sum_{j=1}^{n_k} D_{(1),j}. \end{aligned} \tag{3.15}$$

In the above decomposition, the first summation includes all independent minimums among three correlated Chi-square random variables, denoted as $D_{(1),i}$ with length N_{I}. The second part consists of the repeated observations which are arranged by their length k up to maximal singularity chunk length K_{D}, in which $D_{(1),j}$ is independent of each other. Here, the simplification is replacing the local distance $(x_{s_k+j} - y_{s_k}^*)^2$ with repeated matched element $y_{s_k}^*$ by the local distance $(x_{s_k} - y_{s_k}^*)^2$ with it canonical singularity x_{s_k}. Then there is a constant k indicating the duplicates.

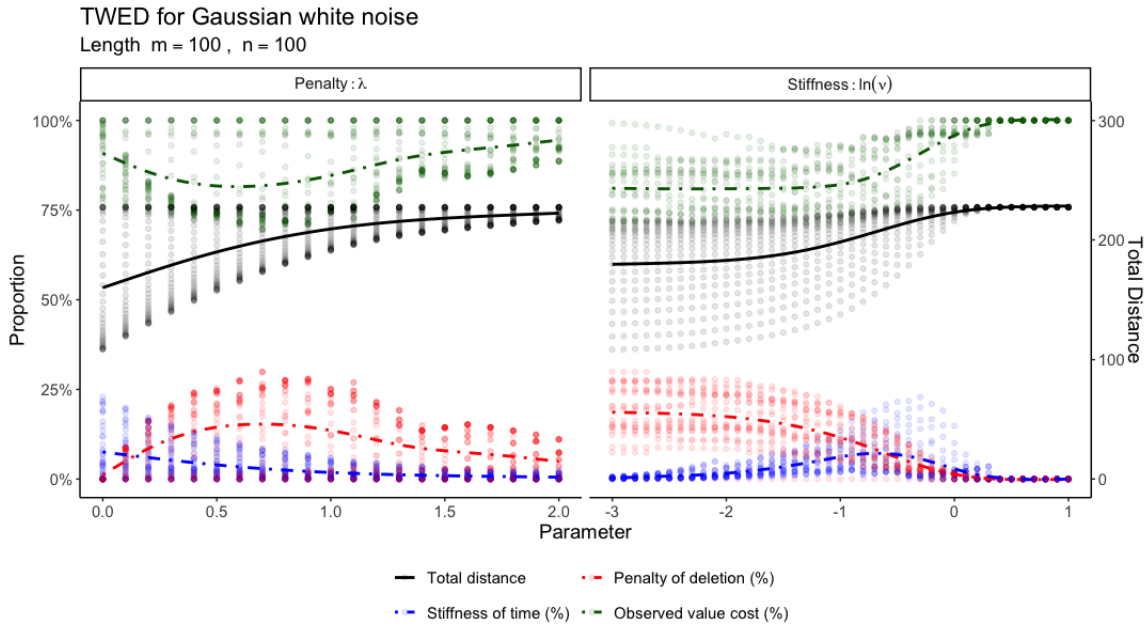
To simulate the above distribution in Eq.(3.15), for 10^6 warping paths of DTW distance, we record the average number $\bar{n}_k$ of each canonical singularity with the same length k,

including the number of independent non-singular local distance N_{I}. We take the k into consideration when parameter $\bar{n}_k \geq 1$ as shown in Table 3.1, which means relatively long singularity chunks are removed in Monte Carlo simulation.Then we replace the distance in Eq.(3.15) by the minimum distance $D_{(1)}$ defined from Eq.(3.14). In the right plot of Figure 3.5, the probability density curve derived from Eq.(3.14) with Gaussian kernel has a good fit of the distribution of DTW distance of Gaussian white noise. In practice the long series when only a tiny shift is observed, e.g., $\bar{n}_1$ in Table 3.1 is very large, we claim that DTW distance follows a Gaussian distribution for simplicity.

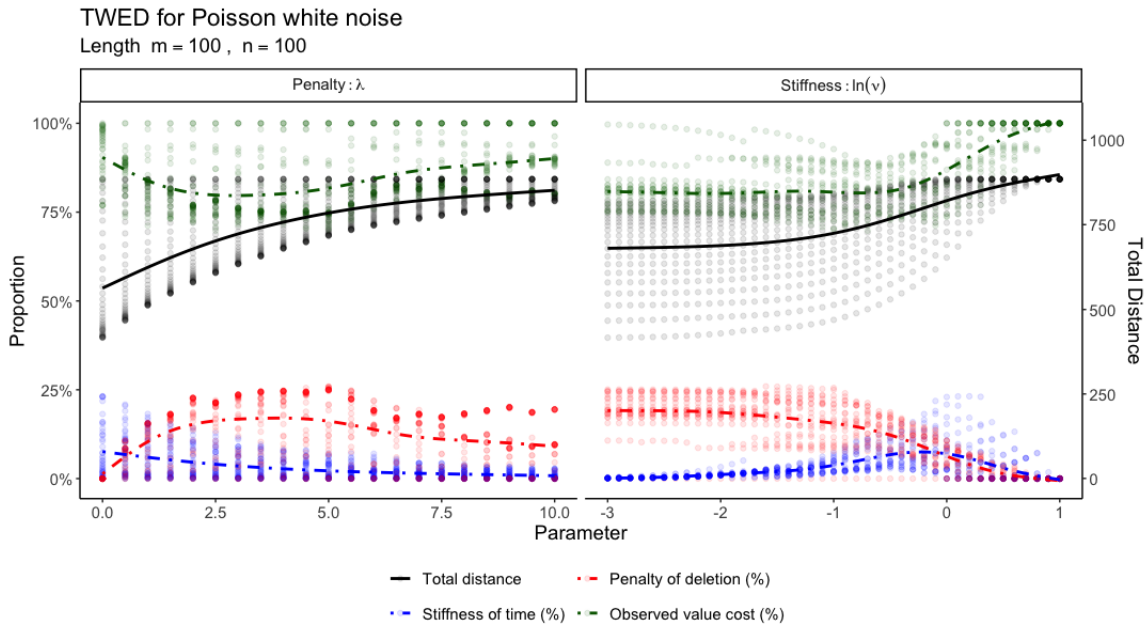
The distribution of TWED distance is more complicated comparing to DTW because it contains the information in time domain, which controls the deletion or matching operation for both query and reference series. However, similar to DTW, we consider that TWED distance follows a Gaussian distribution in the case of long series with tiny shifts. Besides the distribution, the following subsection analyzes the sensitivity of TWED with respect to its parameters.

3.3.2 TWED Parameter Choices

Recall the local distance of TWED in Eq. (3.11), we have two parameters, stiffness coefficient ν of time cost and penalty constant λ of deletion, undetermined due to the relationship of parameter magnitude is unknown to most situations. To reveal the performance of parameter combination, we apply the TWED with various sets of parameters, (λ, ν), to two white noises, Gaussian and Poisson. We then focus on the proportional contribution to the TWED distance by the decomposition of the local distance defined in Eq. (3.11), which is decomposed as

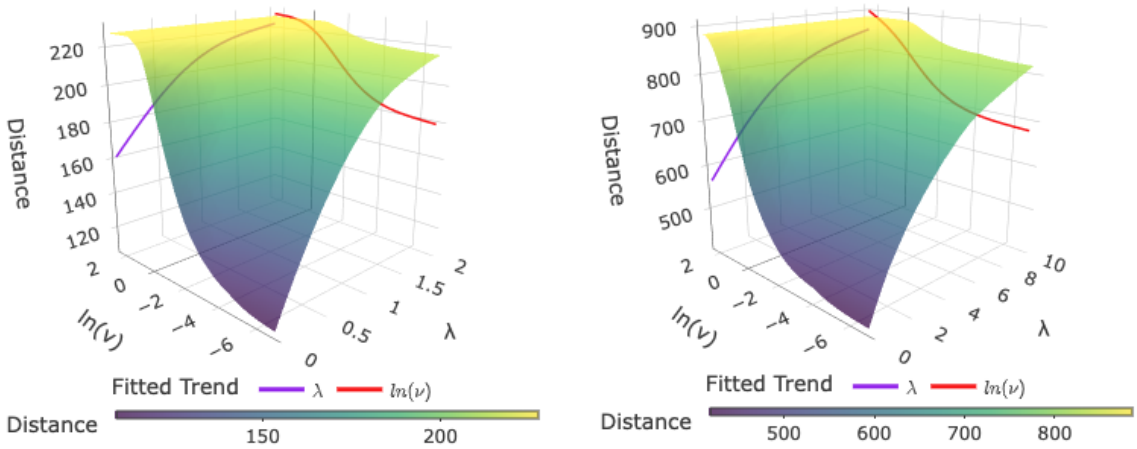


(a) Gaussian white noise defined in Eq.(3.17)



(b) Poisson white noise defined in Eq.(3.18)

Figure 3.6: The TWED distance variation against parameters λ and ν. The TWED is evaluated between two sequences of independent Gaussian white noise and Poisson white noise. The query and reference sequences have the same length ($m = n = 100$). The black curves in both plots are the total distance of TWED. The dashed curves are distance contributions proportional to the total TWED distance, which is defined in Eq.(3.16). All curves are smoothing curve modelled by a generalized additive model with a shrinkage of cubic spline basis. The stiffness parameter is transformed by logarithm for a clear visualization of the changes in distance and each component proportion.



(a) Gaussian white noise defined in Eq.(3.17)

(b) Poisson white noise defined in Eq.(3.18)

Figure 3.7: The TWED distance surface of variation against the combination of parameters λ and ν. The TWED is evaluated between two sequences of independent Gaussian white noise and Poisson white noise. The query and reference sequences have the same length ($m = n = 100$). The red curves in both plots are the fitted trend between the distance of TWED and parameter λ, while the purple curves represent the relationship of parameter $\ln(\nu)$ for a clear visualization. All marginal curves are smoothing curves modelled by a generalized additive model with shrinkage of cubic spline basis.

$$d_{\text{TWED}}(X_k, Y_k^*) = \begin{cases} d(x_k, x_{k-1}) + \nu d(t_k, t_{k-1}) + \underbrace{\lambda}_{\text{Cost of deletion penalty}}, & \text{delete } X_k \\ \underbrace{d(x_k, y_k^*) + d(x_{k-1}, y_{k-1}^*)}_{\text{Cost in observed value}} + \underbrace{\nu d(t_k, s_k^*) + \nu d(t_{k-1}, s_{k-1}^*)}_{\text{Stiffness cost controlled by time}}, & \text{match } X_k \text{ and } Y_k^* \\ d(y_k^*, y_{k-1}^*) + \nu d(s_k^*, s_{k-1}^*) + \lambda, & \text{delete } Y_k^*. \end{cases} \tag{3.16}$$

The cost in observed value, the stiffness cost controlled by time and the cost of deletion penalty are corresponding to terms related to observed value x_k, y_k, time stamps, t_k, s_k and penalty λ. The relationships of distance proportion of each components are evaluated for parameter grids in $\nu \in [10^{-3}, 1]$ and $\lambda \in [0, 10]$(Poisson), $\lambda \in [0, 2]$(Gaussian), as shown in Figure 3.6. From Figure 3.6a, the Gaussian white noise is generated from standard Gaussian distribution in Eq.(3.17). The query sequence is added by 3 in each simulated value.

The Gaussian white noise series are scaled in order to erase the effect of magnitude differences. As a comparison, the Poisson white noise series remain their original magnitude in Eq.(3.18).

$$X = \{(x_i, t_i)|x_i \sim \mathcal{N}(3, 1)\}\,,\ Y = \left\{(y_j, s_j)|y_j \sim \mathcal{N}(0, 1)\right\}. \tag{3.17}$$

$$X = \{(x_i, t_i)|x_i \sim \text{Poisson}(10) + 3\}\,,\ Y = \left\{(y_j, s_j)|y_j \sim \text{Poisson}(10)\right\}, \tag{3.18}$$

$$\text{in which } t_i, s_j \in \{1, \ldots, 100\}.$$

From Figures 3.6 and 3.7, there is a similar increasing trend of TWED distance when λ and ν converge to a certain upper bound when both parameters, penalty parameter λ and stiffness parameter ν increase. The values have a decreasing oscillation, indicating the convergence with decreasing variance. Two parameters contribute to the TWED total distance in two approaches.

- Penalty parameter λ: indicates the proportion trend of match and deletion.

  In the left plot of Figure 3.6a, the main effect of penalty λ in Gaussian white noise is on the penalty cost of deletion: when λ increases, the contribution of the penalty of deletion increases first and then decreases with peak at about $\lambda \approx 0.7$. As a trade-off, the cost in observed value decreases first then increases to compensate the penalty cost of deletion. When $\lambda > 1.5$, the total distance is converging to the maximum, accompanied by decreasing converged penalty cost of deletion.

  A similar behaviour of λ is shown in the left plot of Figure 3.6b. The peak point of the proportion of deletion penalty is achieved when $\lambda \approx 3$. As λ is only shown in the deletion operation, the range of penalty proportion indicates the proportion of deletion operation.

An ideal value of penalty parameter λ is constrained by the tolerance of deletion operation.

- Stiffness parameter ν: provide control on the matching operation.

  In the right plot for stiffness parameter in Figure 3.6a, the increasing $\ln(\nu)$ has a depression effect on the distance proportion of deletion penalty, while stiffness distance proportion of time increases first and then diminishes. The magnitude of oscillation converges to 0 when $\ln(\nu) > 0$, which is $\nu > 1$. The stiffness cost of time has an visible influence from $\ln(\nu)$ when $-2 \leq \ln \nu \leq 0$, which is $0.01 \leq \nu \leq 1$. This range can be roughly estimated by the relative magnitude ratio between a range of observed values and the range of time stamps in the scaled Gaussian white noise, which is about $6/100 = 0.06$. In order to contribute stiffness controlled by time, the stiffness parameter ν should be larger than 0.06, which is $\ln(\nu) \geq -1.22$, as a compensation of magnitude difference between the observed value and time stamp.

  In Poisson white noise in Figure 3.6a, the magnitude ratio between a range of observed values and a range of timestamps is about $20/100 = 0.2$. In order to contribute stiffness controlled by time, the stiffness parameter ν should be larger than 0.2, which is $\ln(\nu) \geq -0.70$.

  The stiffness parameter is ν, which constrains the deletion operation by enlarging the penalty in differences in the time stamp. Though the proportion of stiffness cost contributes little to the total distance, it has a non-negligible influence on the total TWED distance when properly chosen by checking the magnitude ratio between a range of observed values and a range of time stamps.

In practice, k-fold cross-validation is used to find the optimal $\hat{\lambda}$ and $\hat{\nu}$. We can choose the best parameters by evaluating the average score, e.g., accuracy in classification model, among the k splits of test data with a classification model trained by the corresponding training data. Except for the traditional tuning methods, the grid search or randomized

search of parameters, Bayesian optimization [54] can be used with the help of the distance variation as shown in Figures 3.6 and 3.7.

3.4 Skorohod Distance

The Skorohod topology is a powerful tool to investigate stochastic process limits, which allows a uniformly small deformation of the time scale. This is in close analogy to the uncertainty principle [14], which reveals a fundamental lower bound of the precision in measuring the position and momentum of a particle simultaneously. The distortion in time naturally embodies the idea that the timestamp t_i cannot be perfectly measured given that the measured value of the position x_i is accurate enough in one observation.

The Skorohod topology is defined on the Space D, which includes all real functions x on [0,1] that are right continuous and have left-hand limits. These functions are called *càdlàg* functions [14].

1. For $0 \leq t < 1$, $x(t+) = \lim_{s \downarrow t} x(s)$ exists and $x(t+) = x(t)$.
2. For $0 \leq t < 1$, $x(t-) = \lim_{s \uparrow t} x(s)$ exists.

