Transport Transforms for Solving Signal Estimation and Classification Problems

---

A

Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Abu Hasnat Mohammad Rubaiyat

May 2023

APPROVAL SHEET

This

Dissertation

is submitted in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

Author: Abu Hasnat Mohammad Rubaiyat

This Dissertation has been read and approved by the examing committee:

Advisor: Gustavo Rohde

Advisor:

Committee Member: Scott Acton

Committee Member: Aidong Zhang

Committee Member: Haibo Dong

Committee Member: Jonathan Nichols

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science

May 2023

Transport Transforms for Solving Signal Estimation and Classification Problems

Abstract

The ability to solve signal estimation and classification problems has wide-ranging practical applications. With the advent of advanced sensor technology, signal data generation has increased in various fields, creating a need for effective and efficient solutions to the ever-growing signal processing problems. Prior literature offers several approaches to solving estimation and classification problems. However, they often fail to address the underlying nonlinearity of the system of interest, resulting in poor performance when solving nonlinear problems. The proposed research explores the use of transport-based transforms to estimate signal parameters and classify signal data. A transport-based generative model is adopted to define an estimation or classification problem. The mathematical properties of transport transforms are then used to simplify the problem in the transform domain, leading to an effective and efficient solution. The specific objectives of the thesis include: (1) proposing a closed-form solution to certain nonlinear estimation problems using the cumulative distribution transform (CDT), a transport-based signal transformation technique, (2) proposing an end-to-end signal classifier using the signed cumulative distribution transform (SCDT), an extension of the CDT, and (3) exploring a transport-based modeling approach to identify the governing partial differential equation (PDE) of a dynamical system.

To my parents and my beloved wife.

Acknowledgements

First and foremost, I would like to express my heartfelt gratitude to my thesis advisor, Gustavo Rohde, for his invaluable guidance, encouragement, and support throughout my PhD journey. His expertise, insightful comments, and constructive criticism have played a significant role in shaping my research and enhancing the quality of this thesis.

Next, I would also like to thank the members of my thesis committee, Scott Acton, Aidong Zhang, Haibo Dong, and Jonathan Nichols, for their valuable feedback, insightful suggestions, and constructive criticism that helped me refine my research and improve the quality of my thesis.

I am deeply grateful to my present and former lab mates: M Shifat Rabbi, Duy Hoang Tai, Naqib Pathan, Kristofor Pas, Natasha Ironside, Shiying Li, Xuwang Yin, Yan Zhuang, Mohammad Abdul Hassan, Cailey Fitzgerald, who have created a supportive and stimulating environment that has inspired me to pursue my research goals with passion and dedication.

I would like to thank my colleagues and friends at the University of Virginia, who have helped me through challenging times during my PhD.

I would like to thank the administrative staff at ECE for their help and support.

Finally, I would like to thank my family and friends. I am deeply grateful to my wife, Tanjin Taher Toma, for her unconditional love and unyielding support. My heartfelt thanks go to my parents for their unwavering love, encouragement, and support throughout my life. Their sacrifices and hard work have been the driving force behind my success.

My thesis work was supported by the National Institutes of Health (NIH) under grants GM130825 and GM090033, and by the Office of Naval Research (ONR) under grant N000142212505. I am grateful to NIH and ONR for providing financial support that made this research possible.

Thank you all for your support, encouragement, and inspiration.

Table of Contents

List of Tables

List of Figures

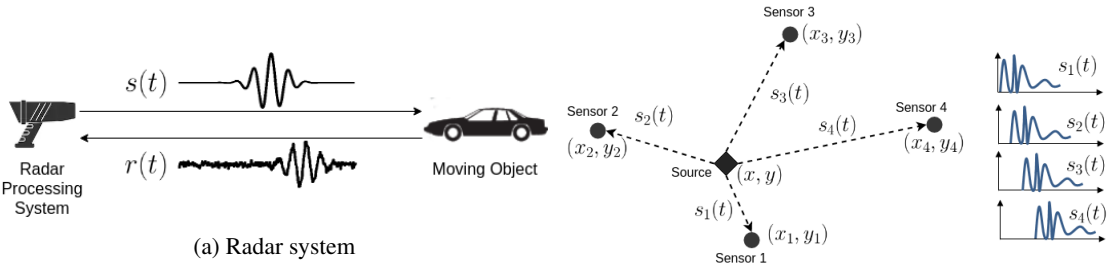
Chapter 1

Introduction

1.1 Signal Estimation & Classification

Estimation and classification problems lie at the core of many signal processing applications. The former is concerned with locating and tracking source signals in various fields, including radar [2], underwater acoustics [3], seismology [4] [5], communication [6], and source localization [7] [8]. Such systems require the estimation of multiple parameters. For instance, in radar systems, the estimated time delay and Doppler stretch between transmitted and received signals help determine the location and speed of a target object, as illustrated in Figure 1.1a. Similarly, in Figure 1.1b, the time difference of arrivals (TDOAs) of signals received by four sensors is estimated to identify the location of an AE (acoustic emission) source. Likewise, signal classification tasks are also widespread in various applications, such as human activity recognition (HAR) [9], physiological signal assessment [10][11], communications [12], structural or machine health monitoring systems [13] [14], and financial modeling [15], among others. Two examples of signal classification problems are shown in Fig. 1.2. The first example involves distinguishing between normal and diseased heartbeats from ECG signals, as depicted in Fig. 1.2a. The second example pertains to hand gesture recognition utilizing accelerometer data, as shown in Fig. 1.2b.

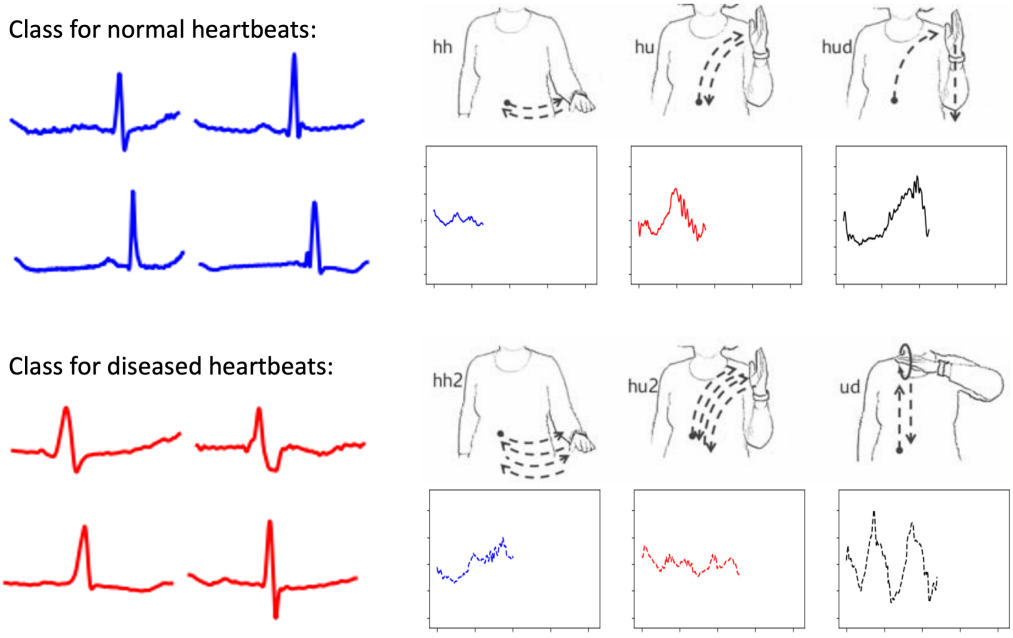
There is a substantial body of literature in signal estimation and classification, but many existing methods are inadequate for modeling nonlinear systems, leading to suboptimal outcomes. Common estimation approaches involve maximizing the likelihood function, which often results in non-convex optimization problems that are difficult and computationally expensive to solve. Maximizing cross-correlation [16], l_p correlation [17], maximizing the magnitude of difference between measured and template signals [18], entropy [19], and mutual information-based methods [20][21] are among the previously proposed techniques. However,



(a) Radar system

(b) Source localization

Figure 1.1: Examples of estimation problems in signal processing applications.



(a) Physiological signal assessment

(b) Human activity recognition

Figure 1.2: Examples of classification problems in signal processing applications.

all of these approaches assume that the only difference between the measured and template signals is time delay and noise, disregarding the possibility of complex parameterized changes in real-world scenarios. For instance, in radar-related estimation problems, the motion of the target object causes linear dispersion, also known as Doppler stretch, in addition to time delay.

The joint estimation of time delay and Doppler stretch has been a topic of study in the literature, with estimation techniques focusing on maximizing the ambiguity function between the measured and template signals [22] [23] [24]. However, these methods typically result in non-convex optimization problems, making it difficult to obtain global minima. A number of subspace-based methods [25] [26] [27] [28] have been developed for the joint estimation of time delay and Doppler parameters. The majority of these methods rely on a narrowband signal approximation, allowing for the modeling of the Doppler effect as a frequency

shift, and subsequently enabling explicit estimation. Nonetheless, like ambiguity function maximization-based techniques, subspace methods necessitate a computationally expensive *search* over a parameter space to estimate the parameters.

Signal classification approaches can be broadly categorized into two categories: feature engineering-based classifiers and end-to-end learning classifiers like convolutional neural networks (CNNs). Feature engineering-based methods [29] [30] [31] typically involve the extraction of numerical features (such as time domain features, frequency domain features, and wavelet features) from the raw signal data, followed by the use of various multivariate regression-based classification techniques such as support vector machines, linear discriminant analysis, random forests, and others. However, the majority of these methods require extensive data preprocessing and crafting specially designed features.

Distance-based techniques [32] [33] are the traditional end-to-end signal classification methods that operate directly on raw time series using similarity measures like Euclidean distance or dynamic time warping (DTW) [34]. One very effective approach in this category is the combination of 1-nearest neighbor (1NN) with DTW distance [35]. However, the computational complexity of this approach is high [36], which limits its applicability to real-world scenarios. Recent studies have explored end-to-end signal classification methods based on deep neural networks, especially convolutional neural networks (CNN) [37]. For example, Wang et al. [38] introduced three standard deep learning benchmark models for time series classification: deep multilayer perceptrons (MLP), fully convolutional networks (FCN), and residual networks (ResNet). Another popular deep learning approach is the multi-scale convolutional neural network (MCNN) [39], which leverages CNNs for end-to-end classification of univariate time series. Karim et al. [40] proposed LSTM-FCN, a combination of FCN with a Long Short Term Recurrent Neural Network (LSTM RNN) sub-module. However, these approaches typically require a significant amount of training data, can be computationally expensive, and are susceptible to out-of-distribution examples.

The main focus of this thesis is to propose novel and efficient methods to address the problems of signal estimation and classification. A generative model-based problem formulation is adopted for the systems where the signals undergo some nonlinear deformations. Additionally, a transport-based signal transformation technique is introduced to linearize certain nonlinear deformations in the transform domain. A least-squares solution is then proposed to solve nonlinear estimation and classification problems in the transformed space. This mathematical model-based approach is further utilized to solve nonlinear system identification problems in structural health monitoring applications.