To evaluate the discontinuity, we introduce the uniform norm, $\|x\|$. $\|x\|$ is bounded in Space D due to the property that there can be at most finitely many points t at which the jump $|x(t) - x(t-)|$ exceeds a given positive number [14].

$$\|x\| = \sup_t |x(t)| < \infty.$$

When x is the real function from $[0, 1]$ to the vector space $\mathbb{R}^p$, $|x(t)|$ is the norm defined on the vector space $\mathbb{R}^p$, with notation $\|x(t)\|_p$. Denote that $x, y \in D$. The Skorohod distance measures the uniform metric $\|x - y\lambda\|$ allowing small perturbations λ in abscissa. Denote that Λ is the set of strictly increasing, continuous mappings of $[0, 1]$ onto itself. If $\lambda \in \Lambda$,

then $\lambda 0 = 0$ and $\lambda 1 = 1$. Let x and y be in the Space D. Denote $d(x, y)$ as Skorohod distance defined as the infimum of those positive ϵ for which there exists a $\lambda \in \Lambda$ satisfying

$$\sup_t |\lambda t - t| = \sup_t |t - \lambda^{-1} t| < \epsilon$$
$$\sup_t |x(t) - y(\lambda t)| = \sup_t |x(\lambda^{-1} t) - y(t)| < \epsilon.$$

By the uniform metric $\|\cdot\|$ and identity map I on [0, 1], a concise form of Skorohod distance is defined as

$$d_{\text{Skorohod}}(x, y) = \inf_\lambda \{\|\lambda - I\| \vee \|x - y\lambda\|\}. \tag{3.19}$$

The Skorohod distance is a metric, and this is proved in Th.A.1. A modified Skorohod distance, which is also a metric, is also defined, and this is proved in Th.A.2.

$$d^*_{\text{Skorohod}}(x, y) = \inf_\lambda \{\|\lambda - I\| + \|x - y\lambda\|\}. \tag{3.20}$$

3.4.1 Fréchet distance

To compute the Skorohod distance between two time series, we need to apply the discrete Fréchet distance, which is defined on the polygonal curve of univariate time series [26]. The Fréchet distance is a measure of similarity between two curves mapping to vector space $\mathbb{R}^p$. Denote that these two curves are $f : [a_f, b_f] \to \mathbb{R}^p$ and $g : [a_g, b_g] \to \mathbb{R}^p$.

$$d_{\text{F}}(f, g) = \inf_{\phi_f, \phi_g} \max_{0 \le t \le 1} \|f(\phi_f(t)) - g(\phi_g(t))\|_p, \tag{3.21}$$

where $\|\cdot\|_p$ is the norm on vector space $\mathbb{R}^p$, and where ϕ_f and ϕ_g are continuous and strictly increasing bijective functions onto $[a_f, b_f]$ and $[a_g, b_g]$ [26]. A similar construction

of Skorohod distance and Fréchet distance is shown in Eq.(3.19) and (3.21). In fact, the two distances are exactly the same when choosing a proper norm on a generalized mapping space of curve, $\mathbb{R}^p \times T$. Inspired by the work [26], from the definition of the generalized Skorohod distance, we define a distance on $(f,t) = (f(t),t) \in \mathbb{R}^p \times T$,

$$\|(f,t)\| = \|f(t)\|_p + |t|, \tag{3.22}$$

where $f : [a_f, b_f] \to \mathbb{R}^p$ and $\|\cdot\|_p$ is the norm on vector space $\mathbb{R}^p$. This is a norm which is proved by Th.A.3. Notice that this norm $\|(\cdot,\cdot)\|$ is different from the uniform norm $\|\cdot\|$ due to its domain. Then we have the following theorem which proves the equivalence between Skorohod distance and Fréchet distance.

Theorem 3.4.1 *Let $x : [a_x, b_x] \to \mathbb{R}^p$ and $y : [a_y, b_y] \to \mathbb{R}^p$ be two polygonal curves. Then we have*

$$d^*_{Skorohod}(x,y) = d_F(x,y).$$

Proof Firstly, notice that x and y are functions defined from $[0,1]$ to $\mathbb{R}^p$, the norm of $x(t)$ and $y(s)$ is $\|x(t)\|_p$ and $\|y(s)\|_p$ correspondingly. Then, we prove that $d^*_{\text{Skorohod}}(x,y) \leq d_{\text{F}}(x,y)$, and $d_{\text{F}}(x,y) \leq d^*_{\text{Skorohod}}(x,y)$.

1. $d^*_{\text{Skorohod}}(x,y) \leq d_{\text{F}}(x,y)$. The time distortion is defined as ϕ_x, ϕ_y in Eq.(3.21). Then,

we use the $\lambda : [a_x, b_x] \to [a_y, b_y]$ defined as $\lambda(t) = \phi_y(\phi_x^{-1}(t))$. Then,

$$\begin{aligned}
\|\lambda - I\| + \|x - y\lambda\| &= \sup_t \left|\phi_y\left(\phi_x^{-1}(t)\right) - t\right| + \sup_t \left\|x(t) - y\left(\phi_y(\phi_x^{-1}(t))\right)\right\|_p \\
&= \max_\theta \left|\phi_y(\theta) - \phi_x(\theta)\right| + \max_\theta \left\|x\left(\phi_x(\theta)\right) - y\left(\phi_y(\theta)\right)\right\|_p \\
&\text{where } \phi_x^{-1}(t) \text{ is denoted as } \theta. \\
&= \max_\theta \left(\left\|x\left(\phi_x(\theta)\right) - y\left(\phi_y(\theta)\right)\right\|_p + \left|\phi_x(\theta) - \phi_u(\theta)\right|\right) \\
&= \max_\theta \left\|\left(x, \phi_y(\theta)\right) - \left(y, \phi_y(\theta)\right)\right\| \\
&= \max_\theta \left\|x\left(\phi_x(\theta)\right) - y\left(\phi_y(\theta)\right)\right\|.
\end{aligned}$$

Thus, for every ϕ_x, ϕ_y, it is naturally to have a warping function that such that $\|\lambda - I\| + \|x - y\lambda\| = \max_\theta \left\|x\left(\phi_x(\theta)\right) - y\left(\phi_y(\theta)\right)\right\|$. Hence $d^*_{\text{Skorohod}}(x, y) \le d_{\text{F}}(x, y)$.

2. $d_{\text{F}}(x, y) \le d^*_{\text{Skorohod}}(x, y)$. The time warping function is defined as $\lambda : [a_x, b_x] \to [a_y, b_y]$. Let $\phi_x : [0, 1] \to [a_x, b_x]$ defined as $\phi_x(\theta) = (1 - \theta)a_x + \theta b_x$. Let $\phi_y : [0, 1] \to [a_y, b_y]$ defined as $\phi_y(\theta) = \lambda\left((1 - \theta)a_x + \theta b_x\right)$. We have $\phi_y(\theta) = \lambda\left(\phi_x(\theta)\right)$. Both ϕ_x and ϕ_y satisfy the condition of a warping function in generalized Skorohod distance. Then,

$$\begin{aligned}
\max_\theta \left\|x\left(\phi_x(\theta)\right) - y\left(\phi_y(\theta)\right)\right\| &= \max_\theta \left\|\left(x, \phi_y(\theta)\right) - \left(y, \phi_y(\theta)\right)\right\| \\
&= \max_\theta \left(\left|\phi_y(\theta) - \phi_x(\theta)\right| + \left\|x\left(\phi_x(\theta)\right) - y\left(\phi_y(\theta)\right)\right\|_p\right) \\
&= \max_\theta \left|\phi_y(\theta) - \phi_x(\theta)\right| + \max_\theta \left\|x\left(\phi_x(\theta)\right) - y\left(\phi_y(\theta)\right)\right\|_p \\
&\text{where } \phi_x(\theta) \text{ is denoted as } t. \\
&= \sup_t \left|\phi_y\left(\phi_x^{-1}(t)\right) - t\right| + \sup_t \left\|x(t) - y\left(\phi_y(\phi_x^{-1}(t))\right)\right\|_p \\
&= \|\lambda - I\| + \|x - y\lambda\|.
\end{aligned}$$

Thus, for every warping function λ, it is naturally to have ϕ_x, ϕ_y such that

$$\max_{\theta} \left\| x\left(\phi_x(\theta)\right) - y\left(\phi_y(\theta)\right) \right\| = \|\lambda - I\| + \|x - y\lambda\|.$$

Hence $d_{\mathrm{F}}(x, y) \leq d^*_{\mathrm{Skorohod}}(x, y)$.

As a result, in order to compute the generalized Skorohod distance, we can use Fréchet distance by taking the norm defined in Eq.(3.22). To fit into the discrete situation, we are using the discrete Fréchet distance derived from Eq.(3.21) to estimate $d_{\mathrm{F}}(\cdot, \cdot)$. Let X and Y be two time series defined in Eq.(2.1), then the discrete Fréchet distance d^*_{dF} is defined as

$$d^*_{\mathrm{dF}}(X, Y) = \min_{\phi}\{\max_{i} \left\| X - Y^*, t_i - s^*_i \right\|\} = \min_{\phi}\{\max_{i}(\|x_i - y^*_i\|_p + |t_i - s^*_i|)\} \tag{3.23}$$

Here the time warping function ϕ is defined as $\phi(t_i) = s^*_i$, and with conditions that $t_i \leq t_{i+1}$ and $s^*_i \leq s^*_{i+1}$. The discrete Fréchet distance is selecting the minimum between all alignments of pairs. An algorithm for computing the above discrete Fréchet distance is also in a dynamic programming scheme [26].

Algorithm 4: Discrete Fréchet distance

Data: Query $X = \{X_i = (x_i, t_i) | i = 1, \ldots, m\}$, Reference $Y = \{Y_j = (y_i, s_j) | i = 1, \ldots, n\}$

Input: Local distance metric, $d(X, Y) = \|X - Y, t - s\|$ in Eq.(3.22)

Output: Distance, $d^*_{\mathrm{dF}}(X, Y) = ca(m, n)$, in which ca is the local distance matrix.

Function `c(`i, j`)`**:**

- **if** $c(i, j) > -1$ **then**
  - **return** $ca(i, j)$;
- **else if** $i = 1$ **and** $j = 1$ **then**
  - $ca(i, j) := d(X_1, Y_1)$
- **else if** $i > 1$ **and** $j = 1$ **then**
  - $ca(i, j) := \max\{c(i - 1, 1), d(X_i, Y_1)\}$
- **else if** $i = 1$ **and** $j > 1$ **then**
  - $ca(i, j) := \max\{c(1, j - 1), d(X_1, Y_j)\}$
- **else if** $i > 1$ **and** $j > 1$ **then**
  - $ca(i, j) := \max\{\min\{c(i - 1, 1), c(i - 1, j - 1), c(i, j - 1)\}, d(X_i, Y_1)\}$
- **else**
  - $ca(i, j) := \infty$
- **return** $ca(i, j)$

for $i = 1$ **to** n **do**

- **for** $j = 1$ **to** m **do**
  - $ca(i, j) := 1$
  - **return** $c(m, n)$

3.4.2 Comparison with TWED

In this section, we will be discussing the relationship between TWED and the generalized Skorohod distance in a special case. The objective of the elastic distance based on a dynamic programming scheme is to find the optimal discrete alignment between two time series, $X = \{X_i = (x_i, t_i)\}$ and $Y = \{Y_j = (y_j, s_j)\}$, by minimizing the cost of the matching function $\Phi(\cdot, \cdot)$,

$$\sum_i d(X_i, Y_i^*).$$

Here the $Y_i^* = (y_i^*, s_i^*)$ is a sequence aligned from the optimal path. That is to say the following mapping $\phi(t_i) = s_i^*$ in the warping path Φ. The local distance matrix is usually L1 and L2 norm. In order to compare with Skorohod distance, we use the L1 norm for the univariate situation in the following analysis, e.g., $d(x, y) = |x - y|$ and $\|x\|_p = |x|$ when $p = 1$. Besides, it is natural to generalize the TWED to the multivariate time series by choosing proper local distance $d(\cdot, \cdot)$ defined for elements in vector space $\mathbb{R}^p$.

Notice both TWED and the generalized Skorohod distance have the timestamp information included. A special case for TWED is matching every component by its diagonal when choosing $\lambda = 0$ and $\nu = 1$. By adding the timestamp difference penalty to the absolute distance, the distance cost of TWED becomes

$$\sum_i |x_i - y_i^*| + |t_i - s_i^*| = \sum_{i=1}^{\max(m,n)} |x(t_i) - y(\phi(t_i))| + |t_i - \phi(t_i)|.$$

We can apply the generalized Skorohod distance, $d^*_{\text{Skorohod}}(\cdot, \cdot)$, to the time series with a timestamp, which is the discretization of an intrinsic stochastic process. The modified version of distance becomes:

$$\inf_{\phi}\{\sup_{i}|t_i-\phi(t_i)|+\sup_{i}|x(t_i)-y(\phi(t_i))|\}.$$

Then we have the following inequality,

$$\begin{aligned}\frac{1}{\max(m,n)}\sum_{i=1}^{\max(m,n)}|x(t_i)-y(\phi(t_i))|+|t_i-\phi(t_i)| &\leq \left(\sup_{i}|x(t_i)-y(\phi(t_i))|+|t_i-\phi(t_i)|\right)\\ &\leq \left(\sup_{i}|x(t_i)-y(\phi(t_i))|+\sup_{i}|t_i-\phi(t_i)|\right).\end{aligned}$$

So, the 'average' TWED distance with time penalty is a lower bound of the summation of the uniform norm in both time and sample value spaces, which means

$$\frac{1}{\max(m,n)}\sum_{i=1}^{\max(m,n)}|x(t_i)-y(\phi(t_i))|+|t_i-\phi(t_i)| \leq \inf_{\phi}\left(\sup_{i}|x(t_i)-y(\phi(t_i))|+\sup_{i}|t_i-\phi(t_i)|\right).$$

The generalized Skorohod distance is the infimum, which is the largest lower bound among $(\sup_i|x(t_i)-y(\phi(t_i))|+\sup_i|t_i-\phi(t_i)|)$. The TWED will provide a more dense grid when mapping the time series information in $\mathbb{R}^p\times T$ to $\mathbb{R}^+$.

3.5 Summary

This chapter introduces the definition, calculation and their relationship with each other. Three types of elastic distance are introduced in this chapter, which are DTW, DTW with window and TWED. Also, taking the time information into consideration, the Skorohod distance are also introduced, which can be calculated by discrete Fréchet distance. Both theoretical analysis and empirical simulations are conducted to all of these distances. They will be applied to the real data for a thorough comparison.

Chapter 4

4 Experiments and Discussion

This chapter lists the comparison between the elastic distance performance in UCR time series database and two astrophysics datasets, photometric LSST light flux observations and Gravitational Wave (GW) strain time series of each event observed in LIGO Hanford detector's Operation Run 1 and 2. In Section 4.1.1, the UCR dataset is used to compare the performance between discrete Fréchet distance and the TWED distance in general cases in order to check their performance. Then we apply the TWED in LSST datasets, with Euclidean and DTW distance for a comparison in classification results in section 4.1.2.

For the data related to the GW detection, we introduce the relevant ideas and detection mechanism in LIGO, Virgo and KAGRA Collaboration at the beginning of Section 4.2, including a direct application of DTW distance to the real data. For the strain time series of events in LIGO Hanford detectors, we establish a new pipeline in Section 4.2.2 which gives an early warning of the arrival of the detection by giving the time region including possible GW waveform by the idea of coincidence. The pros and cons of this early warning pipeline are listed with the discussion on the non-detected events.

Apart from the published dataset, a simulated dataset of template bank of GW detection is also generated to compare the performance of the elastics distance in shrinking the size of the template bank, which is discussed in Section 4.2.1. We discuss the potential of applying the shrunk template bank to the hierarchical detection pipeline, which requires our work in the future.

4.1 Application in Classification

As we already know, 1-Nearest Neighbor (1-NN) classification with DTW achieves state-of-art performance. We apply the 1-NN approach to test the performance of the distance. Let $\{\mathbf{X}_n | 1 \leq n \leq N\}$ be the test dataset and $\{\mathbf{Y}_m | 1 \leq m \leq M\}$ be the training dataset. The class for each object is denoted as $C(\mathbf{X}_n)$. The evaluation pipeline is divided into two parts.

Firstly, we calculate the distance between samples. For distance without any parameter, such as Euclidean distance, DTW and Fréchet distance, we directly calculate the distance between samples from test and training datasets, $d(\mathbf{X}_n, \mathbf{Y}_m)$. For distance with parameters, which are DTW with window and TWED, we follow the leave-one-out cross-validation (LOOCV) in the training datasets to find the optimal parameters corresponding to minimal classification error. Then we calculate the distance between samples from test and training datasets given the optimal parameters.

Secondly, we list the nearest neighbour for each query time series. When there is a tie of the minimal distance between multiple reference time series objects, e.g., $d(\mathbf{X}_n, \mathbf{Y}_{m_1}) = d(\mathbf{X}_n, \mathbf{Y}_{m_2})$ which are both minimal distance, we randomly choose one among those reference time series for simplicity and generality. Predicted by comparing with the training set, the predicted class is denoted as $C^*(\mathbf{X}_n)$. The accuracy is computed by the indicator function $\mathbb{1}(\cdot)$ as follows,

$$\text{Accuracy} = \sum_{n=1}^{N} \mathbb{1}\left(C^*(\mathbf{X}_n) = C(\mathbf{X}_n)\right). \tag{4.1}$$

For a specific distance measure, the larger accuracy provides evidence of a better performance in similarity comparison between two time series in one dataset. We implement one package dtwCpp including the distance algorithm written in Cpp in order to parallelize the

classification pipeline for faster speed.

4.1.1 Baseline: UCR Database

The UEA & UCR Time Series Classification repository is an important database in the domain of Time Series data mining due to its abundance, diversity and authenticity [23].

When comparing the performance of each classifier over all datasets in UCR, we rank the accuracy of each classifier for each dataset in descending order and compare it with the critical difference, CD, defined as follows

$$\mathrm{CD} = q_\alpha \sqrt{\frac{k(k+1)}{6N}}, \tag{4.2}$$

where k is the number of classifiers, N is the number of datasets, and critical values q_α are based on the Studentized range statistic divided by $\sqrt{2}$ as show in Table 5(a) of [24]. We draw the critical difference diagram [24] according to the classification accuracy with significance level $\alpha = 0.05$ to visualize the performance between distance metric performance in 1-NN in Figure 4.1. In the case of 69 datasets from the UCR database, the CD is 0.69.

A lower average rank in descending order means better accuracy among all classifiers. The black solid bars indicate cliques, within which there is no significant difference in rank. Tests are performed with the sign rank test using the Holm correction [24]. The TWED has the leading rank compared with the other four. It outperforms DTW with optimal window w and has an insignificant difference compared with DTW with full window. Though the discrete Fréchet distance has the biggest average accuracy rank, it still outperforms others on a few datasets, such as Coffee, ECG200, MoteStrain, OliveOil, and so on. According to the analysis among these datasets [2, 23], it might be because of the sensitivity of Fréchet distance to the local features, e.g, the sudden changes in the slope, taking ECG200 and

MoteStrain as examples shown in Figure 4.2. These outstanding accuracy ranks indicate the potential of discrete Fréchet distance in certain patterns, e.g., curves showing discontinuity. We will apply these elastic distances in the following astrophysical datasets.

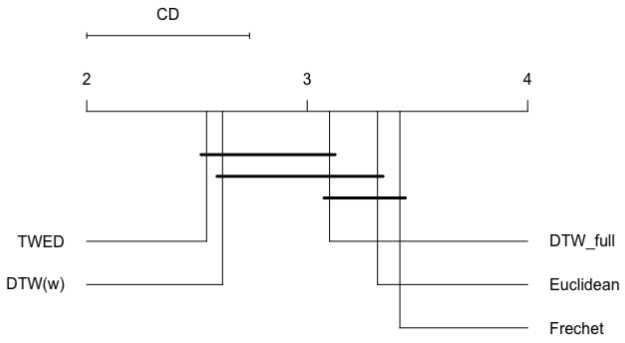
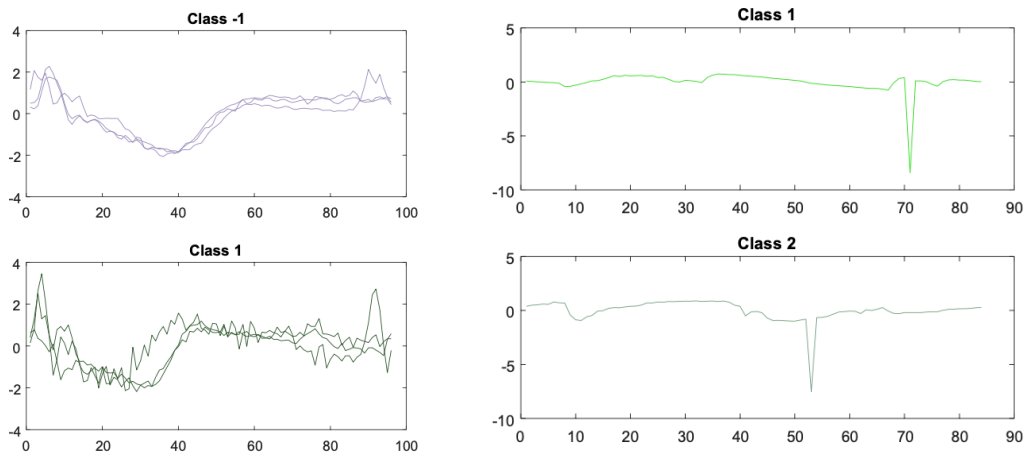


Figure 4.1: The critical difference plot to compare the average accuracy rank in descending order between distance metrics.



(a) ECG200 (b) MoteStrain

Figure 4.2: Examples from two datasets, ECG200 and MoteStrain, from UCR database [2]. In ECG200, Class 1 shows the frequent change in slope with a bumpy curve while Class -1 is relatively smooth. In MoteStrain, both classes show a sudden decrease but at different location.

Table 4.1: Table of accuracy by 1-NN on each test dataset. Four distance evaluation methods are applied. DTW with window and TWED are using trained best parameters in evaluating the test datasets. The first row lists the average of the accuracy rank in descending order.

	Euclidean	DTW(w)	DTW_full	Fréchet	TWED
Average Accuracy Rank (Descending)	3.32	2.62	3.10	3.42	2.54
Adiac	0.61	0.61	0.60	0.58	**0.62**
ArrowHead	**0.80**	**0.80**	0.70	0.69	0.79
Beef	**0.67**	**0.67**	0.63	0.63	**0.67**
BeetleFly	**0.75**	0.70	0.70	0.70	**0.75**
BirdChicken	0.55	0.70	0.75	0.75	**0.8**
BME	0.83	**0.97**	0.89	0.84	0.82
Car	0.73	0.77	0.73	0.65	**0.85**
CBF	0.85	0.99	**1.00**	0.97	**1.00**
Chinatown	0.95	0.95	**0.97**	0.95	0.94
Coffee	0.65	0.65	0.65	**0.96**	0.63
DiatomSizeReduction	**1**	**1**	**1**	0.98	0.96
DistalPhalanxOutlineAgeGroup	0.94	0.94	**0.97**	0.71	0.93
DistalPhalanxOutlineCorrect	0.63	0.63	**0.77**	0.71	0.73
DistalPhalanxTW	0.72	**0.73**	0.72	0.60	**0.73**
ECG200	0.63	0.63	0.59	**0.75**	0.60
ECGFiveDays	0.50	0.49	**1**	0.72	0.56
FaceAll	0.78	0.78	**1**	0.77	0.72
FaceFour	0.83	0.83	**1**	0.56	0.98
FacesUCR	0.88	0.88	0.77	0.73	**0.89**
Fish	0.93	0.92	0.92	0.69	**0.93**

GunPoint	0.80	0.80	0.77	0.79	**0.84**
Ham	0.71	0.81	0.81	0.48	**0.81**
Herring	0.78	0.89	0.83	0.53	**0.97**
ItalyPowerDemand	0.77	0.91	0.91	0.91	**0.97**
Lightning2	0.78	0.84	0.82	0.72	**0.91**
Lightning7	0.68	0.68	**0.76**	0.48	0.71
Meat	0.84	0.84	0.80	0.88	**0.92**
MedicalImages	0.91	0.91	0.91	0.68	**0.96**
MiddlePhalanx-OutlineAgeGroup	0.97	**1.00**	0.98	0.52	0.98
MiddlePhalanx-OutlineCorrect	0.99	0.99	0.98	0.70	**1.00**
MiddlePhalanxTW	**1.00**	**1.00**	**1.00**	0.46	**1.00**
MoteStrain	0.60	0.60	0.47	**0.76**	0.51
OliveOil	0.52	0.53	0.53	**0.77**	0.48
OSULeaf	0.68	**0.84**	**0.84**	0.52	0.79
Plane	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
ProximalPhalanx-OutlineAgeGroup	**1.00**	**1.00**	**1.00**	0.78	**1.00**
ProximalPhalanxOutlineCorrect	0.56	0.59	0.36	**0.79**	0.57
ProximalPhalanxTW	**0.96**	**0.96**	0.95	0.71	0.95
ShapeletSim	0.75	**0.87**	**0.87**	0.57	0.83
SonyAIBORobotSurface1	0.58	0.71	0.73	0.68	**0.74**
SonyAIBORobotSurface2	**0.93**	**0.93**	**0.93**	0.77	**0.93**
SwedishLeaf	0.68	0.75	0.74	**0.75**	0.746
Symbols	**0.95**	0.92	0.88	0.89	0.90
SyntheticControl	0.52	0.52	0.50	**0.97**	0.51

ToeSegmentation1	0.77	0.77	0.70	0.69	**0.77**
ToeSegmentation2	0.51	0.51	0.51	**0.74**	0.52
Trace	0.88	0.87	0.84	**1.00**	0.87
TwoLeadECG	0.87	0.87	0.83	**0.91**	0.83
Wine	0.52	0.61	0.59	0.54	**0.79**
WordSynonyms	0.96	**1.00**	**1.00**	0.53	**1.00**
DodgerLoopDay	0.41	0.40	0.39	0.29	**0.96**
DodgerLoopGame	0.79	0.79	**0.81**	0.80	0.79
DodgerLoopWeekend	0.81	0.79	0.78	**0.87**	0.78
FreezerSmallTrain	0.71	0.76	0.76	**0.81**	0.70
Fungi	**0.64**	**0.64**	0.54	0.57	0.62
GunPointAgeSpan	0.54	0.70	0.65	**0.98**	0.87
GunPointMaleVersusFemale	0.95	0.97	0.83	0.96	**1.00**
GunPointOldVersusYoung	0.70	0.70	0.73	**1.00**	0.69
HouseTwenty	0.86	0.86	0.83	0.54	**0.87**
InsectEPGRegularTrain	0.79	0.85	0.79	**1.00**	0.89
InsectEPGSmallTrain	0.90	0.94	0.95	**1.00**	0.84
MelbournePedestrian	0.88	0.98	**0.993**	0.84	0.98
PowerCons	0.68	0.75	0.77	**0.85**	0.81
Rock	0.81	0.91	0.84	0.50	**0.95**
SmoothSubspace	0.76	0.99	**1.00**	0.63	0.89
UMD	0.75	0.87	0.90	0.88	**0.93**
ChlorineConcentration	0.79	**0.92**	0.85	0.64	0.76
ECG5000	0.61	0.61	0.57	**0.91**	0.65
InsectWingbeatSound	0.62	0.74	0.65	0.21	**0.77**

4.1.2 LSST Light Flux Time Series Classification

The LSST light flux time series dataset is a simulated dataset [37] which was included in a Kaggle competition [57] as a rehearsal of the observations from the Large Synoptic Survey Telescope (LSST). One sample of the time series data with multiple channels is shown in Figure 2.3 in Chapter 2. The objective is to classify astronomical objects by their variation in brightness in multiple passbands. In this particular dataset, a time series object is denoted as $\mathbb{X}_n = \{\mathbb{X}_{1,n}, \mathbb{X}_{2,n}, \ldots, \mathbb{X}_{6,n}\}$, corresponding to a. collection of the univariate time series $\mathbb{X}_{k,n}$ in each passband, where $k = 1, \ldots, 6$. Here passband index k matches the wavelength range *ugrizy* of the filter passband.

To deal with the multivariate time series shown as the light flux from multiple filtering passbands, a loosely univariate classification scheme is applied. For each pair of the objects, $(\mathbf{X}_n, \mathbf{Y}_m)$, we compute the classification error in each dimension of the time series, e.g., compute the distance $d(\mathbf{X}_{1,n}, \mathbf{Y}_{1,m})$ between two univariate time series. Then we apply the 1-NN to find the nearest neighbour class $C^*(\mathbf{X}_n)$ in each passband dimension, denoted as $\{C^*(\mathbf{X}_{1,n}), \ldots, C^*(\mathbf{X}_{6,n})\}$. Notice that the query time series has various patterns among the passband, resulting in a multiple label in the predicted class tuple. E.g., $C^*(\mathbf{X}_{1,n}) = 16$ means $\mathbf{X}_{1,n}$ has a similar time series pattern matched with one object of eclipsing binary stars in passband u, while $C^*(\mathbf{X}_{2,n}) = 88$ means $\mathbf{X}_{2,n}$ has a similar time series pattern matched with one object of active galactic nuclei in passband u. As a tuple of 6 dimensions, two approaches are raised to derive the class. The first approach is to using the mode of $\{C^*(\mathbf{X}_{1,n}), \ldots, C^*(\mathbf{X}_{6,n})\}$,

$$C^*_{\text{Mode}}(\mathbf{X}_n) = \text{Mode}(\{C^*(\mathbf{X}_{1,n}), \ldots, C^*(\mathbf{X}_{6,n})\}).$$

Then we compute the accuracy defined in Eq.(4.1) by $C^*_{\text{Mode}}(\mathbf{X}_n)$.

Distance	Loose Accuracy	Mode Accuracy
Euclidean	0.73	0.42
DTW_full	0.72	0.42
DTW(w)	0.73	0.41
TWED	**0.74**	**0.43**

Table 4.2: Accuracy comparison between four distances with 1-NN.

$$\text{Mode Accuracy} = \sum_{n=1}^{N} \mathbb{1}\left(C^*_{\text{Mode}}(\mathbf{X}_n) = C(\mathbf{X}_n)\right). \tag{4.3}$$

Another approach is that of loosely matching. As long as one of the predicted classes among $\{C^*(\mathbf{X}_{1,n}), \ldots, C^*(\mathbf{X}_{6,n})\}$ matches the real class, we mark this as corrected classified. Then the loose accuracy rate becomes

$$\text{Loose Accuracy} = \sum_{n=1}^{N} \mathbb{1}\left(C(\mathbf{X}_n) \in \{C^*(\mathbf{X}_{1,n}), \ldots, C^*(\mathbf{X}_{6,n})\}\right). \tag{4.4}$$

Notice that the mode accuracy has a more strict matching condition than loose accuracy, leading to a small accuracy. Both of the loose accuracy and mode accuracy are listed in the following table.

From the result shown above, we notice that TWED outperform others in both loose accuracy and mode accuracy. Another observation is that both DTW methods, full window or optimal window, are less accurate than Euclidean distance. This is because the query and reference time series in the dataset have different lengths in the observation. The neglect of the timestamp information leads to multiple singularities in the warping path by DTW. Apart from that, the calculation of discrete Fréchet distance is limited by the problem of out-of-memory. So we exclude Fréchet distance from the LSST dataset at this moment.

4.2 Application in GW Detection

Aiming to discover the application of DTW and its variants in GW detection, this section make numerical experiments of all distances mentioned in Chapter 3 to simulated template data and the open-source GW observatory data [15]. Our application is based on one search pipeline, PyCBC [47, 58], which established the foundation of the first gravitational-wave signal detection by the LIGO-Virgo Scientific Group, event name GW150914, on September 14 2015 [13] during their first observing run (O1). We will first introduce the fundamental components and mechanisms of a standard and valid search pipeline of the gravitational wave.

Each LIGO detector collects the strain data time series $\mathbf{d}$ from the interferometer, which is written as the sum of the GW signal $\mathbf{h}$ and the noise $\mathbf{n}$, such that

$$\mathbf{d} = \mathbf{h} + \mathbf{n} \tag{4.5}$$

The GW signal $\mathbf{h}$ is usually numerically simulated under the result of the solutions of Einstein's equations with parameter $\boldsymbol{\theta}$, denoted as $\mathbf{h}(\boldsymbol{\theta})$, e.g., the Taylor family of the waveform for the spinning coalescing binaries. For most numerical models, the parameter set $\boldsymbol{\theta}$ contains signal amplitude A in the detector, phase ϕ of the sinusoidally-varying signal, masses m and spins of the components, and the arrival time t of the signal [28].