1.2 Structural Health Monitoring

Structural health monitoring (SHM) systems utilize distributed sensors installed at certain regions of a structure and employ diagnostic algorithms to analyze the acquired data. SHM typically consists of four functional levels referred to as technology classification levels (TCLs) [41] [42]:

- Level I: detection of the occurrence of an event (damage/crack)
- Level II: estimation of the location of the event
- Level III: determination of the magnitude (severity) of the event
- Level IV: estimation of the remaining service life/strength of the structure

In each level, the technology employed requires analyzing time series data collected by the onboard sensors in order to effectively accomplish the targeted actions. Figure 1.3 presents a schematic overview of a generic active-sensing SHM system. A body of work aimed at the detection and identification of structural damage

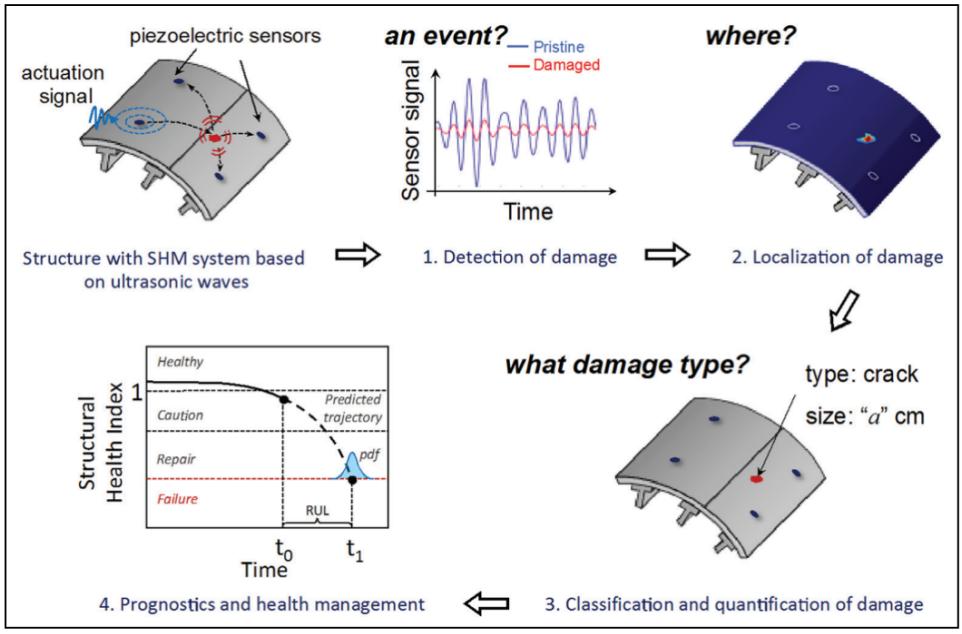


Figure 1.3: Generic active sensing structural health monitoring (SHM) principles of operation [1].

can be found in the field of Structural Health Monitoring (SHM). Typically, this information is inferred from the dynamic response of the structure to ambient or applied excitation [43]. This thesis aims to use the dynamic response measured at a specific location of the structure over time to detect and identify nonlinearity induced by damage.

Although the idea of inferring parametric system equations from time series data has been around for a while [44], traditional linear identification methods such as eigensystem realization algorithm (ERA) [45] and

dynamic mode decomposition (DMD) [46] [47] fail to account for the nonlinear behavior of a dynamical system [48]. To address this limitation, various data-driven nonlinear system identification methods have been proposed in the literature. For instance, NARMAX [49], neural networks [50] [51] [52] [53], equation-free methods [54] [55], and Laplacian spectral analysis [56] are among the methods that have been used. Sparse regression-based methods have also been employed to identify ordinary [57] [58] and partial [59][46][60] differential equations. However, these methods often require significant amounts of data for training, involve tuning a large number of parameters, and can be computationally expensive. This thesis proposes a transport transform-based data-driven approach to solve the parametric system identification problem in a computationally and data-efficient manner.

1.3 Dissertation Overview

The main focus of this thesis is to introduce innovative and efficient solutions for signal estimation and classification problems in systems where signals can be represented as observations of templates subjected to unknown deformations.

First, we present a new method for estimating signal model parameters using the Cumulative Distribution Transform (CDT). The proposed approach minimizes the Wasserstein distance between measured and model signals. We establish some useful properties of the CDT and demonstrate that the estimation problem, which is nonlinear in the original signal domain, becomes a linear least squares problem in the transform domain. Additionally, we discuss the estimator's properties in the presence of noise and present a novel technique for mitigating the noise's impact on the estimates. To evaluate the proposed estimation approach, we apply it to a source localization problem where we localize a crack on a metal plate using sensor measurements. Comparing the results with traditional methods, we demonstrate that the proposed approach provides better estimates of the source location.

Next, we introduce a new end-to-end generic signal classification method which utilizes the signed cumulative distribution transform (SCDT), an extension of the CDT. The proposed approach involves defining the classification problem using a transport generative model and exploiting mathematical properties of the SCDT to simplify the problem in the transform domain. To determine the class of an unknown sample, we employ a nearest local subspace (NLS) search algorithm in the SCDT domain. Experiments show that the proposed method provides high accuracy classification results while being computationally cheap, data efficient, and robust to out-of-distribution samples with respect to the existing end-to-end classification methods.

The next section of this thesis explores the feasibility of utilizing generative model-based solutions for structural health monitoring (SHM) applications. Specifically, the focus is on addressing nonlinear system

identification problems. The proposed approach lies somewhere between model-based system identification techniques and data-driven pattern recognition approaches. We presume some basic knowledge of the structural system, namely that it follows an underlying partial differential equation (PDE) model and that damage results in the presence of nonlinearity in the model. It is also presumed that the solution of this PDE model measured at a particular sensor location conforms to a generative model; meaning, it is generated from a template under some unknown deformations. Given the assumption that the sensor data adheres to the generative model, the SCDT nearest local subspace classifier is employed to identify the nonlinearity parameter of a dynamical system. The potential of the proposed approach in the field of structural health monitoring is demonstrated through comparisons with state-of-the-art deep neural networks and Fourier transform-based approaches.

1.4 Dissertation Outline

The organization of the dissertation is as follows: Chapter 2 details a novel technique for parametric signal estimation using the cumulative distribution transform (CDT). This approach aims to minimize the Wasserstein distance between measured and model signals by transforming the nonlinear estimation problem into a linear least-squares problem in the transform domain. Chapter 3 introduces an end-to-end signal classification method using the signed cumulative distribution transform (SCDT). In this work, a nearest local subspace classifier is employed in the SCDT domain to provide an effective solution for signal classification problems. Chapter 4 proposes a transport transform-based mathematical modeling approach for structural health monitoring applications. This chapter explores a novel technique for recovering the parameters of the governing partial differential equation (PDE) of a dynamical system. Lastly, Chapter 5 provides a concise discussion regarding the knowledge learned from this study, and it serves as the concluding chapter of the dissertation.

Chapter 2

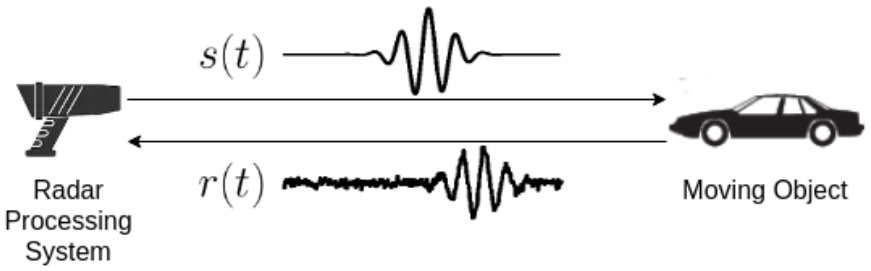
Parametric Signal Estimation Using the Cumulative Distribution Transform

2.1 Introduction

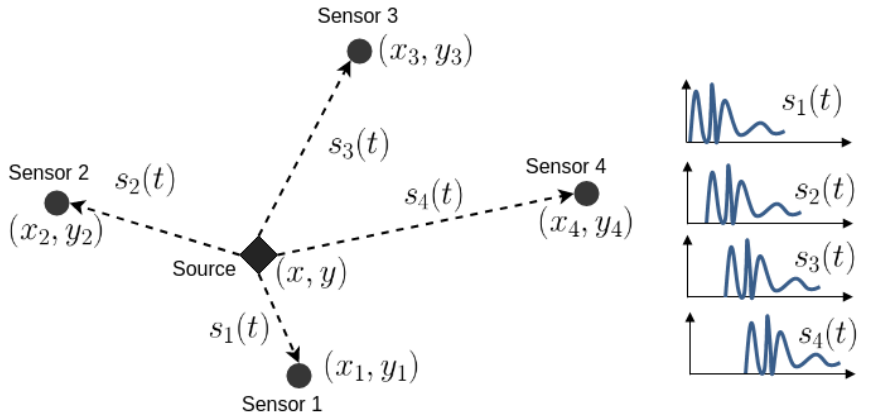
Signal parameter estimation is at the heart of many signal processing applications that involve localisation and tracking of a source signal, e.g. radar [2], underwater acoustics [3], source localization [7] [8], seismology [4] [5], communication [6] etc. All these systems require estimation of the values of a group of parameters. In radar systems, for example, the estimated time delay and Doppler stretch between transmitted and received signals are used to determine the position and speed of a target object (Fig. 2.1a). Fig. 2.1b illustrates another example where the location of an AE (acoustic emission) source is determined using the estimated time difference of arrivals (TDOAs) of the signals received by four sensors. Typical estimation techniques involve maximizing the likelihood function [2], which most often yield to non-convex optimization problems. In this work, we propose to solve the parametric signal estimation problem by minimizing Wasserstein distance between measured and model signals. To solve the ensuing transport problem, we rely on a novel technique called the cumulative distribution transform (CDT) introduced in [61] for the purposes of simplifying the estimation process.

2.1.1 Estimation as a transport problem

We propose to solve certain signal estimation problems borrowing concepts from optimal transport theory [62]. Specifically, we are interested in the case where a strictly positive quantity $s(t)$ is undergoing a param-



(a)



(b)

Figure 2.1: Estimation in signal processing applications: (a) radar system, and (b) source localization.

eterized change of variables

$$s_g(t) = g_p'(t)s(g_p(t)) \tag{2.1}$$

where $g_p(t)$ is a one to one, differentiable function with parameters $\boldsymbol{p}$ (e.g. $g_p(t) = \sum_{k=0}^{K-1} p_k t^k$) and $g_p'(t) = dg_p/dt$. This signal model is particularly pertinent in physical situations where $s(t)$ represents a conserved quantity, e.g., intensity, that evolves in time or space. To emphasize this point we note that Eqn. (2.1) can be restated as:

$$\int_{-\infty}^{t} s_g(u)du = \int_{-\infty}^{g_p(t)} s(u)du \tag{2.2}$$

which underscores the conservation of $s(t)$ under the action of $g_p(t)$. Here, u is the integration variable. Indeed, Eqn. (2.1) is simply a Lagrangian restatement of the continuity equation from continuum mechanics where the function $g_p(t)$ transforms the independent variable according to the problem physics (see e.g., [63, 64]).

Such models are common in wave optics, for example, where Eqn. (2.1) is seen to operate on the squared magnitude of a wavefunction or an electric field (see e.g., Schrodinger equation [65] or paraxial wave equation

[66, 67]). For example, if $s(t)$ represents the time-varying optical intensity of a beam propagating through a lossless medium, $g_p(t)$ captures the influence of the medium to yield the modified intensity $s_g(t)$ [64]. Similar physics can be observed in phase modulated acoustic signals of finite duration (i.e., "pulses") propagating through linear elastic solids [68]. In short, the signal model used in this work is consistent with data collected from a broad and important class of physical systems.

Note that measured signals may or may not conform to the approach outlined above since they may not be strictly positive or conserve energy. These properties can be guaranteed, however, under an appropriate normalization scheme. Denoting the measured signal $z_g(t),\ t \in \Omega_z$, we can associate with this signal a positive probability density function (PDF):

$$s_g(t) = B(z_g)(t) := \frac{z_g^2(t)}{\int_{\Omega_z} z_g^2(t)dt} \tag{2.3}$$

where $B(.)$ is a normalization scheme that transforms raw signals into PDFs. Although this transformation is non-invertible, it guarantees strict positivity and a constant signal energy in accordance with our signal model (2.1). The impact of this normalization scheme on the estimation problem will be discussed in section (2.5). Note that $B(z_g)$ is not one-to-one function of $z_g(t)$, and thus critical phase information may not be retrieved [69]. This would impede us from performing correct estimation in case of estimating the shift of periodic signals (e.g. a pure sign wave); however, in this work we are interested to estimate the parameters of signals that are mostly transient. Therefore our approach does not require recovering $z_g(t)$.

The goal of this work is to illustrate how relationships (2.1) and (2.2) can be leveraged to produce estimates of the parameters $\boldsymbol{p}$ of $g_p(t)$ that governs the modulation or modification of such signals. In particular, we will show how this nonlinear, generally non-convex estimation problem in the time domain can be transformed into a linear least-squares problem in the CDT domain.

2.1.2 Related works

Previously proposed methods include maximizing cross-correlation [16], l_p correlation [17], maximizing the magnitude of difference between measured and template signals [18], entropy [19] and mutual information based methods [20][21]. All these approaches assume that the only difference between the measured and template signals is the time delay, in addition to noise. In real applications, however, signals may undergo complex parameterized changes. In radar related estimation problems, for example, motion of the target object introduces linear dispersion (also called Doppler stretch) along with the time delay. In such cases the above mentioned techniques may produce erroneous estimates.