An ideal assumption on the noise $\mathbf{n}$ can be considered as independent, stationary and Gaussian noise between each independent detector. Denote the $S_{\mathbf{n}}(f)$ as the power spectral density of noise $\mathbf{n}$, which can only be estimated from data among every detector, LIGO and Virgo applied matched-filtering analyses to find the optimal parameterized GW signal $\mathbf{h}(\hat{\boldsymbol{\theta}})$. The matched filter is the optimal filter for detecting a known waveform in stationary Gaussian noise [28], defined as the following noise-weighted inner product:

$$(\mathbf{d} \mid \mathbf{h}(\boldsymbol{\theta})) = 2 \int_0^\infty \frac{\tilde{\mathbf{d}}(f)\tilde{\mathbf{h}}(\boldsymbol{\theta})^*(f) + \tilde{\mathbf{d}}^*(f)\tilde{\mathbf{h}}(\boldsymbol{\theta})(f)}{S_n(f)} \mathrm{d}f \tag{4.6}$$

This notation of noise-weighted inner product is the core of the hypothesis testing to identify candidate signals in LIGO-Virgo data, comparing null hypothesis, $\mathcal{H}_0$, that data contains noise only, to the signal hypothesis, $\mathcal{H}_1$, that data contains both noise and a gravitational-wave signal $\mathbf{h}(\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$. For simplicity, under the null and alternative hypothesis, the probability of the observed strain data $\mathbf{d}$ can be written as $p(\mathbf{d} \mid \mathcal{H}_0) = p_0(\mathbf{d})$ and $p(\mathbf{d} \mid \mathcal{H}_1) = p_1(\mathbf{d})$.

According to Baye's theorem, the probability of a detection $\mathcal{H}_1$ given the observed strain data $\mathbf{d}$, the posterior probability, is

$$p(\mathcal{H}_1 \mid \mathbf{d}) = \frac{p(\mathcal{H}_1)p_1(\mathbf{d})}{p(\mathcal{H}_0)p_0(\mathbf{d}) + p(\mathcal{H}_1)p_1(\mathbf{d})} = \frac{p_1(\mathbf{d})}{p_0(\mathbf{d})}\left[\frac{p_1(\mathbf{d})}{p_0(\mathbf{d})} + \frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)}\right]^{-1}, \tag{4.7}$$

where $p(\mathcal{H}_0)$ and $p(\mathcal{H}_1)$ are prior beliefs of whether the data includes or excludes a signal. The posterior probability is monotonic in the likelihood ratio when fixing the ratio of $\Lambda(\mathbf{d}|\boldsymbol{\theta}) = p(\mathcal{H}_0)/p(\mathcal{H}_1)$. Under the ideal assumption of Gaussian noise, with the noise-weighted inner product notation, the log of the likelihood ratio is split into two parts

$$\log \Lambda(\mathbf{d}|\boldsymbol{\theta}) = (\mathbf{d} \mid \mathbf{h}(\boldsymbol{\theta})) - \frac{1}{2}(\mathbf{h}(\boldsymbol{\theta}) \mid \mathbf{h}(\boldsymbol{\theta})) \tag{4.8}$$

The maximum value of $\Lambda(\mathbf{d}|\boldsymbol{\theta})$ over unknown parameters $\boldsymbol{\theta}$ corresponds to the optimal parameters $\hat{\boldsymbol{\theta}}$ for the template $\mathbf{h}(\boldsymbol{\theta})$. As the linearity between the term of matched-filter $(\mathbf{d} \mid \mathbf{h}(\boldsymbol{\theta}))$ and the log of the likelihood ratio $\log \Lambda(\mathbf{d}|\boldsymbol{\theta})$, the inner product $(\mathbf{d} \mid \mathbf{h}(\boldsymbol{\theta}))$ will be maximized if the optimal GW signal $\mathbf{h}(\hat{\boldsymbol{\theta}})$ exists. Based on the above derivations, a basic procedure for the detection of a signal in LIGO and Virgo can be summarized as follows [28]:

- Establish template bank with the template waveform $\mathbf{h}(\boldsymbol{\theta})$ by a multiple selection among the parameter set $\{\boldsymbol{\theta}\}$.
- Apply the matched filter to each chunk of the strain data for each template.
- Determine the optimal parameter $\hat{\boldsymbol{\theta}}$, including the detection time $\hat{t}$ from matched filter.

Besides the simplified pipeline above, there are more sophisticated methodologies and insightful analyses spreading in the numerous details of GW pipeline. For example, information between detectors is also applied in the confirmation of detection, e.g., the coincidence check which limits the time difference between possible signals in each detector to guarantee the time-shift and exclude the non-GW glitches [58]. In this section we will introduce the aspects of templates and detection, exploring the application of DTW and its variants in the GW domain by starting with a first example.

At first, we concentrate on constructing a DTW-based 'inner-product' like matched-filtering. As the first experiment of the DTW directly to the data, we found that the hidden signal from background noise in the observatory GW data is the major obstacle to applying DTW to the direct detection of a gravitational-wave signal, as shown in Figure 4.3, taking the first detected event of gravitational wave GW150914 [13] as an example. The sampling frequency of the strain data is 4096 Hz and the strain data has a duration of 4 seconds around the detection time, indicated in the red vertical line. The calculation of the DTW distance Applying sliding window of window size = 0.25 second and forward size = 0.03125 second, the DTW distance within each window is calculated between LIGO-Hanford(H1) and LIGO-Livingston(L1). The window size is chosen to cover the whole length of the signal, which is approximately 200ms [3]. The 95% confidence interval is inside two grey lines in the plot. Only a few of the data points are outside of the 95% of the confidence interval, none of them near the middle line of the detection. Indicated from the strain series from Figure 4.3, DTW and its variants are quite sensitive to the shape of both reference and

query sequences, which means a detection purely from the shape information is not enough to capture a GW signal. This first experiment also enlightens us to find in each component, signal model template $h(\boldsymbol{\theta})$, and noise $\mathbf{n}$, instead of finding a similar 'inner-product' directly to the strain data $\mathbf{d}$.

The noise-free signal template $h(\boldsymbol{\theta})$ purely depends on the shape of the template without the randomness of the background noise, leading us to apply the DTW and its variants to shrink the size of the template banks by excluding similar templates. A simple experiment is done in Section 4.2.1 by DTW distance. We can also compromise on the false alarm rate to construct the pipeline of template-free early warning of detection without the knowledge of the templates model, including the idea of 'coincidence' only, with details in Section 4.2.2.

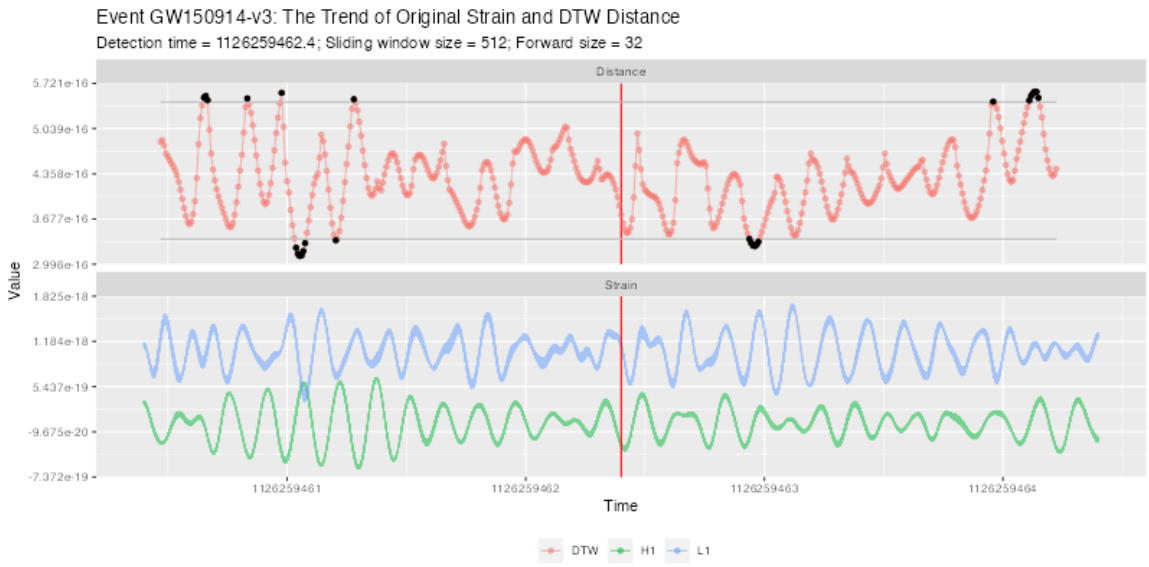


Figure 4.3: The lower graph shows the original strain data from observatory Hanford(H1) and Livingston(L1). The upper graph shows the DTW distance calculated between two strain series.

4.2.1 Shrinking the Template Bank

In this section, we will introduce the application of elastic distance in shrinking the size of the template bank, with the idea of similarity between GW signal $\mathbf{h}$, starting with the

background introduction of the generation of templates used by LIGO and Virgo in the detection of the coalescence of compact binary objects, e.g., the merger of binary black holes.

In solving Einstein's field equation of general relativity, Post-Newtonian (PN) approximation is a perturbative expansion to approximate the metric tensor around a weak gravitational field with increasing accuracy beyond the Newtonian limit [42], e.g., TaylorF2 and SpinTaylorT2 used in our simulated toy dataset. When the waveform propagates to the detectors, the strain data will show a sinusoidally-varying signal [28], expressed by the following equation with time in the detector t, signal amplitude A, the phase ϕ and physical parameters $\boldsymbol{\theta}$,

$$\mathbf{h}(t, \boldsymbol{\theta}) = A\mathbf{p}(t, \boldsymbol{\theta}) \cos\phi + A\mathbf{q}(t, \boldsymbol{\theta}) \sin\phi, \tag{4.9}$$

where $\mathbf{p}(t, \boldsymbol{\theta})$ and $\mathbf{q}(t, \boldsymbol{\theta})$ are in-phase (cosine) and quadrature-phase (sine) waveform.

The process of binary black hole merger has multiple physical parameters which expands into a 17-dimensional parameter space [18], denoted as $\boldsymbol{\Theta}$. the component masses (m_1, m_2), the component spin vectors ($\mathcal{S}_1$, $\mathcal{S}_2$), the eccentricity e and phase of perihelion γ, the right ascension and declination of the source (α, δ), the distance r, the inclination angle η, the polarization phase ϕ, the orbital phase at coalescence Φ_c and the time at coalescence t_c. When emitting as a gravitational wave under certain assumptions, e.g., non-precessing and spin-weighted spherical-harmonic mode, physicists simplify the parameters further to detect GW signals, with only remaining intrinsic parameters by mathematical simplification in maximizing the log-likelihood ratio.

As the linearity between matched-filter ($\mathbf{d} \mid \mathbf{h}(\boldsymbol{\theta})$) and the log likelihood in Eq.(4.8), the parameter $\hat{\boldsymbol{\theta}}$ is estimated with the largest matched-filter ($\hat{\boldsymbol{\theta}}|\mathbf{d}$) by filtering all candidate waveform against the real data $\mathbf{d} = \mathbf{h} + \mathbf{n}$. As the detection of GW is achieved by trial and

error, we need a so-called template bank, $\{\boldsymbol{\theta}\} \subset \boldsymbol{\Theta}$, which is a set of template waves with various parameters inside the parameter space of interest. To cover the possible GW events as much as possible, the template bank should obtain the sensitivity over the complete parameter space $\boldsymbol{\Theta}$. We can evaluate the 'completeness' of a template bank against a putative simulation of waveform $\mathbf{h}(\boldsymbol{\theta}_{\text{sim}})$ with parameter $\boldsymbol{\theta}_{\text{sim}} \in \boldsymbol{\Theta}$ before the real detection in order to find a complete and effectual template bank.

For any parameter in template bank, $\boldsymbol{\theta} \in \{\boldsymbol{\theta}\}$, we define the overlap $O(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{sim}})$ between the template $\mathbf{h}(\boldsymbol{\theta})$ against the putative simulation $\mathbf{h}(\boldsymbol{\theta}_{\text{sim}})$,

$$O(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{sim}}) \equiv \frac{\rho(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{sim}})}{\rho_{\text{opt}}(\boldsymbol{\theta}_{\text{sim}})}, \tag{4.10}$$

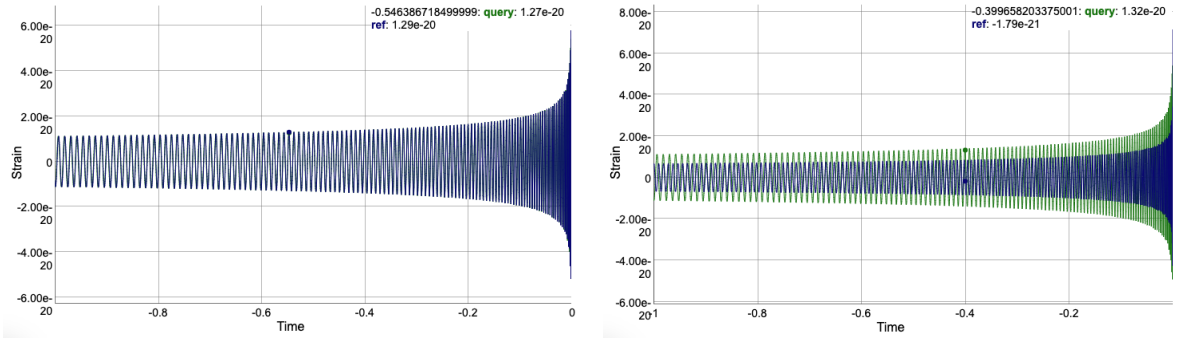
where $\rho(\cdot \mid \cdot)$ is the matched-filter signal-to-noise ratio (SNR) defined as follows

$$\rho(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2) \equiv \frac{|(\mathbf{h}(\boldsymbol{\theta}_1) \mid \mathbf{h}(\boldsymbol{\theta}_2))|}{\sqrt{(\mathbf{h}(\boldsymbol{\theta}_1) \mid \mathbf{h}(\boldsymbol{\theta}_1))}}, \tag{4.11}$$

and ρ_{opt} is optimal SNR when $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{sim}}$, meaning $\rho_{\text{opt}}(\boldsymbol{\theta}_{\text{sim}}) = \sqrt{(\mathbf{h}(\boldsymbol{\theta}_{\text{sim}}) \mid \mathbf{h}(\boldsymbol{\theta}_{\text{sim}}))}$. A larger overlap $O(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{sim}}) \in [0, 1]$ means the template is closer to the simulated waveform. As a criterion of the 'completeness' of a template bank, the 'effectualness' for any putative simulated waveform $\mathbf{h}(\boldsymbol{\theta}_{\text{sim}})$ is defined as the largest overlap between the signal against all templates

$$\mathcal{E}(\{\boldsymbol{\theta}\}) = \max_{\boldsymbol{\theta}} O(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{sim}}). \tag{4.12}$$

Due to the computational cost, a template bank with size n_θ should have enough templates. From a first check of one pair in a toy dataset of template bank as shown in Figure 4.4, we found that the template with maximal DTW distance is quite different compared to the other one in both shape and magnitude. In order to shrink the size of the template bank, we



(a) Two templates with minimal DTW

(b) Two templates with maximal DTW

Figure 4.4: Two pairs of templates with the minimal and maximal DTW distance among all pairs of templates from a template bank. The reference sequence is in blue line while the query sequence is in green. There are large differences between the templates generated.

apply DTW and its variants to the similarity between templates to shrink the size n_θ of the template bank. The procedures are:

Data Preparation Construct the template bank $\{\boldsymbol{\theta}\}$, simulated signals $\{\boldsymbol{\theta}_{\text{sim}}\}$ and the background noise from any power spectral density.

Template Evaluation For each pair of simulated signal and template, calculate the overlap $\mathcal{O}(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{sim}})$ and Effectualness $\mathcal{E}(\{\boldsymbol{\theta}\})$.

Elastic Distance Calculate the elastic distance $D(\mathbf{h}(\boldsymbol{\theta}_i), \mathbf{h}(\boldsymbol{\theta}_j))$ between each pair of templates and the total distance $D(\mathbf{h}(\boldsymbol{\theta}_i)) = \sum_{j \neq i} D(\mathbf{h}(\boldsymbol{\theta}_i), \mathbf{h}(\boldsymbol{\theta}_j))$ of one template with other rest of templates.

Shrinking Template Bank Set a threshold, e.g., 0.05, to remove those with the smallest ascending rank. Compute the Effectualness $\mathcal{E}(\{\boldsymbol{\theta}\}_{\text{shrunk}})$ of the shrunk template bank $\{\boldsymbol{\theta}\}_{\text{shrunk}}$.

For simplicity, in the toy dataset for the experiment, we consider only the one pair of intrinsic parameters of the gravitational wave to form a 2-dimensional parameter space $\boldsymbol{\Theta} = \{m_1, m_2\}$, which are component mass for the simulation and the component mass $m_1, m_2 \in [2\text{M}_\odot, 3\text{M}_\odot]$ and total masses $\text{M}_{\text{total}} \in [4\text{M}_\odot, 5\text{M}_\odot]$ with solar mass $\text{M}_\odot$. The

parameter of interest is placed in a globally flat coordinate system at 3.5 PN order [9]. With the usage of the hexagonal lattice algorithm with the ξ_i coordinate system [17] inherited in the code of non spinning bank placement, `pycbc_geom_nonspinbank` from PyCBC [9], the size of the template bank $n_\theta = 479$.

For the choice of templates, the PN model we used is TaylorF2, which is derived by expanding the phase of the gravitational waveform in terms of post-Newtonian corrections [40]. For the simulated waveform, which is desired to be as precise as possible to simulate the real waveform, we choose SpinTaylorT4. It implements the spin-orbit and spin-spin interactions to higher post-Newtonian orders [40], meaning a higher precision compared with TaylorF2. Two models have the same setting in phase order=3 and spin order=3.

The 10 putative simulations of the waveform are uniformly injected into the 10 seconds of background noise. We generate the background noise from a PSD curve, which is labelled as aLIGOAdVO4T1800545 from PyCBC module [10]. This PSD curve shows the signal's power distribution versus the frequency of the background noise for multiple localization areas [5], during the operation run O4 of Advanced LIGO and Advanced Virgo, which has its first period O4a done from 15:00 UTC 24 May 2023, and ended at 16:00 UTC 16 January 2024 [6, 29].

We list the distribution of m_1 and m_2 with different sizes indicating the value of overlap, the distribution of overlap against chirp mass $\mathcal{M} = M_{\text{total}}(m_1 m_2 / M_{\text{total}}^2)^{3/5}$ and the distance for the templates to be removed with three distance, Euclidean, DTW and TWED. TWED distance has the default setting of penalty $\lambda = 0.1$ and stiffness $\nu = 0.1$. The total distance of the template is 20 seconds with a sampling frequency of 4096 Hz. Due to the computational complexity of DTW and TWED distance, we compute the elastic distance on the last 1 second of data, with 4096 data points, which includes the part of the inspiral and merger phase when parameters only vary in component masses.

According to our simulation result, by setting the threshold of 5% size of shrinkage, the

effectualness of the shrunk template doesn't change for each simulated signal for TWED and Euclidean distance as shown in Figure 4.5 and Figure 4.7. However, there are two issues require further discussion in the application of DTW distance.

For DTW distance in Figure 4.6, the template with parameters $\{m_1 = 2.857\mathrm{M}_\odot, m_2 = 2.104\mathrm{M}_\odot\}$ has the smallest total distance against all other templates but also has a high level of overlap and achieves the largest overlap or the effectualness for one simulated template. One possible reason for this contradiction for this template is due to the inverse exponential relationship between overlap and total DTW distance shown in the right corner graph of Figure 4.6. A small total distance has large varying values of overlap, which cannot be perfectly distinguished inside them. Besides that, another possible reason is the closely distributed simulated waveform as shown in Figure 4.8: two simulation waveforms with dark cyan have similar shapes and magnitudes which both match the template with high overlap values. This corresponds to the cluster of the effectualness pair of template and simulation waveform in Figure 4.6.

Another interesting phenomenon is about the TWED distance shown in Figure 4.7: the distribution of overlap against total TWED distance and against the chirp mass $\mathcal{M}$ are vertically mirror-flipped. This might be because of adding the time domain information in TWED, which adds the phase information between each template. TWED distance has better separation of templates among the group of low total distance and the group of largest overlap compared with DTW. This requires further simulation on a larger template bank and properly distributed simulated waveform.

Other criteria are also available for the discussion of the effectualness of template bank, e.g., minimal match, which requires that a template bank such that any simulated waveform has an effectualness bigger than the minimal match [18], ending up with an effectual template bank. In the practice of LIGO and Virgo templates for detecting GW signals, the minimal match is usually set as 97%. Applying this idea to our experiment, most of the excluded

templates are located in this region, which again reveals the effectiveness of similarity comparison in shrinking the template bank. However, due to the toy template bank we are using, this requires further verification in large-scale simulation.

As shown above, this elastic distance idea is valid in describing the internal structure inside the template bank, which can also be applied in the detection pipeline in a hierarchical approach. As the distance information reveals partially the property of the template, we first compare the signal with those templates with higher total elastic distance and go to the rest of the templates with less total elastic distance, until we find the desired optimal signals or iterate over all templates inside an effectual template bank. The possible drawbacks of this hierarchical model are the usage of shape information in the time domain, the high computational complexity of DTW/TWED as the element-by-element comparison in time series [45], and the lack of physical domain knowledge. It is promising to apply other elastic distance, e.g., fastDTW with almost linear complexity [51], and combining with the physical mechanism of templates, e.g., two-stage hierarchical search pipeline by searching on a smaller and coarser template bank first and then search over a finer parameter space from the first stage search [43, 55] to overcome those drawbacks.

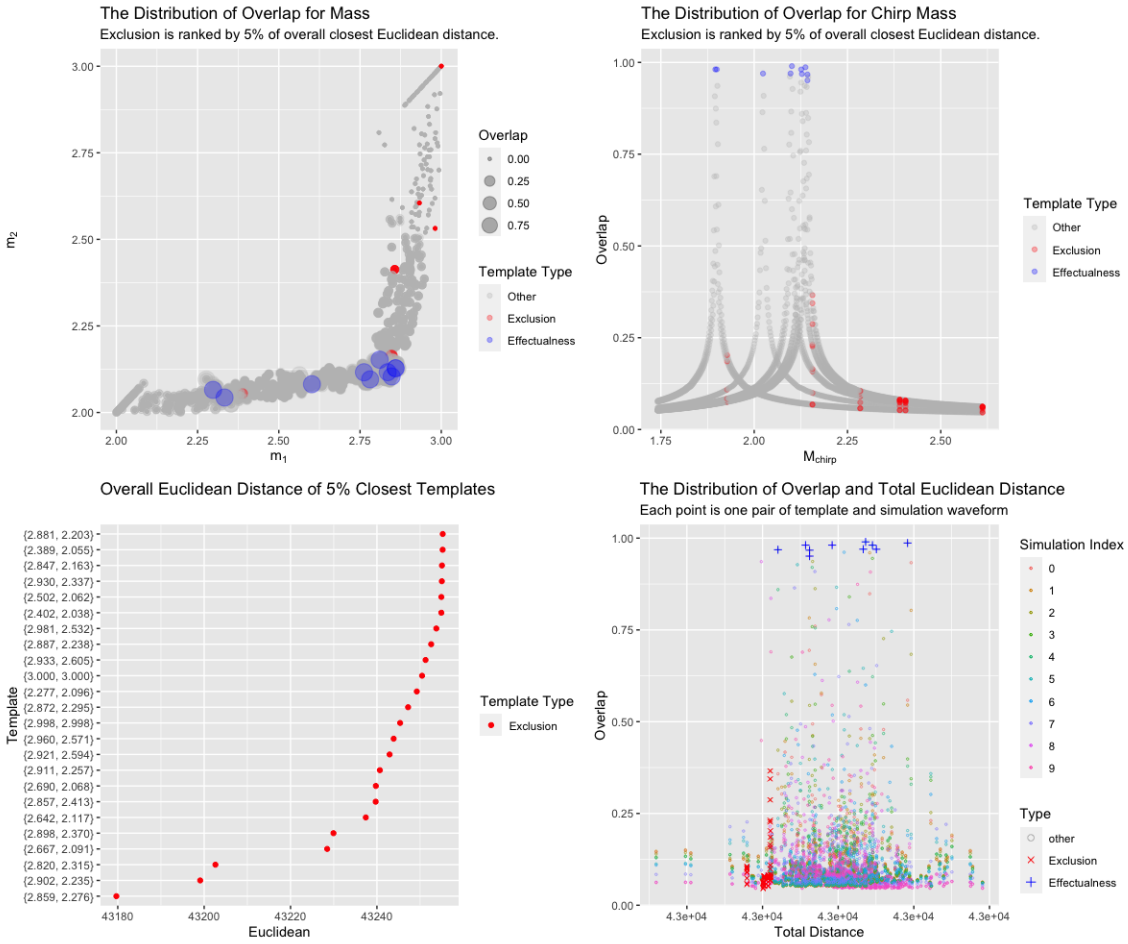


Figure 4.5: Application of Euclidean distance to the template bank and comparison with matched-filter overlap. In the above two distribution graphs, each point represents the matched-filter overlap between one template and one simulated wave. Those templates to be excluded gather around the tail of the larger chirp distribution. The templates which have the largest overlap, the effectualness $\mathcal{E}(\{\boldsymbol{\theta}\})$, are in blue and the templates which have the closest total distance inside the bank are in red. In the lower right corner, the graph shows the distribution between the overlap value and the total Euclidean distance, which is Gaussian-like distributed.

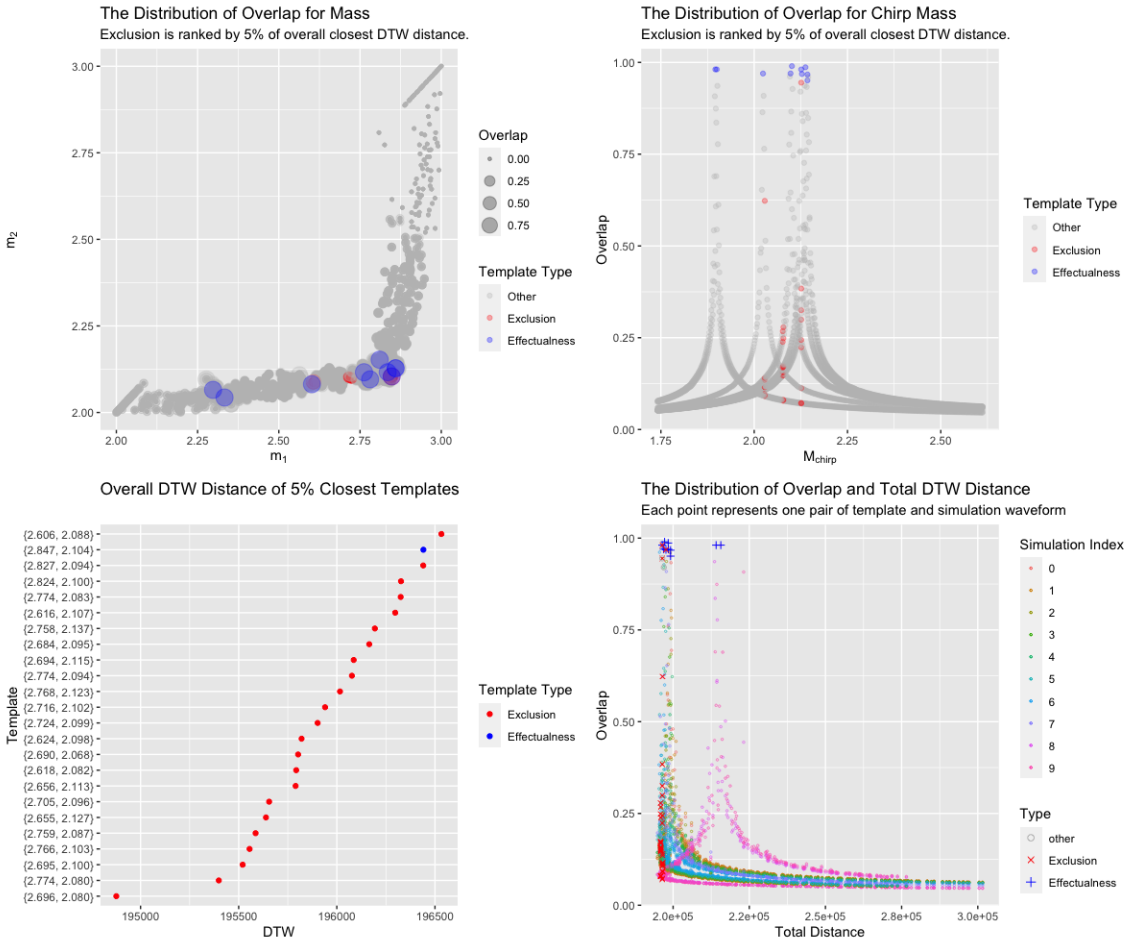


Figure 4.6: Application of DTW distance to the template bank and comparison with matched-filter overlap. There is an overlapping in the templates to be excluded (red) and the templates to have the largest overlap (blue), which is abnormal. The parameter of this template is labelled in the following graphs which have a quite small total distance. The graph in the lower right corner shows the distribution between the overlap value and the total Euclidean distance, which shows an inverse exponential of the total DTW distance against the overlap value.

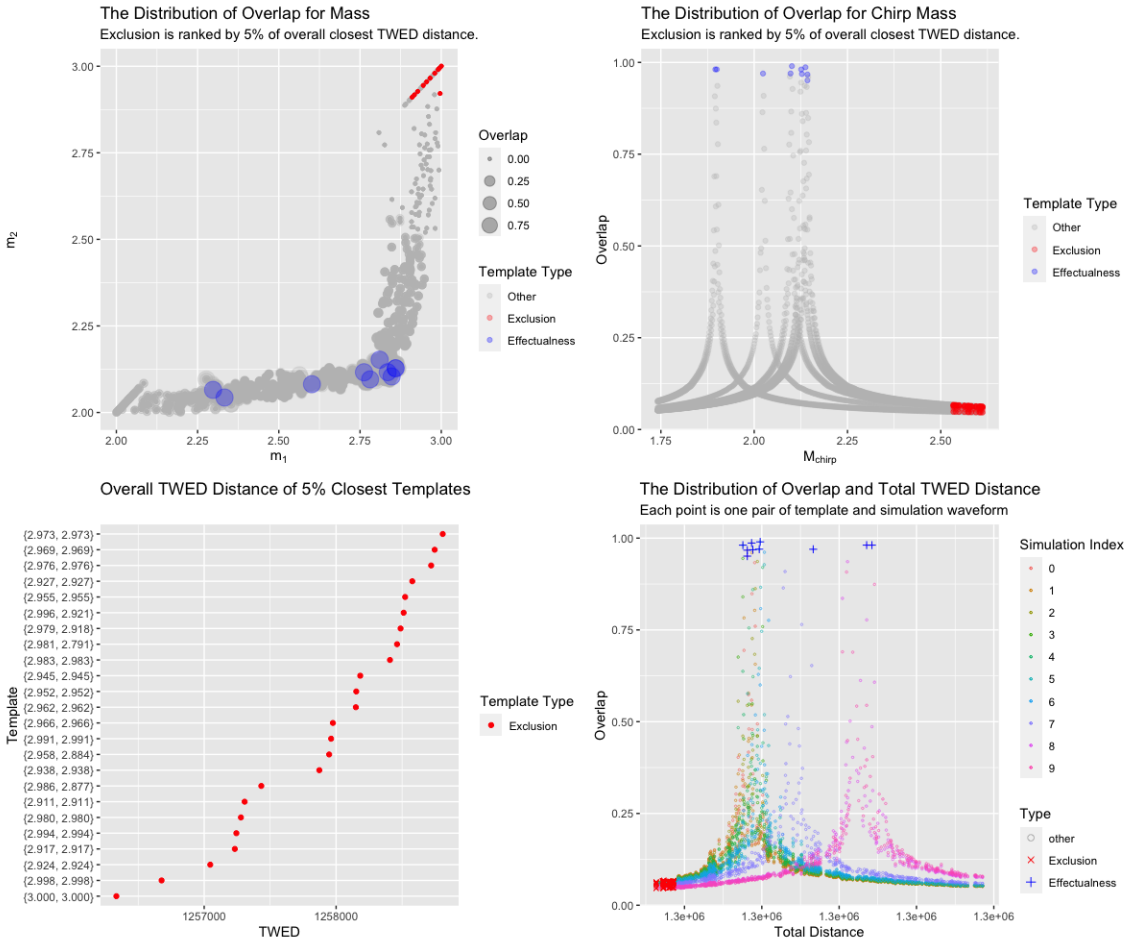


Figure 4.7: Application of TWED distance to the template bank and comparison with matched-filter overlap. A similar vertically mirror-flipped shape can be found in the distribution of overlap against chirp mass and the total TWED distance.