Several subspace based methods [25] [26] [27] [28] have been proposed to jointly estimate time delay and Doppler parameters. Most of the subspace methods exploit narrowband approximation of the signals so that the Doppler effect can be modeled as frequency shift, and hence, can be estimated explicitly. A *search* over a parameter space is still required for time delay estimation, which is computationally expensive. Colonnese et al. [70] proposed a generalized method of moments (GMM) for estimating shift/translation, i.e. location parameters. Applying the GMM to parameters other than shift requires a transformation function to convert it to a location parameter; for example, a transformation realizing exponential warping of the independent axis can be applied to turn a scale parameter into a location one. Moreover, similar to subspace methods the GMM based approach requires computationally expensive *search* to estimate the parameters. Joint estimation of time delay and Doppler stretch has also been studied in the literature where the estimation techniques involve maximizing the ambiguity function between the measured and the template signals [22] [23] [24]. In most cases, ambiguity function-related techniques yield non-convex optimization problems for which global minima may be difficult to produce. The proposed method described in this work involves minimizing Wasserstein distance which yields to a convex problem.

The Wasserstein metric is a well developed concept in the optimal transport theory [71], which measures the difference between two distributions by the optimal cost of rearranging one distribution into the other. It has been proven to be a suitable tool to model and solve problems in the areas of signal processing and machine learning [62]. Nichols et al. [8] proposed an estimator based on the Wasserstein distance for estimating the time delay, but they did not address the linear dispersion or other forms of transformation. In [72], Engquist and Froese first used this metric in the seismic inversion problems. Then, the idea of using the Wasserstein distance to identify a geophysical model from the observations was exploited in [4] and [73], where the convexity property of the Wasserstein metric in the context of model identification was utilized. In this work, we incorporate the CDT, a new transformation technique, along with the Wasserstein distance so that the estimation problem becomes a linear least squares problem in the transform space.

In a prior work [61] the cumulative distribution transform (CDT) was introduced as a useful means of modeling and subsequently classifying observed data. The CDT is a fundamentally nonlinear mapping of the *locations* of the signal values with respect to a particular reference. Put another way, computations performed in the CDT domain alter the independent variable of the signal(s) to produce a desired effect (e.g., matching one signal to another). The advantages of the CDT include its invertibility, ease of computation, and its ability to render certain classification problems linearly separable in transform space (see again, [61]).

2.1.3 Outline and overview of contributions

The main contribution of this work is to describe a Wasserstein distance minimization-approach to parametric signal estimation. The mathematical approach is aided by the CDT [61] to help develop a generic closed form solution to the problem. The solution is general enough to encompass a variety of mass (signal) transport phenomena. In the following sections, we briefly review the definitions of the Wasserstein distance and the CDT, derive two important lemmas to formulate an expression for the estimator, discuss details of implementation, and introduce a strategy for mitigating the influence of noise on the estimates. We will conclude with both numerical and experimental examples comparing the CDT based estimator against some standard techniques. In section 2.7, we present a source localization example where the location of a crack is determined on a metal plate using the estimated TDOAs of the acoustic signals received by the sensors. Our analysis shows that in most cases the accuracy is improved when using the proposed technique, but most importantly, in all cases the computational cost is *orders of magnitude* lower than competing methods.

2.1.4 Note about notation

Throughout this chapter, we deal with real signals s, r, z etc. assuming these to be square integrable in their respective 1D domains. That is, we assume that $\int_{\Omega_s} |s(t)|^2 dt < \infty$, where Ω_s is the domain over which s is defined. In addition, we at times make use of the common notation: $\|s\|^2 =< s, s >= \int_{\Omega_s} |s(t)|^2 dt$. Some necessary symbols used throughout this chapter are described in Table 2.1.

Table 2.1: Description of symbols

Symbols	Description
s	Normalized, strictly positive signal
S	Cumulative distribution function (CDF) of s
s_0	Reference density function
$\widehat{s}$	Cumulative distribution transform (CDT) of s
s_g	Generated from s under the action of g_p
g_p	One-to-one, continuous function with parameter $\boldsymbol{p}$
z_g	Measured signal (not normalized) in absence of noise
r	$B(z_g + \eta)$; normalized measured signal with noise (η)
B	Normalization scheme to transform raw signals into PDFs
r_f	Generated from r under the action of f_p
f_p	g_p^{-1}
$W(s_1, s_2)$	Wasserstein distance between density functions s_1 and s_2

2.2 Estimation as a Transport Problem

First, let $g_p(t)$ be a differentiable and strictly increasing mapping (i.e. $g'_p = dg_p/dt > 0$) between $\Omega = [0, 1]$ and Ω_z, where $\boldsymbol{p}$ refers to a parameter vector. It is easy to see that g_p is one-to-one and hence invertible. For example, polynomials $g_p(t) = \sum_{k=0}^{K-1} p_k t^k$ of different degrees will be used in the estimation problem described in this work. This polynomial is able to capture events such as time delay and dispersion in the physics of wave propagation. Moreover, such polynomial model of the transformation $g_p(t)$ is commonly used in many signal processing applications [74] [75] [76]. In applications where $g_p(t)$ is unknown, polynomial approximations are often used to model the transformation [77].

The goal in our estimation problem is to then find the parameter vector $\boldsymbol{p}$ that generated some measured, normalized data $r(t) = B(z_g(t) + \eta(t))$ where $\eta(t)$ is a noise process (see again section 2.5). Typical estimation techniques try to solve this problem by finding the parameters of a model, e.g. $s_g(t) = g'_p(t)s(g_p(t))$ that best matches the measured signal $r(t)$ [78] [79]. Alternatively, this problem can also be stated as finding the parameter vector $\boldsymbol{p}$ such that some measure of a 'match' between $r_f(t) = f'_p(t)r(f_p(t))$ and $s(t)$ is maximized, where $f_p(t) = g_p^{-1}(t)$. In this work, we adopt the alternative approach as it helps generating closed form solution (discussed in section 2.4) for our estimation problem even when $g_p(t)$ is a higher order polynomial. Here we propose to solve the signal estimation problem by finding the parameters of $g_p(t)$ such that the Wasserstein distance [62] between $r_f(t)$ and $s(t)$ is minimized:

$$W^2(r_f, s) = \inf_h \int_{\Omega_s} |h(u) - u|^2 s(u) du \tag{2.4}$$

where $W(.,.)$ is the Wasserstein distance between two PDFs and

$$\int_{\inf(\Omega_r)}^{h(t)} f'_p(u) r(f_p(u)) du = \int_{\inf(\Omega_s)}^{t} s(u) du. \tag{2.5}$$

Thus, we have implicitly defined a "match" as the minimum distance $h(u) - u$, for all possible $h(\cdot)$, over which the original signal values $s(t)$ must be moved in order to form $r_f(t)$. The quantity (2.4) features prominently in the field of optimal mass transport where the minimizer is used to define the optimal transport map $h(u)$ for moving the "mass" $s(u)$ over a distance $h(u) - u$ [62].

We note that because we are looking at 1D signals $s(t)$, there is only one h satisfying the equation above. Moreover, by Lemma III.1 (see next section) we will never need to explicitly compute h, instead, its influence is embedded in the respective CDTs of the relevant signals.

In what follows we demonstrate the benefits of defining the cost function in this manner for the parameter estimation problem.

2.3 The Cumulative Distribution Transform

In this section we show that using the cumulative distribution transform (CDT) [61] we can derive a solution to the signal estimation problem expressed in equation (2.4) above. Let $s_0(y)$, $y \in \Omega_{s_0}$ define a reference signal pattern defined on the domain Ω_{s_0} which is in general different from the signal domain Ω_s. Without loss of generality we use $s_0(y) = 1$ and $\Omega_{s_0} = [0, 1]$ in this work. The transform of $s(t)$ is then defined to be the function $\hat{s}(y)$ that solves

$$\int_{\inf(\Omega_s)}^{\hat{s}(y)} s(u)du = \int_{\inf(\Omega_{s_0})}^{y} s_0(u)du. \tag{2.6}$$

Now define the cumulative distribution functions (CDFs) $S(t) = \int_{-\infty}^{t} s(u)du$ and $S_0(y) = \int_{-\infty}^{y} s_0(u)du$. Note that because $s(t), s_0(y) > 0$ for $t \in \Omega_s, y \in \Omega_{s_0}$, it follows that S, S_0 are one to one continuous maps. Furthermore, if s, s_0 are continuous, S, S_0 will be differentiable [61]. Therefore, an alternative expression for $\hat{s}$ is

$$\hat{s}(y) = S^{-1}(S_0(y)). \tag{2.7}$$

The CDT is therefore seen to inherit the domain of the reference signal. Moreover, given our particular choice of reference, $s_0(y) = 1$, $S_0(y) = y$ and $\hat{s}(y) = S^{-1}(y)$. That is to say, *the CDT is the inverse of the cumulative distribution function of* $s(t)$. This definition is similar to the *Quantile Function* [80] [81] in statistics, although the similarity does not hold if non-uniform reference $s_0(y)$ is used to calculate the CDT. The inverse formula can then be defined as

$$s(t) = \left(\hat{s}^{-1}\right)'(t) s_0(\hat{s}^{-1}(t)) \tag{2.8}$$

where $\hat{s}^{-1}(t) = S(t)$. Fig. 2.2 illustrates the process of calculating the CDT for a normalized, strictly positive quantity $s(t)$ and uniform reference signal $s_0(y)$. Physically, the CDT is a coordinate transformation acting on the independent variable (in this case time) in such a way as to preserve the total signal energy while morphing the distribution $s_0(y)$ into the distribution $s(t)$. Note that this definition is slightly different from that used in [61] where the CDT was defined in terms of the coordinate deviation $\hat{s}(y) - y$. In summary, the CDT and inverse CDT map continuous positive PDFs to diffeomorphism, and vice versa [61].

Given these definitions, we can now describe the proposed cost function (2.4) in the CDT domain. The following lemma [8] helps link the Wasserstein distance between s and r, and $\|\hat{s} - \hat{r}\|_{\ell_2}^2$

Lemma 2.3.1. *Let $\hat{s}$ and $\hat{r}$ be the CDTs of s and r, respectively. We then have that $W^2(s, r) = \|\hat{s} - \hat{r}\|_{L_2}^2$.*

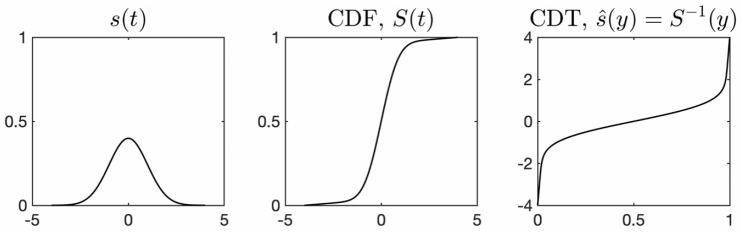


Figure 2.2: Steps of calculating the CDT of a signal $s(t)$, given the uniform reference signal $s_0(y)$.

Proof: Given, $\hat{s}$ and $\hat{r}$ are the CDTs of s and r, respectively. That is:

$$\hat{s}'(y)s(\hat{s}(y)) = \hat{r}'(y)r(\hat{r}(y)) = s_0(y)$$

Let $h'(t)s(h(t)) = r(t)$. Plugging $t = \hat{r}(y)$ into this equation, we have $r(\hat{r}(y)) = h'(\hat{r}(y))s(h(\hat{r}(y)))$ and $\hat{s}(y) = h(\hat{r}(y))$. Consequently

$$\begin{aligned} W_2^2(s,r) &= \int_{\Omega_r} (h(u) - u)^2 r(u) du \\ &= \int_{\Omega_{s_0}} (h(\hat{r}(y)) - \hat{r}(y))^2 \hat{r}'(y) r(\hat{r}(y)) dy \\ &= \int_{\Omega_{s_0}} (\hat{s}(y) - \hat{r}(y))^2 s_0(y) dy \\ &= \|\hat{s} - \hat{r}\|_{L_2}^2 \end{aligned}$$

The lemma above simply states that for 1D signals which are PDFs, the CDT naturally embeds the Wasserstein distance. As in the computations all the signals are discrete, L_2 (norm of functions on real line) will be replaced by ℓ_2 (norm of sequences) in what follows. Therefore, in discrete cases we have,

$$W^2(s,r) = \|\hat{s} - \hat{r}\|_{\ell_2}^2$$

In addition, we also have the following useful functional composition lemma.