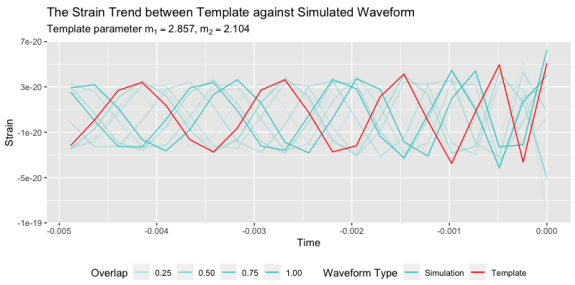


Figure 4.8: The strain trend of the abnormal template shown in Figure 4.6 against 10 simulated waveform in the last 0.005 seconds. There are two possible injections with high overlap values shown as two darker cyan-blue lines. Though there are time lags between them and the template which can be eliminated by estimating detection time, the higher overlap between the abnormal template and these two simulated waveforms is due to the similar shape at the end of the waveform, representing the merger phase.

4.2.2 Early Warning of a Detection

Due to the obstacle of direct detection of GW by DTW, we establish an early warning pipeline of the detection without the prior knowledge of the waveform based on the coincidence signal among the network of detectors. For example, the distance of 3,002 km between LIGO Hanford and LIGO Livingston gives the estimation of the longest time of gravitational wave signal arrival between two sites, which is about 10 milliseconds. In the practice of detection pipeline, PyCBC, any similar episodes larger than 10 15 milliseconds should be ignored for further analysis: 10 milliseconds travel time between LIGO Hanford and LIGO Livingston and 5ms buffer left for timing errors [13, 58]. As assumed Gaussian background noise, the distance of DTW and its variants $D(X, Y)$ follow the Gaussian distribution approximately according to Chapter 3 Section 3.3.1.

We apply Euclidean distance, DTW distance without window, Fréchet distance and TWED distance to the centralized data of the original strain data from two detectors on the first category of confident detection of GW signals during the O1, Gravitational-Wave Transient Catalog-1 (GWTC-1) [27]. This dataset includes an overlapped time period between the first two LIGO's interferometer detectors located at Hanford and Livingston. In the GWTC-1, 11 episodes of data were recorded 4096 seconds near the coincident observations from 2015 September to 2017 August. The early warning pipeline conducted the following procedures:

Data Aggregation Split and centralize the strain data in each 60s chunk from overlapped time period among multiple detectors.

Elastic Distance Apply each distance to the pair of detectors within 4 seconds chunk with a fixed size of sliding window and forward size. Also, calculate each distance of the nearby 30-second data as baseline distribution.

Hypothesis Test Apply the two-sided t-test with significance level $\alpha = 0.0001$ to each

chunk of DTW, e.g. 20 distance $D(X, Y)$. Record the consecutive significant time region which lasts for 1 second.

Performance Evaluation For any random period, we evaluate the significant duration ratio. For GWTC-1 data with confident detection, we evaluate the detection ratio of whether the signal is included inside a significant time region.

Hypothesis testing is applied to determine which period of time contains possible signals. Null hypothesis H_0 is given when there is only noise included in the window period so the distance $D(X, Y)$ follows a Gaussian distribution. This is a simplified condition from Chapter 3 Section 3.3.1 under the assumption that noise follows a Gaussian distribution and tiny shift when calculating DTW distance. Alternative hypothesis H_1 means that there is a 'coincident' signal included in the data, $D(X, Y)$ be deviated from the mean of the noise-based distance distribution. So we need the baseline distribution of the 30 seconds data adjacent to the 4 seconds chunk for detection. This baseline distribution also excludes the outliers based on the rule of 1.5 inter-quartile range because of a high false-alarm rate when only shape information is considered.

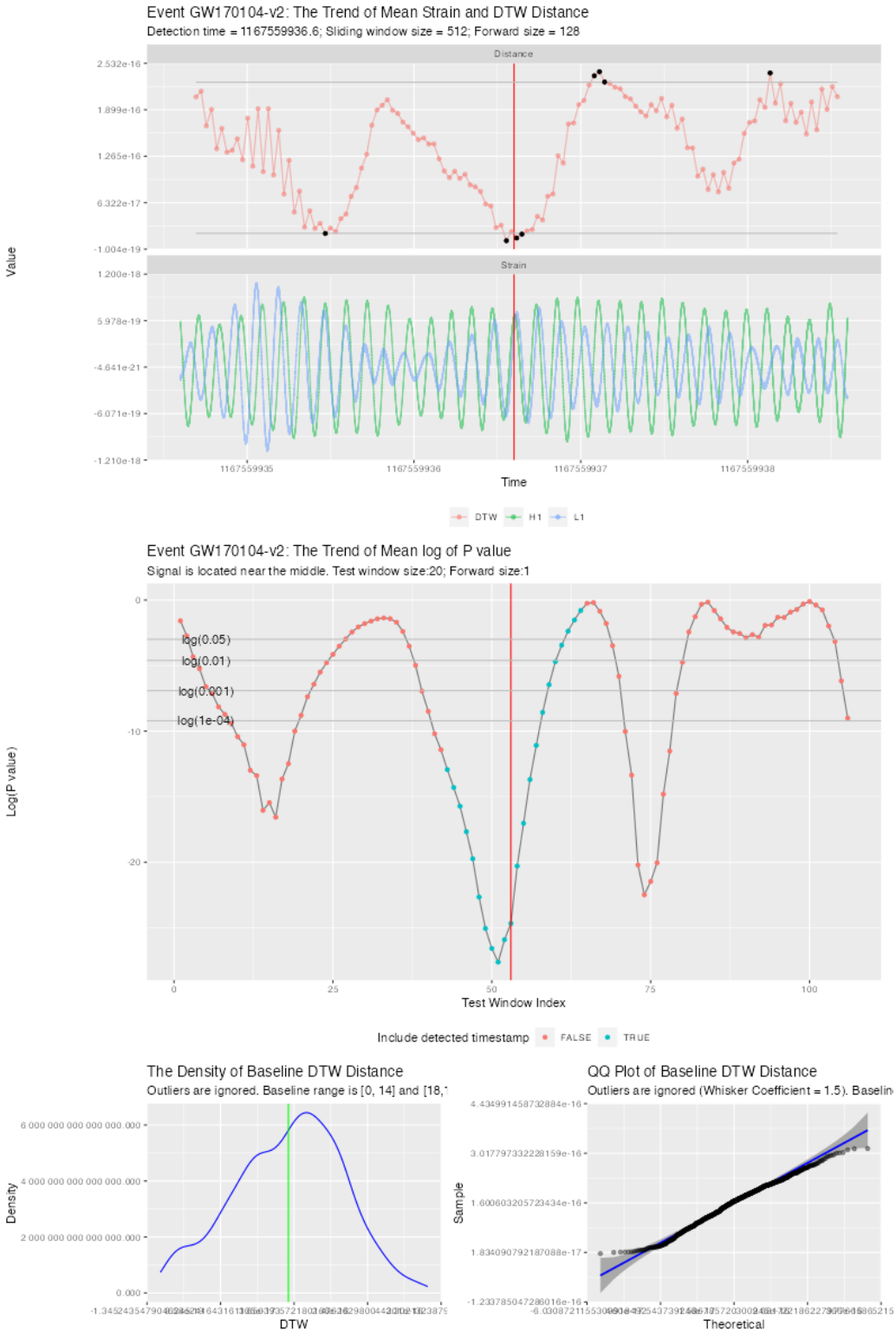


Figure 4.9: The result of the pipeline to the GW event GW170104. The top two plots show the DTW distance calculated from centralized strain data from H1 and L1 shown in the strain plot. The middle plot shows the trend of log p-values with blue points indicating that this time period includes the timestamp of the detected signal, coloured with a red vertical line. The last two plots show the Gaussian distribution of the nearby 30 seconds of data excluding the outliers.

As an example, the result of DTW distance of the GW event GW170104 in GWTC-1 is shown in Figure 4.9. The centralization of the original strain data reduces the magnitude difference between two time series, increasing the sensitivity of the result. Besides that, the baseline distribution of noise is almost Gaussian distributed. It also contains the detection time (red vertical line) inside the significant time period. The period with a steeper log of p-value before the detection also shows the drawbacks of the DTW distance, depending on the shape information only.

The detection result among each distance is listed in Table 4.3. The most powerful distance here is the DTW distance with a higher detection ratio, including 9 out of 11 events' detection time inside the significant region. TWED and Euclidean distance both have 7 out of 11 events detected inside the significant region. The Fréchet distance is not as powerful as the other three distances, with only 6 out of 11 events inside the significant region. Adding the time information reduces the significant ratio in Table 4.4 as shown in the Fréchet distance. The TWED distance has a wider significant range due to a dense grid provided compared with Fréchet distance in Chapter 3 Section 3.4.2.

To further look into the non-detected signals, we take GW event GW170818 as an example shown in Figure 4.10. Though the baseline distribution of the nearby period is almost Gaussian, the p-value of the periods which include the detected timestamp is not smaller than the significance level of 0.0001. This might be because of the magnitude between two time series as shown in the top strain plot: the magnitude of LIGO Livingston (L1) is 10 times less than that of LIGO Hansford (H1). All of four distances cannot claim a significant detection in this region, revealing the shape-only approach is not strong enough because of the shortage of prior information, e.g., a template. Another reason for the non-detected signals might because of the disturbance of the background noise. According to the reconstructed signal of GW170818 in Figure 4.11, an oscillation with large magnitude is hidden behind the whitened reconstructed strain data in the Hanford detector compared

to that in Livingston [4]. So the comparison of the coincidence still requires whitening of the data or filtering of a certain frequency in future practice.

Regard to whitening or filtering, the power spectral density is estimated from each detector in the time of no detection. Though similar to the idea of GW signal-free range, in our procedures of early detection, the assumption of Gaussian noise distribution is an idealized case of the data, of which the original data deviates from this assumption as shown in the graph of Q-Q plot of DTW distribution from Figure 4.9. In practice, LIGO and Virgo Collaboration handle noise with more care, considering all possibilities of inevitable instrumental noise, e.g., quantum sensing noise and gravity gradient noise [11], and searching the source transient disturbances which cause spectral lines, e.g., human activity and earthquake [12]. With all meticulous classification of GW signals and glitches, the LIGO-Virgo strain data in each detector is approximated as Gaussian and stationary in the whitened data [28]. A comparison between whitened data and the original sinusoidally-varying wave is also shown in Figure 2.2, indicating a stationary and Gaussian noise after whitening.

Apart from the pipeline, the experiments and simulation reveal the potential of elastic distance $D(x, y)$ as a metric in the detection of GW signal in the Deep Learning approach, e.g., convolutional neural networks (CNN) with binary cross entropy as the loss function to classify signal from noise [31]. Our experiment of DTW in LIGO data reveals the potential of the loss function including the idea of similarity and coincidence, which can be combined with other auxiliary and environmental channels to establish a pipeline which distinguishes signals from noise [46]. Considering the first experiment shown in Figure 4.3, there are two possibilities of the elastic distance, taking DTW as an example:

- An extremely small DTW distance shows the synchronized shape shown in two detectors, indicating the coincidence of the GW detection from the same source, or glitches.
- An extremely large DTW distance shows the mismatch between two detectors, in-

dicating the mismatch between signal compared to noise, in which signal might be hidden from the signal.

Either situation cannot rule out the possibility of detection. Due to that, the minimum between the distance and its reciprocal value, min{$D(x, y)$, 1/D(x,y)}, which finds a balance between the minimal and maximal value of a statistic $D(x, y)$. Implementing this kind of shape information and time shift will be beneficial to Deep Learning in representing a more sophisticated loss function to increase the classification accuracy of signals from glitches. Besides, with the new value min{$D(x, y)$, 1/D(x,y)}, it is worthy of deriving new statistics for the one-sided hypothesis test with higher accuracy and power for the further improvement and upgrade of the procedures of early warning.

Still, as mentioned the sacrifice in the false-alarm rate, we can still apply this method as a preliminary early warning before the standard searching pipeline, e.g., PyCBC [47], which reduces the calculation of the matched filter over the whole series. If DTW is applied, we only need to search over 40% of data for the detection given the low significance level. To guarantee the detection of a signal, a higher significance level should be applied, resulting in a larger false alarm rate and a wider period of the standard searching pipeline.

Event	Detection Time	DTW	Euclidean	Fréchet	TWED
GW150914	1126259462.40	[-0.18,1.07]	[-0.21,1.16]	[-0.21,1.16]	[-0.21,1.19]
GW151012	1128678900.40	[-0.78,0.29]	[-0.78,0.32]	[-0.78,0.41]	[-0.78,0.32]
GW151226	1135136350.60	[-0.38,0.49]	-	[-0.38,0.53]	-
GW170104	1167559936.60	[-0.73,0.46]	[-0.98,0.34]	[-0.82,0.28]	[-0.85,0.34]
GW170608	1180922494.50	[-0.88,0.50]	[-0.66,0.69]	[-0.72,0.44]	[-0.66,0.72]
GW170729	1185389807.30	[-0.24,0.83]	[-0.43,0.45]	-	[-0.46,0.51]
GW170809	1186302519.80	-	-	-	-
GW170814	1186741861.50	[-0.84,0.50]	[-0.72,0.63]	[-0.72,0.38]	[-0.719,0.59]
GW170817	1187008882.40	[-0.24,0.76]	[-0.18,0.73]	-	[-0.18,0.73]
GW170818	1187058327.10	-	-	-	-
GW170823	1187529256.50	[-0.88,1.13]	[-0.88,0.38]	[-0.84,0.47]	[-0.88,1.06]
Detection Ratio		81.82	72.73	63.64	72.73

Table 4.3: The possible detection time range around the true detection timestamp for all 11 GW events in GWTC-1. A null value means no detection is made according to our early warning pipeline.

Event	File Start Time	DTW	Euclidean	Fréchet	TWED
GW150914	1126257415	45.24	43.68	38.93	43.18
GW151012	1128676853	48.50	51.77	51.18	53.55
GW151226	1135134303	31.43	34.49	34.65	33.93
GW170104	1167557889	58.68	59.05	31.93	58.24
GW170608	1180920447	45.34	57.05	27.15	56.71
GW170729	1185387760	32.15	37.27	23.37	39.21
GW170809	1186300472	31.31	36.90	12.84	36.68
GW170814	1186739814	35.31	47.21	30.99	48.43
GW170817	1187006835	26.34	37.49	22.62	39.65
GW170818	1187056280	48.12	46.40	19.31	48.21
GW170823	1187527209	40.68	49.09	27.31	53.71
Average Significant Ratio (%)		40.28	45.49	29.12	46.50

Table 4.4: The shrink ratio (%) of each confident event in GWTC-1. Each file stores 4096 seconds of strain data. Each value indicates the significant duration ratio among the 4096 seconds.

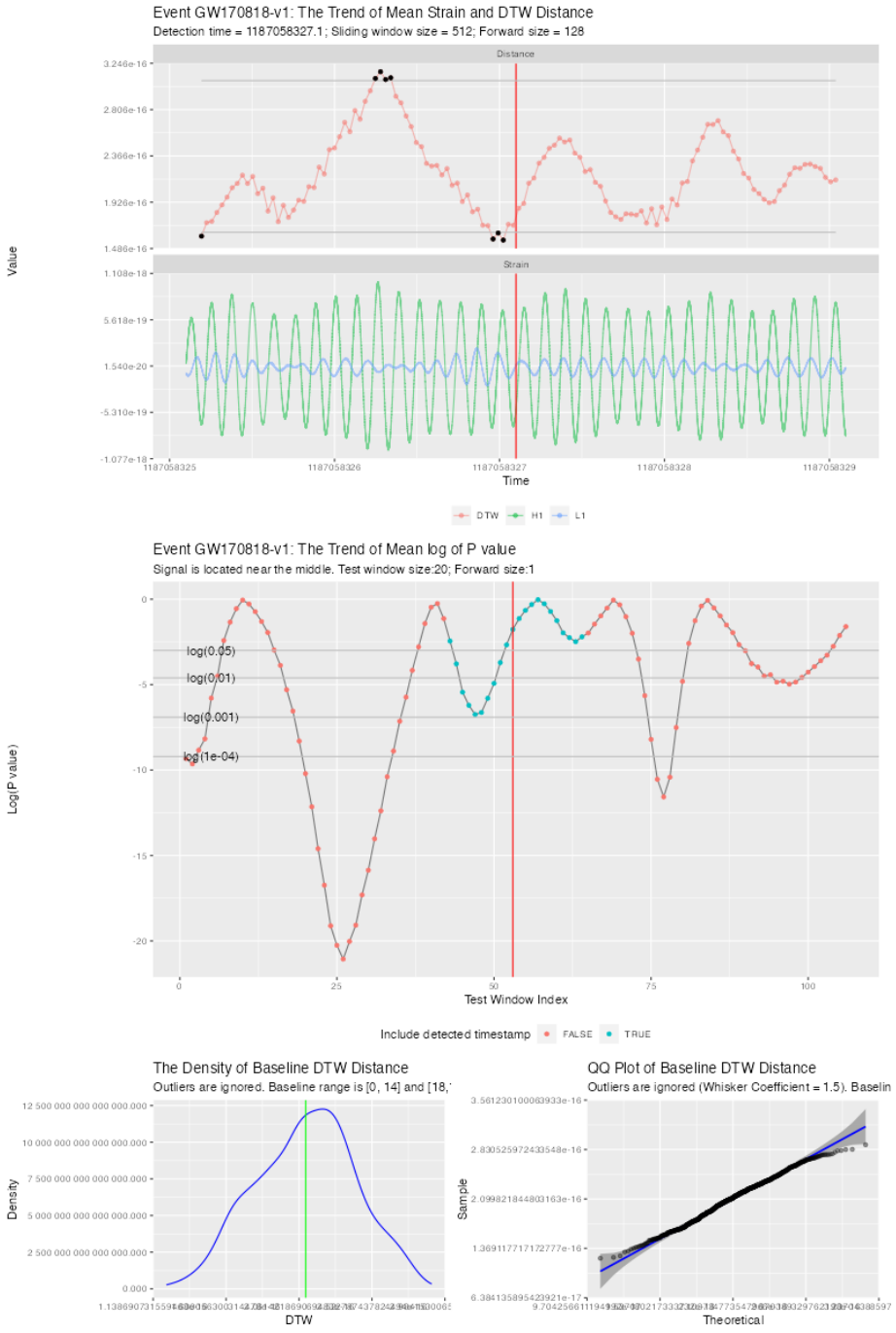


Figure 4.10: The result of the pipeline to the GW event GW170818. The top two plots show the DTW distance calculated from centralized strain data from H1 and L1 shown in the strain plot. The middle plot shows the trend of log p-values with blue points indicating that this time period includes the timestamp of the detected signal, coloured with a red vertical line. The last two plots show the Gaussian distribution of the nearby 30 seconds of data excluding the outliers.

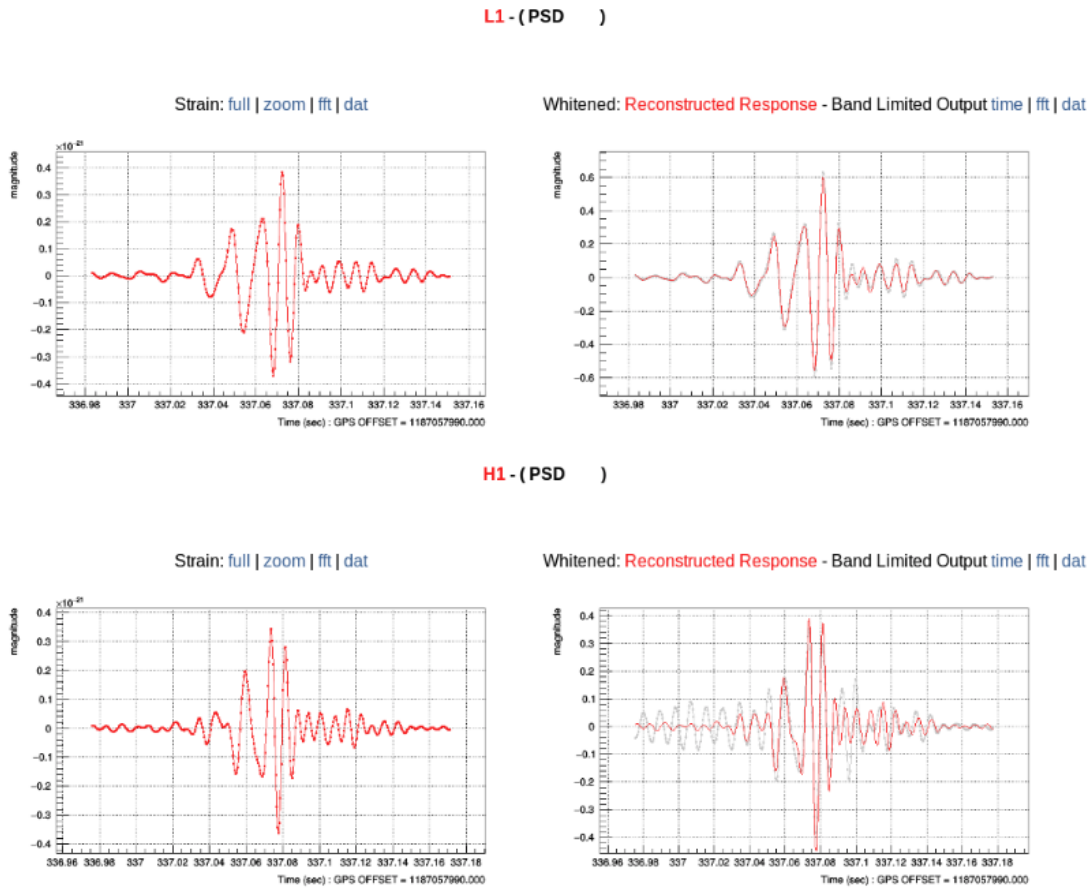


Figure 4.11: The reconstruction of the GW170818 in Livingston and Hanford detectors from cWB [4]. The whitened response in the Livingston detector is a perfect match with the signal, while there are large oscillations before the detection time in the Hanford detector.

Chapter 5

5 Conclusion

In this thesis, we examined the elastic distance and its application in astrophysical data, either the simulated dataset or the real detection, leveraging the information in shape and time information from time series. Except for DTW and TWED distance, as a similar distance measure which allows distortion in time, the Skorohod distance is also added in the comparison by its equivalent discrete distance, discrete Fréchet distance. The comparison between Euclidean distance, DTW, TWED and Fréchet distance reveals the applicability of the idea of similarity in astrophysical time series, and the drawbacks of elastic distance in practice.

Firstly, all three elastic distances are capable of classification shown in the UCR dataset in Section 4.1.1 and simulated LSST dataset by feeding the distance metric into the 1-NN classifier in Section 4.1.2 of Chapter 4. TWED distance outperforms the state-of-the-art DTW distance and other distances because of the inclusion of warping in the timestamp. However, a 1-NN classifier with elastic distance only is not enough to have an overall optimal classification accuracy, which requires better time series classification models, such as Long short-term memory (LSTM) networks or convolutional neural networks (CNN).

Recalling the matched-filtering of templates versus strain data in Gravitational-Wave (GW) detection, the elastic distance shows the potential for optimizing the template bank by shrinking its size in Section 4.2.1, and the early warning for a possible GW signal by the idea of coincidence in the LIGO, Virgo and KAGRA collaboration network of detectors in Section 4.2.2. We make a few experiments with the Python module, PyCBC [47].

Specifically in shrinking template bank, we shrink the size of a toy template bank with mass-only parameters of interest. Without the loss of the effectualness $\mathcal{E}$ of the template

bank, TWED shows the similar parameter space distribution of the chirp mass in a toy dataset of template bank which simulates the merger of binary black holes as shown in Figure 4.7. This knowledge shall be applied to a larger template bank with effectual templates, faster algorithms and a hierarchical detection framework for thorough verification. However, DTW distance shows an abnormal template in Figure 4.6 which represents both the effectualness with the largest overlap and the removal one with a small overall DTW distance, requiring a further test over a wider range of time period and more simulated signals to determine whether it is an inherent lack of discrimination or by an accidental outlier.

In the early warning of detection by the idea of coincidence, we reduced the size of the strain data by DTW for a second-stage matched-filtering into 40% to guarantee a better detection rate or accuracy in the GWTC-1 confident events strain data, shown in Table 4.4. However, in Table 4.3, the defect of low accuracy and high false-alarm rate is obvious due to the noise-dominated data which requires pre-processing of the data, e.g., whitening by strains power spectral density to exaggerate the magnitude ratio between signal and noise. Because DTW and other elastic distances are quite sensitive to the shape. Furthermore, with slight modification into $\min\{D(x, y), 1/\mathrm{D(x,y)}\}$, the DTW distance has the potential to become a novel metric in detecting GW signals by Deep Learning, which shall be discussed in the future work.

About the computational cost, one R package dtwCpp is developed for a fast implementation in parallel computing to accelerate the computation. Though the complexity of dynamic programming prevents DTW and its variants from becoming the first choice in the previous practice in LIGO data [45], we still explore the several applications of elastic distance.

To sum up, the elastic distance, which allows time distortion in matching, provides a quick and data-oriented similarity metric for astronomical time series data without preliminary

knowledge. In the four datasets of experiments, we conclude that Time Warping Edit Distance (TWED) outperforms other elastic distances with results in Table 5.1. TWED only fails to find one detection compared to DTW distance in early warning of a GW signal detection as shown in Table 4.3. Nevertheless, domain knowledge is always required for fine-tuning and integration in the current pipeline to exploit strengths and avoid weaknesses of these elastic distance 'metrics'.

Application	Evaluation Criteria	DTW-full	Euclidean	Fréchet	TWED
UCR Database	Average Accuracy Rank (Descending)	3.10	3.32	3.42	**2.54**
LSST Dataset	Loose Accuracy	0.72	0.73	-	**0.74**
	Mode Accuracy	0.42	0.42	-	**0.43**
GW Template Bank	Effectualness	The shrinkage doesn't affect effectualness, but decreases one effectualness in DTW.			
GW Early Warning	Detection Ratio	81.82	72.73	63.64	72.73
	Average Significant Ratio	40.28	45.49	29.12	46.50

Table 5.1: The overall comparison between elastic distances we choose, DTW, Fréchet distance and TWED. We take the Euclidean distance as a reference. TWED outperforms other distance. Fréchet distance is used to calculate the Skorohod distance in discrete time series. LSST dataset doesn't include Fréchet distance due to its poor performance in UCR database and high computational complexity.

References

[1] Archive of dataset o1 16khz of detector h1. https://gwosc.org/archive/links/O1_16KHZ/H1/1126051217/1137254417/stats/. Accessed: 2024-04-10.

[2] Briefing document of ucr time series classification archive. https://www.cs.ucr.edu/%7Eeamonn/time_series_data_2018/BriefingDocument2018.pdf. Accessed: 2024-06-10.

[3] Gw150914: Factsheet. https://www.ligo.org/detections/GW150914/fact-sheet.pdf. Accessed: 2024-04-11.

[4] Gw170818: Gwtc-1 catalog (gwosc) - coherent event display (hl). https://gwburst.gitlab.io/info/gw170818_ced_gwosc/. Accessed: 2024-04-18.

[5] Ligo document t2000012-v2, noise curves used for simulations in the update of the observing scenarios paper. https://dcc.ligo.org/LIGO-T2000012/public. Accessed: 2024-04-20.

[6] Ligo, virgo and kagra observing run plans. https://observing.docs.ligo.org/plan/#. Accessed: 2024-04-20.

[7] Observation of gravitational waves from a binary black hole merger. https://www.ligo.org/science/Publication-GW150914/. Accessed: 2020-08-02.

[8] Photometric lsst astronomical time-series classification challenge (plasticc). https://plasticc.org. Accessed: 2020-08-02.

[9] Pycbc template bank generation documentation. https://pycbc.org/pycbc/latest/html/tmpltbank.html#the-module-s-source-code. Accessed: 2024-04-20.

[10] pycbc.psd package. `https://pycbc.org/pycbc/latest/html/pycbc.psd.html#pycbc.psd.analytical.aLIGOAdVO4T1800545`. Accessed: 2024-04-20.

[11] B. P. et.al. Abbott. Advanced ligo. *Classical and Quantum Gravity*, 32(7):074001, mar 2015.

[12] B. P. et.al. Abbott. Characterization of transient noise in advanced ligo relevant to gravitational wave signal gw150914. *Classical and Quantum Gravity*, 33(13):134001, jun 2016.

[13] B. P. et.al. Abbott. Observation of gravitational waves from a binary black hole merger. *Phys. Rev. Lett.*, 116:061102, Feb 2016.

[14] Patrick Billingsley. *Convergence of probability measures*. Wiley series in probability and statistics. Probability and statistics section. Wiley, 2nd ed edition, 1999.

[15] the Virgo Collaboration B.P. Abbott et al. (The LIGO Scientific Collaboration and the KAGRA Collaboration). Open data from the third observing run of ligo, virgo, kagra, and geo. *The Astrophysical Journal Supplement Series*, 267(2):29, jul 2023.

[16] P.J. Brockwell, R.A. Davis, S.E. Fienberg, J.O. Berger, J. Gani, K. Krickeberg, I. Olkin, and B. Singer. *Time Series: Theory and Methods: Theory and Methods*. Springer Series in Statistics. Springer New York, 1991.

[17] Duncan A. Brown, Ian Harry, Andrew Lundgren, and Alexander H. Nitz. Detecting binary neutron star systems with spin in advanced gravitational-wave detectors. *Phys. Rev. D*, 86:084017, Oct 2012.

[18] Collin Capano, Ian Harry, Stephen Privitera, and Alessandra Buonanno. Implementing a search for gravitational waves from binary black holes with nonprecessing spin. *Phys. Rev. D*, 93:124007, Jun 2016.

[19] Lei Chen. On the marriage of lp-norms and edit distance. page 12.

[20] LSST Science Collaboration and et.al. Paul A. Abell. Lsst science book, version 2.0, 2009.

[21] Neil J Cornish and Tyson B Littenberg. Bayeswave: Bayesian inference for gravitational wave bursts and instrument glitches. *Classical and Quantum Gravity*, 32(13):135012, jun 2015.

[22] Marco Cuturi and Mathieu Blondel. Soft-DTW: a differentiable loss function for time-series.

[23] Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The UCR time series archive.

[24] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[25] Dave Deriso and Stephen Boyd. A general optimization framework for dynamic time warping. page 23.

[26] Thomas Eiter and Heikki Mannila. Computing discrete fréchet distance. Technical report, Citeseer, 1994.

[27] B.P. Abbott et al. (LIGO–Virgo–KAGRA Collaboration). Gwtc-1: A gravitational-wave transient catalog of compact binary mergers observed by ligo and virgo during the first and second observing runs. *Phys. Rev. X*, 9:031040, Sep 2019.

[28] B.P. Abbott et al. (LIGO–Virgo–KAGRA Collaboration). A guide to ligo–virgo detector noise and extraction of transient gravitational-wave signals. *Classical and Quantum Gravity*, 37(5):055002, feb 2020.

[29] B.P. Abbott et al. (LIGO–Virgo–KAGRA Collaboration). Prospects for observing

and localizing gravitational-wave transients with advanced ligo, advanced virgo and kagra. *Living Reviews in Relativity*, 23(1):3, Sep 2020.

[30] Ada Wai-Chee Fu, Eamonn Keogh, Leo Yung Hang Lau, Chotirat Ann Ratanamahatana, and Raymond Chi-Wing Wong. Scaling and time warping in time series querying. 17(4):899–921.

[31] Hunter Gabbard, Michael Williams, Fergus Hayes, and Chris Messenger. Matching matched filtering with deep networks for gravitational-wave astronomy. *Phys. Rev. Lett.*, 120:141103, Apr 2018.

[32] Shaghayegh Gharghabi, Shima Imani, Anthony Bagnall, Amirali Darvishzadeh, and Eamonn Keogh. Matrix profile xii: Mpdist: A novel time series distance measure to allow data mining in more challenging scenarios. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 965–970, 2018.

[33] Toni Giorgino. Computing and visualizing dynamic time warping alignments in *R* : The **dtw** package. 31(7).

[34] Young-Seon Jeong, Myong K. Jeong, and Olufemi A. Omitaomu. Weighted dynamic time warping for time series classification. 44(9):2231–2240.

[35] Yihang Jiang, Yuankai Qi, Will Ke Wang, Brinnae Bent, Robert Avram, Jeffrey Olgin, and Jessilyn Dunn. EventDTW: An improved dynamic time warping algorithm for aligning biomedical signals of nonuniform sampling frequencies. 20(9):2700.