Lemma 2.3.2. *Let $\hat{r}$ and $\hat{r}_f$ be the CDTs of signals r and r_f respectively, where $r_f = f_p' r \circ f_p$. The CDT of r_f is then given by $\hat{r}_f = f_p^{-1} \circ \hat{r}$.*

Proof: Consider again the reference signal $s_0(y)$ and the signal $r(t)$. The relationship between these two signals can be defined in terms of CDT as,

$$\int_{\inf(\Omega_r)}^{\hat{r}(y)} r(u) du = \int_{\inf(\Omega_{s_0})}^{y} s_0(u) du. \tag{2.9}$$

Similarly, can we relate $r_f(t)$ and $s_0(y)$ via

$$\int_{\inf(\Omega_{r_f})}^{\hat{r}_f(y)} r_f(u)du = \int_{\inf(\Omega_{s_0})}^{y} s_0(u)du. \tag{2.10}$$

We replace $r_f(t)$ with $f_p'(t)r(f_p(t))$ (where $f_p(t) = g_p^{-1}(t)$), in which case

$$\int_{f_p^{-1}(\inf(\Omega_r))}^{\hat{r}_f(y)} f_p'(u)r(f_p(u))du = \int_{\inf(\Omega_{s_0})}^{y} s_0(u)du. \tag{2.11}$$

Applying the change of variables $f_p(u) = y$ and $dy = f_p'(u)du$ to the left hand side,

$$\int_{\inf(\Omega_r)}^{f_p(\hat{r}_f(y))} r(y)dy = \int_{\inf(\Omega_{s_0})}^{y} s_0(u)du. \tag{2.12}$$

From Eqn. (2.9) and (2.12) it can be stated,

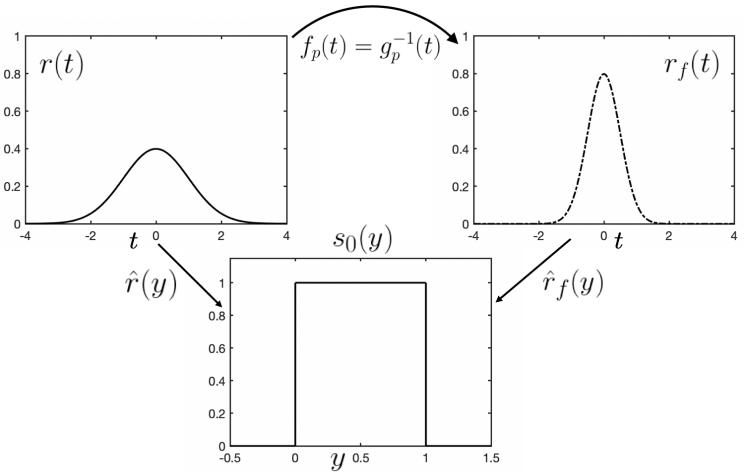
$$\int_{\inf(\Omega_r)}^{f_p(\hat{r}_f(y))} r(y)dy = \int_{\inf(\Omega_r)}^{\hat{r}(y)} r(u)du. \tag{2.13}$$

For this statement to be true, the upper bound of the left hand side must be equal to the upper bound of the right hand side, i.e. $f_p(\hat{r}_f(y)) = \hat{r}(y)$. Since $f_p(t) = g_p^{-1}(t)$ and $g_p(t)$ are invertible, we can finally write $\hat{r}_f = f_p^{-1} \circ \hat{r} = g_p \circ \hat{r}$.

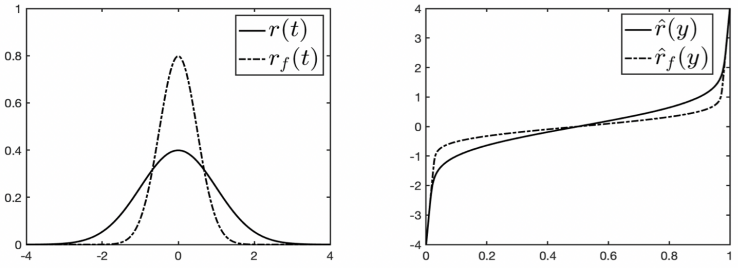
Simply stated, the CDT composition lemma says that changes along the independent variable (e.g., shifts in time $t - \tau$ or dispersions ωt) become changes in the dependent variable in transform domain (refer to Fig. 2.3b). Fig. 2.3 illustrates the relationships between the signals and the transforms for a signal undergoing parametric change. Each of the constituent CDTs transforms their respective signals into the reference signal $s_0(y)$. Similarly, the mapping $f_p(t)$ transforms $r(t)$ into $r_f(t)$.

2.3.1 Numerical Implementation of the CDT

Recall that the CDT is defined for continuous-time signals in contiguous, finite domain. Here we describe the numerical method for approximating the CDT given discrete data. As the CDT $\hat{s}(y)$ is the inverse of the CDF of $s(t)$ for a particular choice of reference signal ($s_0(y) = 1$ for $y \in [0, 1]$), we need to compute the cumulative function first.



(a)



(b)

Figure 2.3: Example case relating the different signals and transforms used in the CDT. Let s_0 be the reference signal, r be the measured signal, and r_f be the manipulated signal. The transforms and their directions are also given. Plots in (b) show that the transformation (linear dispersion in this case) along the independent axis in signal space becomes a transformation along the dependent axis in CDT space.

Let $\boldsymbol{s} = [s_1, s_2, ..., s_N]^T$ be a N-point discrete-time signal, where $s[n] = s_n, \forall n = 1, 2, ..., N$ is the n^{th} sample of $\boldsymbol{s}$. Then the numerical approximation of the cumulative function is given by,

$$S[n] = \sum_{i=1}^{n} s[n], \qquad n = 1, 2, ..., N \tag{2.14}$$

The CDT is then calculated by taking the inverse of the CDF using interpolation.

2.4 Signal Estimation in CDT Domain

This section will demonstrate the use of CDT in estimating signal parameters. We will highlight the relevance of the transform with respect to time delay, linear dispersion, and quadratic dispersion. For each case, we

will leverage lemmas 2.3.1 and 2.3.2 above, so that the cost function in (2.4) becomes

$$W^2(r_f, s) = \|f_p^{-1} \circ \hat{r} - \hat{s}\|_{\ell_2}^2 = \|g_p \circ \hat{r} - \hat{s}\|_{\ell_2}^2. \tag{2.15}$$

It is evident that when $g_p(t)$ is a monotonically increasing polynomial, the problem above is simply a linear least squares problem. The advantage of signal estimation in the CDT domain is that only the function $g_p(t)$ needs to be computed before proceeding to the estimation problem. We note the following specific examples that can be derived from the above.

2.4.1 Time delay estimation

In the time delay estimation problem $g_p(t) = t - \tau$, hence the cost function (2.15) becomes,

$$W^2(r_f, s) = \|\hat{r} - \tau - \hat{s}\|_{\ell_2}^2.$$

The translation value τ that minimizes the equation above is then given by:

$$\widetilde{\tau} = \frac{1}{|\Omega_{s_0}|} \int_{\Omega_{s_0}} [\hat{r}(u) - \hat{s}(u)]\, du \tag{2.16}$$

Note that, as already mentioned in [8] the problem above is convex on τ, hence a global minimizer is possible and given in closed form. Furthermore, utilizing the fact that the center of mass of s can be estimated by $\mu_s = \int_{\Omega_s} ts(t)dt = \int_{\Omega_s} 1 - S(t)dt$ one can also show that $\frac{1}{|\Omega_{s_0}|} \int_{\Omega_{s_0}} \hat{s}(u)du = \mu_s$ and thus the solution of the time delay problem is also given by

$$\widetilde{\tau} = \mu_r - \mu_s.$$

2.4.2 Linear Dispersion Estimation

In the linear dispersion problem we have that $g_p(t) = \omega t$, and thus

$$W^2(r_f, s) = \|\omega\hat{r} - \hat{s}\|_{\ell_2}^2. \tag{2.17}$$

This problem is convex on ω and possesses closed form solution. The minimizer for the equation above (following linear least squares on ω) is

$$\widetilde{\omega} = \frac{< \hat{s}, \hat{r} >}{\|\hat{r}\|_{\ell_2}^2} \tag{2.18}$$

where $< \cdot, \cdot >$ denotes the inner product.

2.4.3 Time delay and Linear Dispersion estimation

In the joint estimation of time delay and linear dispersion we have that $g_p(t) = \omega t - \tau$, thus $g_p \circ \hat{r} = \omega \hat{r} - \tau$. Hence, the cost function (2.15) becomes,

$$W^2(r_f, s) = \|\omega \hat{r} - \tau - \hat{s}\|_{\ell_2}^2 = \|\alpha \hat{r} + \beta - \hat{s}\|_{\ell_2}^2 \tag{2.19}$$

where, $\alpha = \omega$, and $\beta = -\tau$. Once again, this is a linear least squares problem, from which ω and τ can readily be recovered. The closed form solution to this problem is given by,

$$\left[\widetilde{\alpha}, \widetilde{\beta}\right]^T = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \hat{s} \tag{2.20}$$

where $\mathbf{X} \equiv \left[\vec{\hat{r}}, \vec{\mathbf{1}}\right]$ is an $N \times 2$ matrix.

2.4.4 Quadratic dispersion estimation

In the quadratic dispersion estimation problem we have $g_p(t) = \kappa t^2$, so $g_p \circ \hat{r} = \kappa \hat{r}^2$. Consequently

$$W^2(r_f, s) = \|\kappa \hat{r}^2 - \hat{s}\|_{\ell_2}^2. \tag{2.21}$$

Again we have convexity with the solution being

$$\widetilde{\kappa} = \frac{< \hat{s}, \hat{r}^2 >}{\|\hat{r}^2\|_{\ell_2}^2} \tag{2.22}$$

2.4.5 Quadratic dispersion with time delay

The quadratic dispersion with time delay can be expressed as $g_p(t) = \kappa t^2 - \tau$. Therefore the Wasserstein distance is:

$$W^2(r_f, s) = \|\kappa \hat{r}^2 - \tau - \hat{s}\|_{\ell_2}^2 = \|\alpha \hat{r}^2 + \beta - \hat{s}\|_{\ell_2}^2 \tag{2.23}$$

Similar to joint time delay and linear dispersion estimation described in 2.4.3, this problem is also convex and possesses closed form solution which is given by

$$\left[\widetilde{\alpha}, \widetilde{\beta}\right]^T = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \hat{s} \tag{2.24}$$

where $\mathbf{X} \equiv \left[\vec{\hat{r}^2}, \vec{\mathbf{1}}\right]$, $\alpha = \kappa$, and $\beta = -\tau$.

2.4.6 Higher order polynomial

For any general polynomial, i.e. $g_p(t) = \sum_{k=0}^{K-1} p_k t^k$, $g_p \circ \hat{r} = \sum_{k=0}^{K-1} p_k \hat{r}^k$. Thus, the cost function in (2.15) becomes,

$$W^2(r_f, s) = \| \sum_{k=0}^{K-1} p_k \hat{r}^k - \hat{s} \|_{\ell_2}^2 \tag{2.25}$$

which can be stated as a linear least squares problem,

$$\widetilde{\vec{p}} = \underset{\vec{p}}{\operatorname{argmin}} \|\mathbf{X}\vec{p} - \vec{\hat{s}}\|_{\ell_2}^2 \tag{2.26}$$

where $\vec{p} = [p_0, p_1, ..., p_{K-1}]^T$, $\mathbf{X} \equiv \left[\vec{\mathbf{1}}, \vec{\hat{r}}, \vec{\hat{r}}^2, ..., \vec{\hat{r}}^{K-1}\right]$. The Hessian of (2.26) is $2\mathbf{X}^T\mathbf{X}$ which is positive semi-definite. Therefore the estimation problem described in equation (2.26) is convex. Moreover, the columns of matrix $\mathbf{X}$ are linearly independent, that means Hessian is positive definite and $\mathbf{X}^T\mathbf{X}$ is invertible. Hence, (2.26) possesses closed form solution which is given by,

$$\widetilde{\vec{p}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{s} \tag{2.27}$$

In this section, we have shown that the estimation problem, while non-linear in time domain, can be transformed into a linear least squares problem with closed form solution using CDT. In the next section we will address the influence of noise in the estimation process and a strategy to mitigate it.