[36] Eamonn J. Keogh and Michael J. Pazzani. Derivative dynamic time warping. In *Proceedings of the 2001 SIAM International Conference on Data Mining*, pages 1–11. Society for Industrial and Applied Mathematics.

[37] R. Kessler, G. Narayan, A. Avelino, E. Bachelet, R. Biswas, P. J. Brown, D. F. Chernoff, A. J. Connolly, M. Dai, S. Daniel, R. Di Stefano, M. R. Drout, L. Galbany,

S. González-Gaitán, M. L. Graham, R. Hložek, E. E. O. Ishida, J. Guillochon, S. W. Jha, D. O. Jones, K. S. Mandel, D. Muthukrishna, A. O'Grady, C. M. Peters, J. R. Pierel, K. A. Ponder, A. Prša, S. Rodney, and V. A. Villar. Models and simulations for the photometric LSST astronomical time series classification challenge (PLAsTiCC). 131(1003):094501.

[38] R Kessler, G Narayan, A Avelino, E Bachelet, Rahul Biswas, PJ Brown, DF Chernoff, AJ Connolly, M Dai, S Daniel, et al. Models and simulations for the photometric lsst astronomical time series classification challenge (plasticc). *Publications of the Astronomical Society of the Pacific*, 131(1003):094501, 2019.

[39] Jessica Lin, Sheri Williamson, Kirk D. Borne, and David. 1 chapter 1 pattern recognition in time series. 2011.

[40] M. Maggiore. *Gravitational Waves: Volume 1: Theory and Experiments*. OUP Oxford, 2007.

[41] P. Marteau. Time warp edit distance with stiffness adjustment for time series matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):306–318, Feb 2009.

[42] Charles W. Misner, Kip S. Thorne, and John Archibald Wheeler. *Gravitation*. 1973.

[43] S. D. Mohanty and S. V. Dhurandhar. Hierarchical search strategy for the detection of gravitational waves from coalescing binaries. *Phys. Rev. D*, 54:7108–7128, Dec 1996.

[44] Abdullah Mueen, Nikan Chavoshi, Noor Abu-El-Rub, Hossein Hamooni, and Amanda Minnich. AWarp: Fast warping distance for sparse time series. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 350–359. IEEE.

[45] S Mukherjee, R Obaid, and B Matkarimov. Classification of glitch waveforms in gravitational wave detector characterization. *Journal of Physics: Conference Series*, 243(1):012006, aug 2010.

[46] Soma Mukherjee and (on behalf ofthe LIGO Scientific Collaboration). Preliminary results from the hierarchical glitch pipeline. *Classical and Quantum Gravity*, 24(19):S701, sep 2007.

[47] Alex Nitz, Ian Harry, Duncan Brown, Christopher M. Biwer, Josh Willis, Tito Dal Canton, Collin Capano, Thomas Dent, Larne Pekowsky, Gareth S Cabourn Davies, Soumi De, Miriam Cabero, Shichao Wu, Andrew R. Williamson, Bernd Machenschalk, Duncan Macleod, Francesco Pannarale, Prayush Kumar, Steven Reyes, dfinstad, Sumit Kumar, Márton Tápai, Leo Singer, Praveen Kumar, veronica villa, maxtrevor, Bhooshan Uday Varsha Gadre, Sebastian Khan, Stephen Fairhurst, and Arthur Tolley. gwastro/pycbc: v2.3.3 release of pycbc, January 2024.

[48] E. Parzen. *Time Series Analysis of Irregularly Observed Data: Proceedings of a Symposium Held at Texas A & M University, College Station, Texas February 10–13, 1983*. Lecture Notes in Statistics. Springer New York, 1984.

[49] Chotirat Ann Ratanamahatana and Eamonn Keogh. Three myths about dynamic time warping data mining. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 506–510. Society for Industrial and Applied Mathematics.

[50] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. 26(1):43–49.

[51] Stan Salvador and Philip Chan. FastDTW: Toward accurate dynamic time warping in linear time and space. page 11.

[52] Samsu Sempena, Nur Ulfa Maulidevi, and Peb Ruswono Aryan. Human action recog-

nition using dynamic time warping. In *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, pages 1–5. IEEE, 2011.

[53] A. V. Skorokhod. Limit theorems for stochastic processes. *Theory of Probability & Its Applications*, 1(3):261–290, 1956.

[54] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[55] Kanchan Soni, Bhooshan Uday Gadre, Sanjit Mitra, and Sanjeev Dhurandhar. Hierarchical search for compact binary coalescences in the advanced ligo's first two observing runs. *Phys. Rev. D*, 105:064005, Mar 2022.

[56] Sofia Suvorova, Jade Powell, and Andrew Melatos. Reconstructing gravitational wave core-collapse supernova signals with dynamic time warping. *Physical Review D*, 99(12):123012, 2019.

[57] The PLAsTiCC team, Tarek Allam Jr., Anita Bahmanyar, Rahul Biswas, Mi Dai, Lluís Galbany, Renée Hložek, Emille E. O. Ishida, Saurabh W. Jha, David O. Jones, Richard Kessler, Michelle Lochner, Ashish A. Mahabal, Alex I. Malz, Kaisey S. Mandel, Juan Rafael Martínez-Galarza, Jason D. McEwen, Daniel Muthukrishna, Gautham Narayan, Hiranya Peiris, Christina M. Peters, Kara Ponder, Christian N. Setzer, The LSST Dark Energy Science Collaboration, The LSST Transients, and Variable Stars Science Collaboration. The photometric LSST astronomical time-series classification challenge (PLAsTiCC): Data set.

[58] Samantha A Usman, Alexander H Nitz, Ian W Harry, Christopher M Biwer, Duncan A Brown, Miriam Cabero, Collin D Capano, Tito Dal Canton, Thomas Dent,

Stephen Fairhurst, Marcel S Kehl, Drew Keppel, Badri Krishnan, Amber Lenon, Andrew Lundgren, Alex B Nielsen, Larne P Pekowsky, Harald P Pfeiffer, Peter R Saulson, Matthew West, and Joshua L Willis. The pycbc search for gravitational waves from compact binary coalescence. *Classical and Quantum Gravity*, 33(21):215004, oct 2016.

[59] M.S Waterman, T.F Smith, and W.A Beyer. Some biological sequence metrics. 20(3):367–387.

[60] Byoung-Kee Yi, Hosagrahar V Jagadish, and Christos Faloutsos. Efficient retrieval of similar time sequences under time warping. In *Proceedings 14th International Conference on Data Engineering*, pages 201–208. IEEE, 1998.

[61] Jiaping Zhao and Laurent Itti. shapeDTW: shape dynamic time warping.

Appendices

A Proof of Theorems

The Skorohod topology [14] is a powerful tool to investigate stochastic process limits. The Skorohod topology is defined in the Space D, which includes all real functions x on [0,1] that are right continuous and have left-hand-limits. These functions are called *càdlàg* functions [14].

1. For $0 \leq t < 1$, $x(t+) = \lim_{s \downarrow t} x(s)$ exists and $x(t+) = x(t)$.
2. For $0 \leq t < 1$, $x(t-) = \lim_{s \uparrow t} x(s)$ exists.

Theorem A.1 *The Skorohod distance is a metric. [14].*

Proof There are four conditions to be satisfied for a metric.

1. Non-negativity: $d(x, y) \geq 0$ due to that the infimum of positive ϵ is non-negative.
2. Symmetry: $d(x, y) = d(y, x)$ is derived from the definition of λ. As λ is strictly increasing, continuous mapping onto itself in [0,1], the inverse of λ, λ^{-1} exists. From the definition, there are intrinsic symmetry in $d(x, y)$.
3. Identity of indiscernibles: $d(x, y) = 0 \Leftrightarrow x = y$.

   First, assume $x = y$, then $|x(t) - y(\lambda t)| = 0$ when $\lambda t = t$, which is $\lambda = I$. Then the infimum of the max between $\|\lambda - I\|$ and $\|x - y\lambda\|$ is always 0 when $\lambda = I$.

   Secondly, assume $d(x, y) = \inf_{\lambda}\{\sup_t |\lambda t - t| \vee \sup_t |x(t) - y(\lambda t)|\} = 0$. Then $\forall \epsilon_0$, $\exists \lambda_0$, such that $\sup_t |\lambda_0 t - t| < \epsilon_0$ and $\sup_t |x(t) - y(\lambda_0 t)| < \epsilon_0$. From the definition of supremum , $|\lambda_0 t - t| \leq \epsilon_0$ and $|x(t) - y(\lambda_0 t)| \leq \epsilon_0$ for any t. When $\lambda = I$, $x(t) = y(t) (= y(t+))$. When $\lambda \neq I$, say $\lambda(t_0) = t_0 - \epsilon_0$ at $t = t_0$. Then $|x(t_0) - y(t_0 - \epsilon_0)| \leq \epsilon_0$ indicates that $x(t_0) = y(t_0-)$. Then for any such t_0 which is the jump points at y, we all have $y(t_0-) = x(t_0) = y(t_0)$. This also implies that $x(t) = y(t)$ in Space D.

4. Triangular inequality: $d(x,z) \leq d(x,y) + d(z,y)$.

   Two lemmas are necessary here. The first is $\|\lambda_1\lambda_2 - I\| \leq \|\lambda_1 - I\| + \|\lambda_2 - I\|$.

$$\begin{aligned}
\|\lambda_1\lambda_2 - I\| &= \sup_t |\lambda_1\lambda_2 t - t| \\
&= \sup_t |\lambda_1\lambda_2 t - \lambda_2 t + \lambda_2 t - t| \\
&\leq \sup_t |\lambda_1\lambda_2 t - \lambda_2 t| + \sup_t |\lambda_2 t - t| \\
&= \sup_s |\lambda_1 s - s| + \sup_t |\lambda_2 t - t| \\
&= \|\lambda_1 - I\| + \|\lambda_2 - I\|
\end{aligned}$$

   The second lemma is $\|x - z\lambda_1\lambda 2\| \leq \|x - y\lambda_2\| + \|y - z\lambda_1\|$.

$$\begin{aligned}
\|x - z\lambda_1\lambda_2\| &= \sup_t |x(t) - z(\lambda_1\lambda_2 t)| \\
&= \sup_t |x(t) - y(\lambda_2 t) + y(\lambda_2 t) - z(\lambda_1\lambda_2 t)| \\
&\leq \sup_t |x(t) - y(\lambda_2 t)| + \sup_t |y(\lambda_2 t) - z(\lambda_1\lambda_2 t)| \\
&= \sup_t |x(t) - y(\lambda_2 t)| + \sup_s |y(s) - z(\lambda_1 s)| \\
&= \|x - y\lambda_2\| + \|y - z\lambda_1\|
\end{aligned}$$

   Combine the above two equation together for given λ_1 and $\lambda 2$, we have

$$\max\{\|\lambda_1\lambda_2 - I\|, \|x - z\lambda_1\lambda_2\|\} \leq \inf_{\lambda_1}\{\max\{\|\lambda_1 - I\|, \|x - y\lambda_1\|\}\} + \inf_{\lambda_2}\{\max\{\|\lambda_2 - I\|, \|x - z\lambda_2\|\}\}$$

   Then, from the definition of the $d(x,z)$, we have $d(x,z) \leq \max\{\|\lambda_1\lambda_2 - I\|, \|x - z\lambda_1\lambda_2\|\}$, which derives the triangular inequality,

$$d(x,z) \leq d(x,y) + d(y,z)$$

A novel modification on Skorohod distance is replacing maximum by addition of time perturbation and the difference between two stochastic processes, $x(t)$ and $y(t)$,

$$d^*(x,y) = \inf_{\lambda}\{||\lambda - I|| + ||x - y\lambda||\}$$

Theorem A.2 *$d^*(x,y)$ is a metric defined in Space D.*

Proof There are four conditions to be satisfied for a metric.

1. Non-negativity: $d^*(x,y) \geq 0$ due to that the infimum of positive ϵ is non-negative.

2. Symmetry: $d^*(x,y) = d^*(y,x)$ is derived from the symmetric of $d(x,y)$.

3. Identity of indiscernibles: $d^*(x,y) = 0 \Leftrightarrow x = y$.

   First, assume $x = y$, then $|x(t) - y(\lambda t)| = 0$ when $\lambda t = t$, which is $\lambda = I$. Then the summation, $||\lambda - I|| + ||x - y\lambda||$ is always 0 when $\lambda = I$.

   Secondly, assume $d^*(x,y) = \inf_\lambda\{||\lambda - I|| + ||x - y\lambda||\} = 0$. Then $\forall\epsilon_0$, $\exists\lambda_0$, such that $\sup_t |\lambda_0 t - t| + \sup_t |x(t) - y(\lambda_0 t)| < \epsilon_0$, which indicates that $|\lambda_0 t - t| + |x(t) - y(\lambda_0 t)| \leq \epsilon_0$ for any t. From the non-negativity of $||\cdot||$, we have the same small enough values for $|\lambda_0 t - t| \leq \epsilon_0/2$ and $|x(t) - y(\lambda_0 t)| \leq \epsilon_0/2$. When $\lambda = I$, $x(t) = y(t)(= y(t+))$. When $\lambda \neq I$, say $\lambda(t_0) = t_0 - \epsilon_0/2$ at $t = t_0$. Then $|x(t_0) - y(t_0 - \epsilon_0)| \leq \epsilon_0$ indicates that $x(t_0) = y(t_0-)$. Then for any such t_0 which is the jump points at y, we all have $y(t_0-) = x(t_0) = y(t_0)$. This also implies that $x(t) = y(t)$ in Space D.

4. Triangular inequality: $d^*(x,y) \leq d^*(x,z) + d^*(z,y)$.

   The same two lemmas are required, which is proved in the previous proof. Add the both sides of the inequality, we have

$$||\lambda_1\lambda_2 - I|| + ||x - z\lambda_1\lambda_2|| \leq \inf_{\lambda_1}\{||\lambda_1 - I|| + ||x - y\lambda_1||\} + \inf_{\lambda_2}\{||\lambda_2 - I|| + ||x - z\lambda_2||\}$$

Then, we derive the triangular inequality,

$$d(x,z) \leq d(x,y) + d(y,z)$$

Theorem A.3 *Let $f : [a_f, b_f] \to \mathbb{R}^p$, the distance $||(f,t)|| = ||f(t)||_p + |t|$ defined in $\mathbb{R}^p \times T$ is a norm.*

Proof There are three conditions to be satisfied for a norm.

1. Zero vector: If $||(f,t)|| = 0$, either $||f(t)||_p + |t| = 0$ or $|t| \leq ||f(t)||_p = 0$, which means $f(t) = \vec{0}$ and $t = 0$.

2. Absolutely scalable: $||(|a|\cdot f, |a|\cdot t)|| = |a|\cdot||f(t)||_p + |a|\cdot|t| = |a|\cdot(||f(t)||_p + |t|) = |a|\cdot||(f,t)||$.

3. Triangular inequality:

$$\begin{aligned} ||(f,t)|| + ||(g,s)|| &= (||f(t)||_p + |t|) + (||g(s)||_p + |s|) \\ &= (||f(t)||_p + ||g(s)||_p) + (|t| + |s|) \\ &\geq ||(f(t),t) + (g(s),s)|| \\ &\geq ||(f,t) + (g,s)||. \end{aligned}$$

Curriculum Vitae

Name:	Xiyang Zhang
Post-Secondary Education and Degrees:	Sichuan University Chengdu, Sichuan, China 2014 - 2018 B.Sc. University of Western Ontario London, ON 2018 - 2024 M.Sc.
Honours and Awards:	Mitacs Globalink Graduate Fellowship 2018-2019
Related Work Experience:	Teaching Assistant The University of Western Ontario 2019 - 2020 Data Scientist Intern Tencent Video, Tencent Holdings Ltd. 2020 - 2023

Publications:

F. A. Chishtie, X. Zhang, and S.R. Valluri. An analytic approach for the study of pulsar spindown. Classical and Quantum Gravity, 35(14):145012, June 2018.

S. R. Valluri, V. Dergachev, X. Zhang, and F. A. Chishtie. Fourier transform of the continuous gravitational wave signal. Phys. Rev. D, 104:024065, Jul 2021.

V. Upadhyaya, X. Li, X. Zhang, S. R. Valluri. The Role of r-Modes in Pulsar Spindown, Pulsar Timing and Gravitational Waves. arXiv:2307.11270, Nov 2023.