2.5 Signal Estimation in Noise

In the previous sections, we defined the CDT, provided the relationship between the CDTs of signals related by a transformation of the independent variable, and then demonstrated linearity of the Wasserstein cost function with respect to the signal parameters that define several such transformations. These relationships were derived without explicit consideration of the corrupting noise source and how it influences the associated estimation problem.

In this section, we consider the impact of additive Gaussian noise on the CDT and on the subsequent parameter estimation. Assume the received signal is corrupted by zero mean, i.i.d Gaussian noise values, $\eta(t) \sim \mathcal{N}(0, \sigma^2)$ so that the measured data are $z_\eta(t) = z_g(t) + \eta(t)$ and

$$r(t) = B(z_\eta)(t) = \frac{(z_g(t) + \eta(t))^2}{\|z_g(t) + \eta(t)\|_{\ell_2}^2}. \tag{2.28}$$

The normalization therefore results in three terms, two of which involve the signal noise. This additional signal "mass" alters the Wassserstein distance and biases the resulting signal parameter estimates. In what follows we propose a simple solution for removing the influence of the additive noise directly in the signals' CDFs.

2.5.1 Noise corrected CDF

Using Eqn. (2.28) as a starting point, in expectation the effects of additive, zero-mean Gaussian noise on the CDF are modelled as (detailed in A.1)

$$E\left[R(t)\right] = \frac{\mathcal{E}_z S_g(t) + \sigma^2(t - t_1)}{\mathcal{E}_z + \sigma^2(t_N - t_1)}, \quad t_1 \leq t \leq t_N. \tag{2.29}$$

Here, $S_g(t)$ and $R(t)$ are the CDFs associated with $s_g(t) = B(z_g)(t)$ and $r(t) = B(z_\eta)(t)$ respectively and the term $\mathcal{E}_z = \|z_g(t)\|_{\ell_2}^2$ is the total energy of the noise free signal. An expression for the noise corrected CDF in expectation is then obtained from Eqn. (2.29) as

$$\widetilde{S}_g(t) = \frac{E[R(t)]\{\mathcal{E}_z + \sigma^2(t_N - t_1)\} - \sigma^2(t - t_1)}{\mathcal{E}_z} \tag{2.30}$$

where $\mathcal{E}_z + \sigma^2(t_N - t_1)$ is the expected energy of noisy signal (see A.1). Alternatively we can define the signal-to-noise ratio $SNR = \mathcal{E}_z/\sigma^2(t_N - t_1)$ in which case (2.30) becomes

$$\widetilde{S}_g(t) = \frac{E[R(t)]\left[SNR + 1\right] - \frac{t - t_1}{t_N - t_1}}{SNR}, \quad t_1 \leq t \leq t_N \tag{2.31}$$

In short, the influence of additive, i.i.d noise is seen as the addition of a constant slope to the CDF. Moreover, under our chosen normalization scheme (2.3), this slope is the noise variance. Thus, a simple strategy for denoising in the CDT domain is to first estimate σ^2 using a "noise only" portion of the signal, and then apply Eqn. (2.30). This method is effectively filtering the signal in the CDF domain.

To illustrate, consider a Gaussian pulse subject to the coordinate transformation $g_p(t) = \omega t - \tau$ with $\omega = 2$, $\tau = 2$. The noise free input signal PDF is therefore $s(t) = A^2 \exp(-t^2/2b_w^2)$ which, after the transformation, becomes $s_g(t) = A^2\omega \exp(-(\omega t - \tau)^2/2b_w^2)$. For this example the corresponding CDFs can be determined analytically and are shown in Fig. 2.4 for $A = b_w = 1$. The noisy signal was taken as $z_g(t) = s_g^{1/2}(t) + \eta(t)$ where each $\eta(t) \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.15$. The associated CDF $R(t)$ and noise corrected version ($\widetilde{S}_g(t)$) are also shown. The SNR for this example was taken as $SNR = \mathcal{E}_z/(\sigma^2(t_N - t_1)) = 4$. The noise corrected CDF ($\widetilde{S}_g(t)$) is seen to match almost exactly the true CDF ($S_g(t)$).

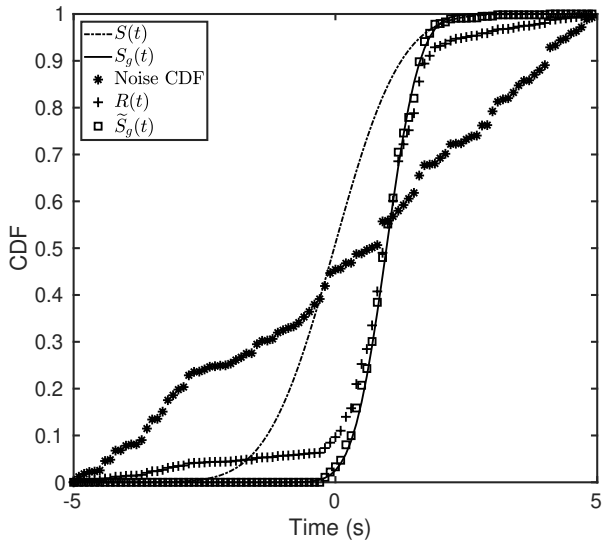


Figure 2.4: CDFs associated with a Gaussian pulse before, $S(t)$, and after, $S_g(t)$, transformation by $g_p(t)$. Also shown are the noisy, $R(t)$, and noise corrected, $\widetilde{S}_g(t)$, CDFs. The true and corrected CDFs match almost exactly, even for this relatively high ($SNR = 4$) level of noise.

2.5.2 Distribution of the CDT

Even after the expected noise is removed via (2.30), there will remain residual fluctuations that will impact our parameter estimates. The joint PDF of the values $\hat{r}(y)$ will dictate the degree to which the cost function (2.15) can be expected to produce good estimates.

Following the derivation described in A.2, the distribution for the CDT values associated with each observation in the signal $z_\eta(t_k),\ k = 1 \cdots N$ is shown to be approximated by the PDF

$$p_{\hat{R}_k}(\hat{r}_k) = \frac{e^{-\frac{(S(\hat{r}_k)-S(\hat{s}_k))^2}{2\Sigma^2(\hat{s}_k)}}}{\sqrt{2\pi}\Sigma(\hat{s}_k)} \frac{\partial S(\hat{r}_k)}{\partial \hat{r}_k},$$
$$1 \leq k \leq N. \tag{2.32}$$

where the variance

$$\Sigma^2(t_k) = \frac{\sigma^4\,(2k + 4\lambda_k)}{\mathcal{E}_{z_\eta}^2} \tag{2.33}$$

is a function of the total noisy signal energy, $\mathcal{E}_{z_\eta} = \sum_k z_\eta^2(t_k)$, and the cumulative sum of the noise-free signal, $\lambda_k = \sigma^{-2} \sum_{i=1}^{k} z^2(t_i)$. Note that both mean and variance in (2.32) are evaluated at the fixed, noise-free CDT values $\hat{s}_k$ (as opposed to the independent variable $\hat{r}_k$). Both can be obtained by simply interpolating

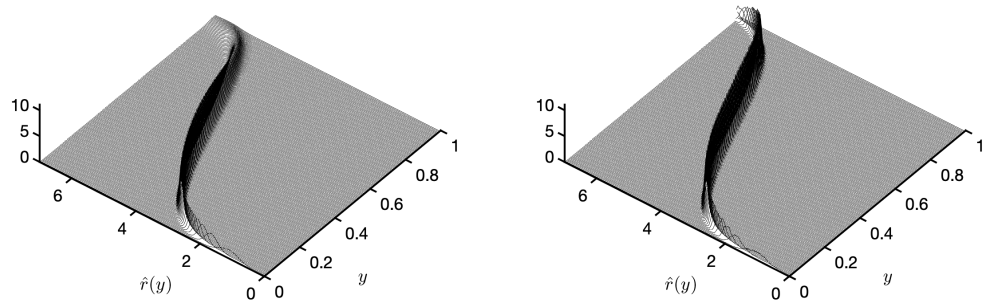
the functions $S(t_k),\ \Sigma(t_k) \rightarrow S(\hat{s}_k),\ \Sigma(\hat{s}_k)$. The resulting distribution is a peaked function that is largely symmetric and centered on the noise-free CDT $\hat{s}(y)$.

As an example, consider signals for which the CDF is well-approximated by the logistic model

$$S(t) = \frac{1}{1 + e^{-at+b}}. \tag{2.34}$$

This is the exact CDF for the logistic distribution, however has also been used to model other CDFs [82]. For appropriate choice of $a,\ b$ this model almost exactly captures the behavior of the CDF for signals such as those shown in Fig. 2.4. Using the general expression (2.32) with this logistic model, we can readily obtain the distribution of CDT values.

We also examined the empirical distribution of the $\hat{r}(y)$ via Monte Carlo simulation. To this end we simulated 1500 realizations of the Gaussian pulse of the previous example, using the parameter values $\omega = 2,\ \tau = 9,\ A = 1,\ b_w = 2$. The associated normalized signals $r(t)$ each consisted of $N = 200$ points sampled at $dt = 0.05s$ and where each additive noise value was taken independently from $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 0.023$. The CDT $\hat{r}(y)$ was then estimated for each realization. The resulting empirical PDF of the CDT values is shown in Fig. 2.5 along with the predicted distribution given by Eqn. (2.32) with $a = 1.68,\ b = 7.54$ and plotted as a function of the dimensionless reference variable $y_k = k/N,\ k = 1 \cdots N$.



(a) Theoretical distributions (given a logistic CDF)

(b) Empirically obtained distributions

Figure 2.5: (a) Theoretical and (b) empirically obtained distributions for the CDT $\hat{r}(y)$ given a logistic CDF (Eqn. (2.34)) consistent with the example of Fig. 2.4. The distributions appear peaked with a single maximum and are well-approximated by a Gaussian function.

The analytical distribution captures precisely the behavior shown in empirically in Fig. 2.4, that is to say, the distributions for CDT values near $y = 0,\ 1$ are skewed while those near the middle of the CDT curve are approximately normally distributed. The difference between simulation and theory near $y = 1$ is due to the fact that the inverse logistic (logit) transformation is only valid for $0 < R(t) < 1$, however our derivation

(described in A.2) assumed constant noisy signal energy, leaving the possibility of values $R(t) > 1$ near the end of the signal.

Nonetheless, the uncertainty in the estimate is indeed well-approximated by a Gaussian distribution as evidenced by Fig. 2.5 and the functional form of (2.32). Moreover, for signals in additive, i.i.d. Gaussian noise, the proposed cost function (2.4) yields a maximum likelihood estimate (MLE) of the parameters. As such, the estimate is guaranteed to reach the Cramer-Rao lower bound (CRLB) asymptotically as $N \to \infty$. The CRLB places a lower bound on the covariance of the parameter estimates and is given by

$$C(p_i, p_j) = \mathbf{F}(\vec{p})^{-1} \tag{2.35}$$

where $\mathbf{F}(\vec{p})$ is the Fisher Information Matrix (FIM). For signals $z_g(t)$ in additive, i.i.d. Gaussian noise the FIM is defined as [2]

$$\begin{aligned} \mathbf{F}(\vec{p}) &\equiv F_{ij} \\ &= \frac{1}{\sigma^2 \Delta t} \left(\int_{t_s}^{t_f} \frac{\partial z_g(t)}{\partial p_i} dt \right) \left(\int_{t_s}^{t_f} \frac{\partial z_g(t)}{\partial p_j} dt \right). \end{aligned} \tag{2.36}$$

Carrying out the integrals and subsequent inversion for the Gaussian pulse of the previous example then yields the full covariance matrix (2.35). We are most interested in the diagonal elements $C(p_i, p_i),\ i = 1 \cdots P$ as these represent the variances of the associated parameter estimates. In what follows we compare our estimates to the CRLB for several of the cases described in the prior section.

2.6 Impact of SNR on Quality of the Estimator

Here we explore the quality of the CDT-based estimation procedure through a series of numerical experiments. Specifically, we compare the quality of the various estimators described in section (2.4) in terms of mean square error (MSE) as a function of SNR.

The signal of interest is taken as the apodized sinusoid

$$z(t) = Ae^{-(t-t_c)^2/(2b_w^2)} \sin(2\pi f t) \tag{2.37}$$

of width b_w and frequency f. The SNR can be well-approximated by

$$SNR = \frac{A^2 \sqrt{\pi} b_w}{2\sigma^2 T} \tag{2.38}$$

so long as the signal length T is large enough to include the entire non-zero portion of the pulse envelope. In our examples we will take $A = b_w = f = 1$ and set $t_c = 0$. We are again assuming that the received signals are corrupted by zero mean additive Gaussian noise, $\eta(t) \sim \mathcal{N}(0, \sigma^2)$.

2.6.1 Time Delay Estimation

From Eqn. (2.16) one can estimate the delay as the difference in the average values of the CDTs $\hat{r},\ \hat{s}$ taken over the domain $\Omega_{s_0} = [0, 1]$. Computationally, we have simply

$$\widetilde{\tau} = \frac{1}{N} \sum_{i=1}^{N} (\hat{r}(u_i) - \hat{s}(u_i)) \tag{2.39}$$

where the CDTs are defined on the discrete grid $u_i = i/N,\ i = 1 \cdots N$. These estimates can then be compared to those obtained via the more familiar cross-correlation estimator applied in the time domain [8]. To this end we simulated 1000 realizations of the signal $z_n(t)$ for varying noise levels and a delay of $\tau = 0.2575\ s$. The linear dispersion was fixed at $\omega = 1$.

To evaluate the performance of the estimator we compute mean squared error (MSE) and compare the results with cross-correlation (XC) based estimator. Although cross-correlation is known to be an MLE for delay estimates in additive Gaussian noise [20], it is a discrete estimator. To implement a continuous delay estimator, an optimization problem is designed that provides maximum likelihood estimates,

$$\widetilde{\tau} = \underset{\tau}{\operatorname{argmax}} \sum_{i=0}^{N-1} z_\eta(t_i) z(t_i - \tau) \tag{2.40}$$

To solve this optimization problem we exploit a gradient based nonlinear programming solver *fmincon* in MATLAB [83]. As *fmincon* solves minimization problems, equation (2.40) is written as,

$$\widetilde{\tau} = \underset{\tau}{\operatorname{argmin}} \quad - \sum_{i=0}^{N-1} z_\eta(t_i) z(t_i - \tau) \tag{2.41}$$

As equation (2.41) is a non-convex problem, this gradient based solver may get stuck in local minima. To resolve this issue, another MATLAB function *GlobalSearch* can be integrated, which repeatedly runs local solver *fmincon* with random starting point to generate global optimal solution. Therefore, in our experiments two approaches are adopted to solve this optimization problem: (i) using *fmincon* only, and (ii) using *GlobalSearch* and *fmincon* together. To compare with a subspace based method, the ESPRIT (estimation of signal parameters via rotational invariance techniques) based time delay estimation technique has been implemented following the approach described in [25]. The MSE for different delay estimators are plotted in Fig. 2.6. We

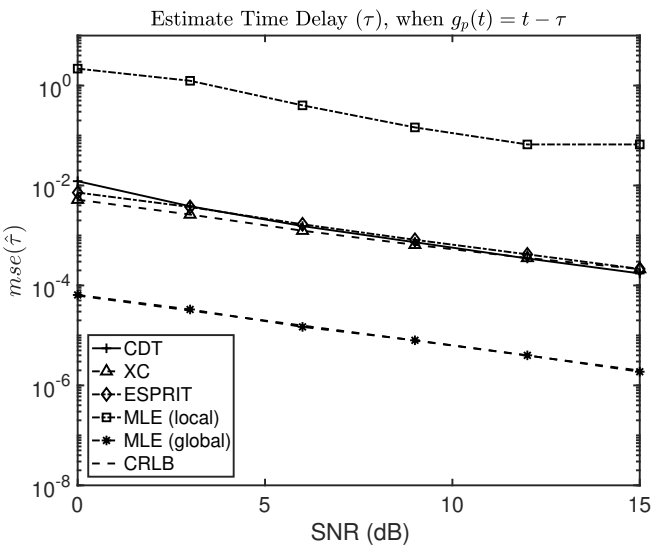


Figure 2.6: MSE in the delay estimates as a function of SNR. For comparison, the estimates produced using MLE, XC, ESPRIT, and the CRLB are also shown. In case of MLE, results for both local and global solvers are plotted.

also plot the CRLB as baseline. The plot shows that the performance of the proposed CDT based estimator is similar to the cross-correlation and ESPRIT estimators, although none of the techniques have reached the CR bound. The MLE obtained via global optimum search reaches the bound, but the estimation using local solver without *GlobalSearch* shows very poor performance.

2.6.2 Joint Estimation of Time Delay and Linear Dispersion

In this example we considered the joint estimation problem for both time delay $\tau = 0.2575s$ and linear dispersion (time scale) $\omega = 0.75$. Again, the estimation problem (2.19) possesses the closed form solution given in section 2.4.3. The MSE of the joint delay and linear dispersion estimates for different estimators are plotted in Fig. 2.7. For comparison, the cross-correlation and ESPRIT based estimators are used again to estimate the delay parameter only. In this case, both XC and ESPRIT estimators perform poorly as these techniques do not take linear dispersion into account. Another subspace based method, the MUSIC (multiple signal classification) algorithm discussed in [25], has been implemented to estimate time delay when both delay and dispersion are present. The time delay estimates from the ESPRIT estimator have been used in the initialization stage of this algorithm. Although this method shows very good performance in significantly high SNR (15 dB), the proposed technique outperforms it in highly noisy cases. Again, as this algorithm requires narrowband approximation of the signal, it gives incorrect estimates of the linear dispersion parameter (ω) for transient signals used in these experiments. Hence, only time delay (τ) estimates are reported for MUSIC

algorithm in Fig. 2.7. For joint time delay and linear dispersion estimation, the another commonly used approach is to locate the peak of Wide-band Ambiguity Function (WBAF) of the received signal [22] [23] [24]. The WBAF between the measured signal $z_\eta(t)$ and the known signal $z(t)$ is given by [22],

$$A_{z_\eta,z}(\tau,\omega) = \sqrt{\omega}\int_{-\infty}^{\infty} z_\eta(t)z^*(\omega t-\tau)dt \tag{2.42}$$

where $(*)$ denotes the complex conjugation which does not have any impact in our experiments as real valued signals are used. Then the joint estimates of τ and ω are given by,

$$\widetilde{\tau},\widetilde{\omega} = \underset{\tau,\omega}{\operatorname{argmax}} \quad |A_{z_\eta,z}(\tau,\omega)|^2$$

which can also be written as,

$$\widetilde{\tau},\widetilde{\omega} = \underset{\tau,\omega}{\operatorname{argmin}} \quad -|A_{z_\eta,z}(\tau,\omega)|^2 \tag{2.43}$$

Similar to time delay estimation discussed in 2.6.1, both the local and global solvers have been exploited to estimate optimum τ and ω using equation (2.43). While global solver performs better than the local solver, the proposed CDT based estimator yields better estimation of the delay parameter (τ) than both solvers (Fig. 2.7). In case of linear dispersion (ω) estimation, it does not outperform the global solver, but the results are still competitive. In both (τ and ω estimates) cases, CDT estimator outperforms local solver based estimator.

The concept of convexity can help understand the superior performance of the proposed technique over the WBAF based estimator. Fig. 2.8 illustrates that in CDT based estimator we are dealing with a convex problem while $-|A_{z_\eta,z}(\tau,\omega)|^2$ is clearly non-convex with several local minima. Although *GlobalSearch* is designed to handle this kind of problem, it is not always accurate to find the global minimal point. It should also be noted that global solver is computationally very expensive as it runs several local solvers. The solution provided by the CDT based estimator, on the other hand, is closed form, hence computational cost is very low. Even running single local solver takes more time than the proposed estimator. The MUSIC algorithm also requires an iterative search over parameter space. Moreover, it uses the result of ESPRIT algorithm for initialization, which contributes to the computational cost. As a result, MUSIC based estimation is also a computationally expensive technique relative to the proposed approach. Fig. 2.9 shows the average times taken by CDT, MUSIC, and WBAF based estimators to jointly estimate the time delay (τ) and the linear dispersion (ω) parameters.

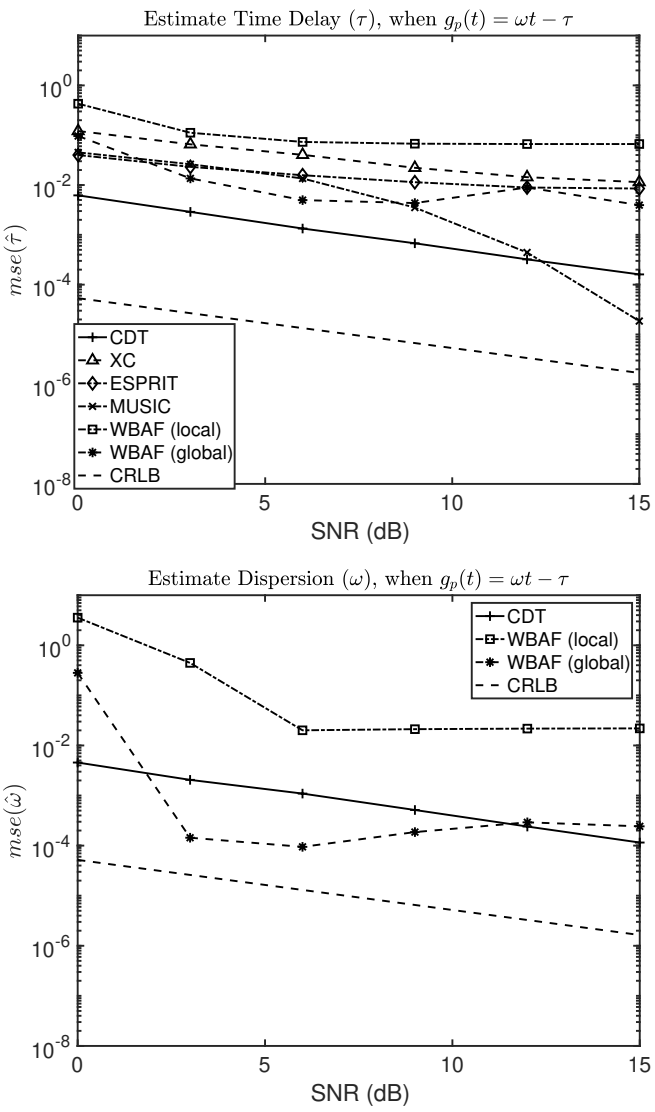


Figure 2.7: MSE in the joint time delay (top) and linear dispersion (bottom) estimates obtained via linear least squares in the CDT domain as compared to the CRLB, XC, ESPRIT, MUSIC, and WBAF (using both local and global solvers). In case of XC, ESPRIT, and MUSIC algorithms, only delay estimates are reported.

2.6.3 Quadratic Dispersion & Delay

As a final illustration we consider the problem of jointly estimating both the quadratic dispersion coefficient (κ) and time delay (τ), i.e. $g_p(t) = \kappa t^2 - \tau$ with $\kappa = 0.5$ and $\tau = 1.2575s$. As discussed in section 2.6.3, the proposed estimation approach described by equation (2.23) possesses closed form solution which is given by equation (2.24). Fig. 2.10 shows the MSE of the estimates of time delay and quadratic dispersion coefficients jointly estimated using the proposed approach. The plots show that CDT based estimator could not reach CR bounds. But the performance of time delay estimation using proposed estimator is better than the cross-correlation based estimator, as cross-correlation does not correct the effect of quadratic dispersion.

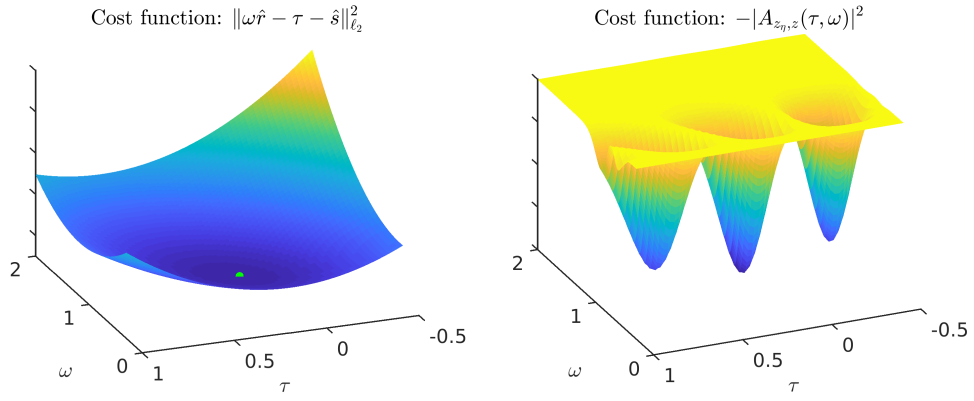


Figure 2.8: Cost functions associated with (a) proposed CDT based estimator (green dot shows the global minimum point), and (b) joint time delay and linear dispersion estimation using WBAF.

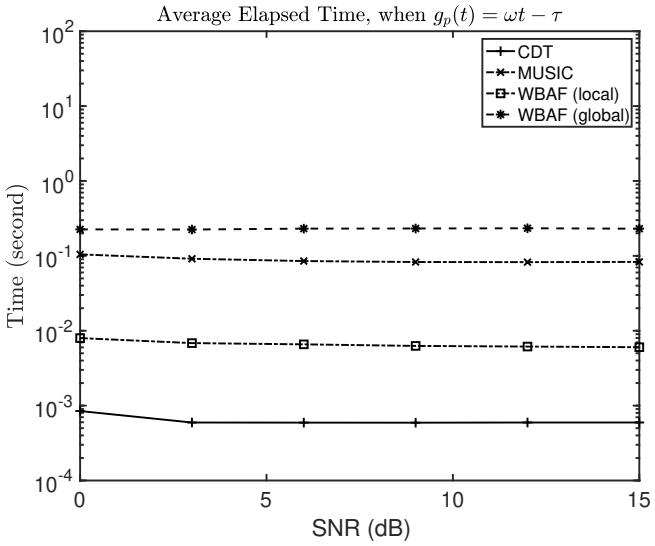


Figure 2.9: Average elapsed time for CDT, MUSIC, and WBAF based estimators. Experiments were run using MATLAB version: 9.4.0 (R2018a) on a computer with an Intel Xeon(R) CPU E5-2630 v3 processor running at 2.40 GHz using 32 GB of RAM.

2.7 Application: Source Localization

As we have mentioned, the estimation approach we have described is appropriate for signals undergoing an invertible transformation of the independent variable, i.e, $z(t) \to g_p'(t)z(g_p(t))$. This is a reasonable model in situations where the signal becomes distorted as it propagates through a medium.

One such situation is the propagation of acoustic signals in solids. Fig. 2.11 shows a metal plate with crack emanating from the end of a horizontal "slot". As the crack propagates it gives off acoustic emissions, loosely defined as a spatially localized release of energy. The result is a short elastic wave "pulse", similar to those used in the preceding numerical examples. By measuring these pulses at different locations on the

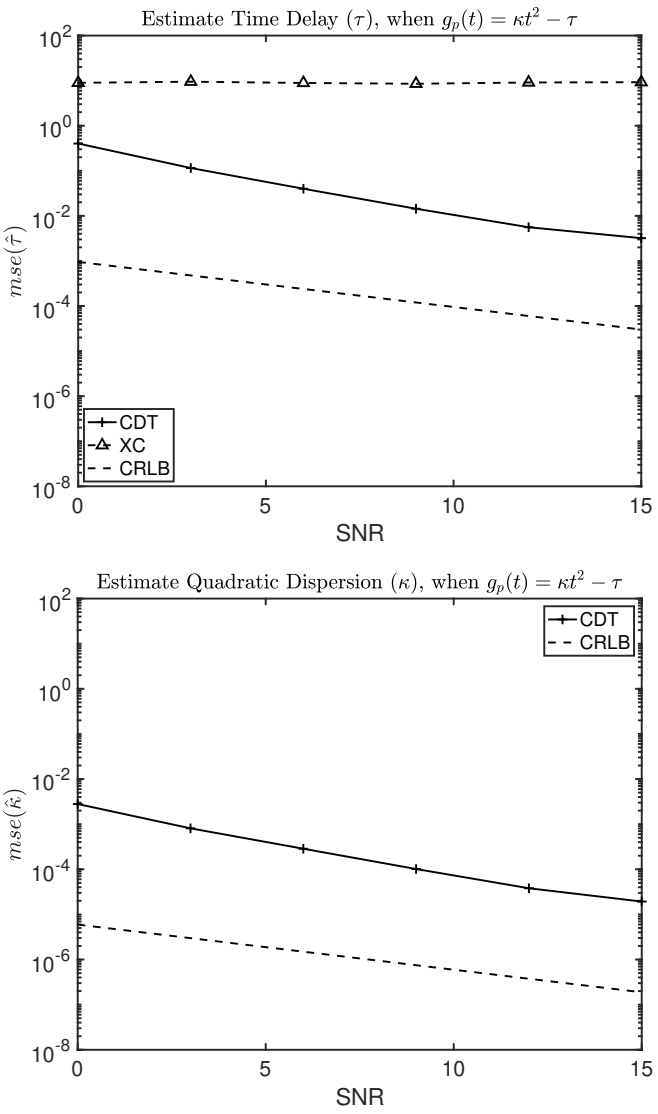


Figure 2.10: MSE associated with the joint estimates of time delay (top) and quadratic dispersion (bottom) parameters

plate and estimating time difference of arrival one can in principle localize the source i.e., the crack tip. In this experiment we used four fiber-optic strain sensors arranged in a diamond pattern, (see Fig. 2.11) [84]. Sample time series from an acoustic emission event are also shown in Fig 2.11.

The challenge is that such signals are difficult to detect and are affected by more than just a time delay during propagation. For example, dispersion is known to influence such pulses during transit [85]. By including dispersion in the model we hypothesize an improved ability to estimate the time delay. Moreover, because this estimation problem reduces to linear least-squares in the CDT domain, the inclusion of this additional term incurs no computational penalty (see again Fig. 2.9)

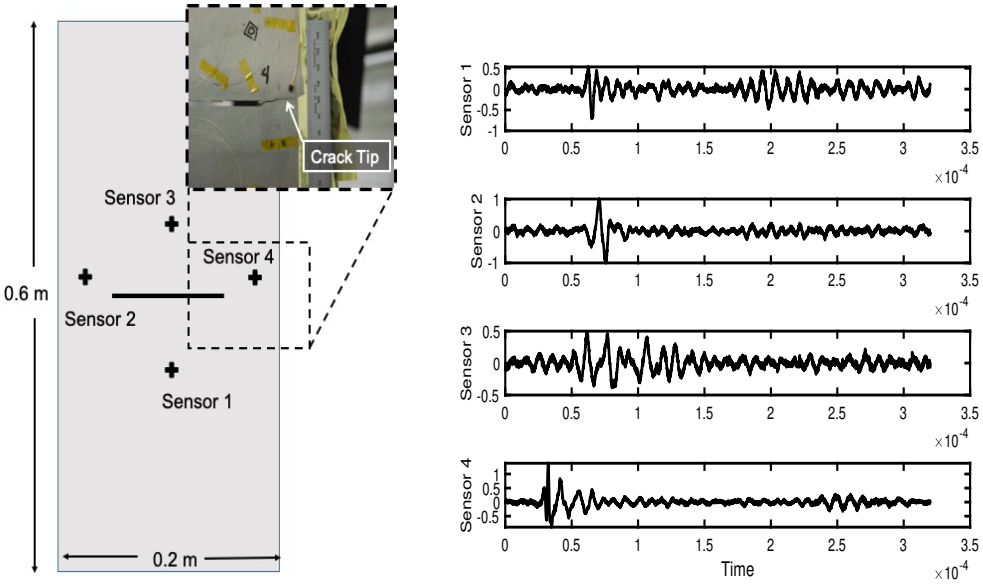


Figure 2.11: (left) Slotted aluminum plate with a crack emanating from the right end of the slot. As the crack propagates it gives off acoustic emission pulses which can be measured at different locations on the plate (right). The time delay of arrival between recorded pulses can then provide the location of the crack tip.

Using the estimation procedure outlined herein, we estimated time delay of arrival among the four sensors. Based on the obtained delays, we then used the source localization algorithm described in [86] to estimate the location of the crack tip. Fig. 2.12 shows the results of these estimates for six different data sets. Specifically, we show the cost function associated with the localization algorithm for a typical realization, along with the estimated minimum which should denote the location of the acoustic emission (i.e., the crack tip). The localization algorithm depends on delay estimates among the four sensors shown as black numbers. To obtain the required delay estimates we used the delay-only estimator (section 2.6.1) as well as the joint delay and linear dispersion estimator (section 2.6.2).

For each of the six data sets, the addition of linear dispersion in the signal model yielded a more well-defined cost function and provided modest improvement in localizing the source of the emission. In fact, one of the location estimates obtained using the "delay only" approach placed the crack tip at the edge of the plate. The source localization results are compared to those obtained using the cross-correlation and the WBAF based estimators. Fig. 2.12 shows that the proposed CDT based technique outperforms both the estimators in estimating the location of the crack tip.

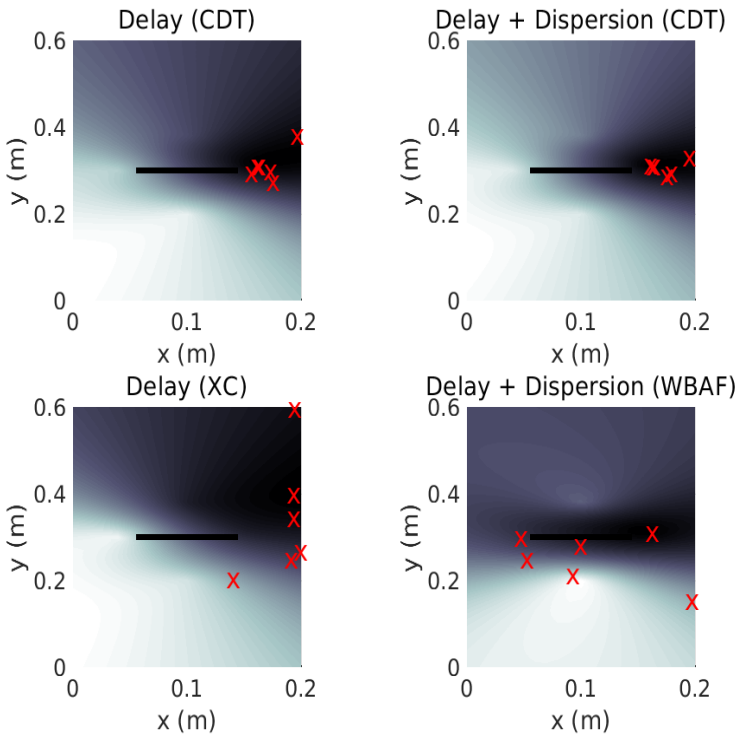


Figure 2.12: Typical cost function for the source localization problem superimposed on the physical plate dimensions. In this case the estimated source location provides the crack tip location (denoted 'X'). Four fiber-optic strain sensors record the data and the location estimate is based on the delay estimates among the sensors as obtained via the cross-correlation (XC), WBAF and CDT. Shown are the locations that rely on the delay-only estimators (left) and the delay + linear dispersion estimator (right).

2.8 Discussion

We considered a class of signal estimation problems for which a positive valued signal is altered by a transformation of the independent variable. Such transformations are common in wave propagation, where the energy of a signal or field is modified by the medium through which it travels, but is ultimately conserved.

We proposed using the Wasserstein distance between the modified received signal and the model signal as a cost function for estimating the parameters that govern the transformation. The idea is to select those parameters that minimize the amount of work it takes to transform the model signal into the received signal. It was then shown that by using the cumulative distribution transform (CDT) the Wasserstein distance becomes a linear, convex function of the desired parameters and possesses a closed form (least squares) solution.

A series of numerical experiments were then conducted to assess the quality of the estimator. The CDT was found empirically to be approximately Gaussian distributed, hence the Wasserstein estimator is an approximate MLE for this class of problem. Indeed, the proposed estimator performs well in comparison to

other approaches in terms of estimator MSE. Moreover, because the estimator is linear in CDT space, the computational cost is orders of magnitude lower than competing methods.

The noise model used in the aforementioned experiments is assumed to be zero mean, i.i.d. Gaussian. While not considered explicitly, we expect other noise models to yield similar results. The reason is that the CDF is the summation of random variables, and with enough such variables the central limit theorem is expected to be applicable. Thus, even under other additive noise models, our CDT distribution will likely remain Gaussian (see Eqn. 2.32); this is a topic of ongoing work.

The numerical experiments also demonstrated the proposed noise reduction method, which works by subtracting the i.i.d. noise cumulative distribution function (CDF) from the total noisy signal CDF prior to effecting the transformation into the CDT domain.

Finally, the estimator was used to localize the source of acoustic emissions in a thin metal plate. It was shown that including linear dispersion in the signal model offered modest improvement in the ability to localize the source without incurring a computational penalty.

2.9 Conclusion

In this work, we proposed a parametric signal estimation approach by minimizing Wasserstein distance between measured and model signals. This approach, aided by the use of the cumulative distribution transform [61], was shown to produce generic closed form solution to the estimation problem. Several numerical experiments showed that the proposed approach not only performs well in comparison to existing methods but also is significantly more computationally efficient compared to the competing methods. In short, by using the CDT and the Wasserstein cost, one can easily and accurately estimate the parameters that govern the modification of signal energy during propagation.

Acknowledgements

This work is done in collaboration with Kyla M. Hallam, Dr. Jonathan M. Nichols, Dr. Meredith N. Hutchinson, Dr. Shiying Li, and Dr. Gustavo K. Rohde, to whom I express my deep gratitude for their contributions to this chapter.

Chapter 3

End-to-End Signal Classification in Signed Cumulative Distribution Transform Space

3.1 Introduction

Signal (time series) classification is considered a challenging problem in data science. It refers to the automatic prediction of the class label of an unknown time series event using the information extracted from the corresponding signal intensities. Time series data classification tasks can be found in many applications such as human activity recognition (HAR) [9], physiological signal assessment [10][11], communications [12], structural or machine health monitoring systems [13] [14], financial modeling [15], and others. In many applications (e.g., HAR, ECG, etc.) the time series events of interest can be modeled as instances of a certain (often unknown) template or prototype pattern observed under unknown time warps [87]. Here we propose a new end-to-end tool for classification of signals or time series events of this type using the signed cumulative distribution transform (SCDT), a new mathematical signal transform introduced in [88].

Existing signal classification approaches can be categorized into two broad, and at times overlapping, categories: 1) feature engineering-based classifiers and 2) end-to-end learning classifiers, such as convolutional neural networks (CNNs). Feature engineering-based methods [29] [30] [31] usually rely on the extraction of numerical features (e.g., time domain features, frequency domain features, wavelet features) from the raw signal data, and then the application of different multivariate regression-based classification methods includ-

ing linear discriminant analysis, support vector machines, random forests, and others. Deep learning-based signal classification methods [89] [90] [91], on the other hand, connect the raw input data to the output class label by utilizing a large number of hidden layers. These methods have widely been studied recently as they have shown high accuracy in certain classification tasks.

Feature engineering signal classification approaches primarily differ in the types of features chosen to characterize each signal. The bag-of-features framework [92] extracts interval features using fixed- and variable-length intervals, and trains a classifier on the extracted features. Ensemble-based approaches such as COTE [93], HIVE-COTE [94], time series forest (TSF) [95], and others combine different features and classifiers to achieve high classification accuracy. Most of these methods require crafting especially designed features (feature engineering) as well as some amount of data preprocessing.

Traditional end-to-end signal classification techniques include distance-based methods [32] [33] that work directly on raw time series with some similarity measures such as Euclidean distance or dynamic time warping (DTW) [34]. Particularly, a combination of 1-nearest neighbor (1NN) with DTW distance is known to be a very effective time series classification approach [35]. However, it is known to have high computational complexity [36]. Approaches based on deep neural networks, especially convolutional neural networks (CNN) [37], have been explored in recent years for end-to-end signal classification. Wang et al. [38], for example, provided three standard deep learning benchmark models for time series classification: deep multilayer perceptrons (MLP), fully convolutional networks (FCN), and residual networks (ResNet). The method known as multi-scale convolutional neural network (MCNN) [39] is another deep learning approach that takes advantage of CNNs for end-to-end classification of univariate time series. Karim et al. [40] proposed the LSTM-FCN, an improvement over FCN by augmenting the FCN module with a Long Short Term Recurrent Neural Network (LSTM RNN) sub-module. Though they can produce accurate results in many instances, these methods tend to require extensive amounts of training data, are computationally expensive, and often vulnerable to out-of-distribution examples. Furthermore, existing end-to-end classifiers often lack a proper data model and a clear formulation of the classification problem, making the classification models difficult to interpret. In particular, the lack of an underlying mathematical foundation often leads to uncertainty as to which exact situations or applications they will work, and when they will fail.

In recent years, some effort has been made to exploit transport transforms for signal classification [62] as an alternative to the techniques mentioned above. The cumulative distribution transform (CDT), based on the 1D Wasserstein embedding, was introduced in [96] as a means of classifying strictly positive signals following the linear optimal transport framework proposed in [97]. Aldroubi et al. [88] extended the CDT to general signed signals and proposed the signed cumulative distribution transform (SCDT), which is related to the generalized Wasserstein embedding [98] and can be viewed as an extension of the Wasserstein metric

to signed densities and measures. Both transforms have a number of properties that allow one to solve nonlinear classification problems using linear classifiers, such as Fisher discriminant analysis, support vector machines, or logistic regression, in signal transform space. The classification method proposed in [87] utilizes the nearest subspace method to classify signals in SCDT domain. Assuming the data corresponding to a particular class as compositions of a single template, this method formed a linear subspace for each class. The method we propose here is an extension of this approach.

In this work, we propose a new transport generative model to represent the signal data such that the signals from each class can be seen as observations of a set of unknown templates under some unknown deformations. We then formulate a supervised signal classification problem for the data that follows the transport generative model and employ a nearest local subspace search algorithm in SCDT domain to devise a solution. We demonstrate the advantages of the method over state-of-the-art deep learning methods in terms of classification accuracy on ten datasets with comprehensive experiments. The proposed method also provides superior performance in classifying signals with significantly low computational cost with respect to a distance-based end-to-end system (1NN-DTW). In addition, experiments highlight other interesting properties of the method compared to alternative end-to-end solutions, including data efficiency and robustness to out-of-distribution conditions. Note that the proposed classifier does not take the templates or the deformations present in a data class as inputs since these are usually unknown in real applications. It utilizes the training samples to search for the nearest local subspace in the SCDT space to classify an unknown signal. It should also be noted that the term "transport generative model" used in this work links the signal classes to the physical processes that generate the time series data from the corresponding classes. It differs from the "statistical generative model" used in machine learning, which is usually referred to as learning the underlying data distribution.

The remaining of this chapter is organized as follows: in section 3.2, we briefly review the definitions and properties of the CDT and the SCDT. In section 3.3, we state the classification problem and the proposed solution. Experimental setup, datasets, and results are described in section 3.4, with the discussion of the results in section 3.5. Finally, section 3.6 provides concluding remarks.

3.2 Preliminaries

3.2.1 Notation

Throughout this chapter, we work with L_1 signals s, i.e. $\int_{\Omega_s} |s(t)|dt < \infty$, where $\Omega_s \subseteq \mathbb{R}$ is the domain over which s is defined. We use $s_{j,m}^{(c)}$ to represent a signal generated from the m-th template of class c under

Table 3.1: Description of symbols

Symbols	Description
$s(t)$	Signal
$s_0(y)$	Reference signal to calculate the transform
$\widehat{s}(y)$	SCDT of signal $s(t)$
$g(t)$	Strictly increasing and differentiable function
$s \circ g$	$s(g(t))$: composition of $s(t)$ with $g(t)$
$\mathcal{T}$	Set of all possible increasing diffeomorphisms
$\mathbb{S}/\widehat{\mathbb{S}}$	Set of signals/SCDT of the signals

deformation g_j. We denote the m-th template from class c as $\varphi_m^{(c)}$. Some symbols used throughout this chapter are listed in Table 3.1.

3.2.2 The Cumulative Distribution Transform

The Cumulative Distribution Transform (CDT) was introduced in [96] for positive smooth normalized functions. It is an invertible nonlinear 1D signal transform from the space of smooth positive probability densities to the space of diffeomorphisms, which can be described as follows: let $s(t), t \in \Omega_s \subseteq \mathbb{R}$ and $s_0(y), y \in \Omega_{s_0} \subseteq \mathbb{R}$ define a given signal and a reference signal, respectively, such that $\int_{\Omega_s} s(u)du = \int_{\Omega_{s_0}} s_0(u)du = 1$ and $s_0(y), s(t) > 0$ in their respective domains. The CDT of the signal $s(t)$ is then defined to be the function $s^*(y)$ that solves,

$$\int_{\inf(\Omega_s)}^{s^*(y)} s(u)du = \int_{\inf(\Omega_{s_0})}^{y} s_0(u)du. \tag{3.1}$$

Now considering the cumulative distribution functions (CDFs) $S(t) = \int_{-\infty}^{t} s(u)du$ and $S_0(y) = \int_{-\infty}^{y} s_0(u)du$, an alternative expression for $s^*(y)$ is given by,

$$s^*(y) = S^{-1}(S_0(y)). \tag{3.2}$$

The CDT is therefore seen to inherit the domain of the reference signal. Moreover, if the uniform reference signal is used (i.e., $s_0(y) = 1$ in $\Omega_{s_0} = [0, 1]$), we can write $S_0(y) = y$ and $s^*(y) = S^{-1}(y)$. That is to say, the CDT is the inverse of the cumulative distribution function of the given signal $s(t)$. Note that the definition of the CDT described above is slightly different from the formulation used in [96]. For simplicity, here we use the CDT definition described in [99]. The CDT is invertible, and the inverse formula is defined in differential form as:

$$s(t) = \left(s^{*^{-1}}(t)\right)' s_0(s^{*^{-1}}(t)). \tag{3.3}$$

Although the CDT can be used in solving many classification [96] and estimation [99] problems, the framework described above is defined only for positive density functions. Aldroubi et al. [88] extended the CDT to general finite signed signals and named the new signal transformation technique as the signed cumulative distribution transform (SCDT).

3.2.3 The Signed Cumulative Distribution Transform

The signed cumulative distribution transform (SCDT) [88] is an extension of the CDT, which is defined for general finite signed signals with no requirements on the total mass. First, the transform is defined for the non-negative signal $s(t)$ with arbitrary mass as:

$$\widehat{s}(y) = \begin{cases} (s^*(y), \|s\|_{L_1}), & \text{if } s \neq 0 \\ (0,0), & \text{if } s = 0, \end{cases} \tag{3.4}$$

where $\|s\|_{L_1}$ is the L_1 norm of signal s and s^* is the CDT (defined in eqn. (3.2)) of the normalized signal $\frac{s}{\|s\|_{L_1}}$ with respect to a strictly positive reference signal s_0.

Now for a signed signal, the Jordan decomposition [100] is used to define the transform. The Jordan decomposition of a signed signal $s(t)$ is given by $s(t) = s^+(t) - s^-(t)$, where $s^+(t)$ and $s^-(t)$ are the absolute values of the positive and negative parts of the signal $s(t)$. The SCDT of $s(t)$ is then defined as:

$$\widehat{s}(y) = \left(\widehat{s^+}(y), \widehat{s^-}(y)\right), \tag{3.5}$$

where $\widehat{s^+}(y)$ and $\widehat{s^-}(y)$ are the transforms (defined in eqn. (3.4)) for the signals $s^+(t)$ and $s^-(t)$, respectively. Fig. 3.1 shows an example of the SCDT of a signal. Like the CDT, the SCDT is also an invertible operation, with the inverse being,

$$\begin{aligned} s(t) =& \|s^+\|_{L_1} \left((s^+)^{*^{-1}}(t)\right)' s_0((s^+)^{*^{-1}}(t)) \\ & - \|s^-\|_{L_1} \left((s^-)^{*^{-1}}(t)\right)' s_0((s^-)^{*^{-1}}(t)). \end{aligned} \tag{3.6}$$

Moreover, the SCDT has a number of properties that will help us simplify the signal classification problems.

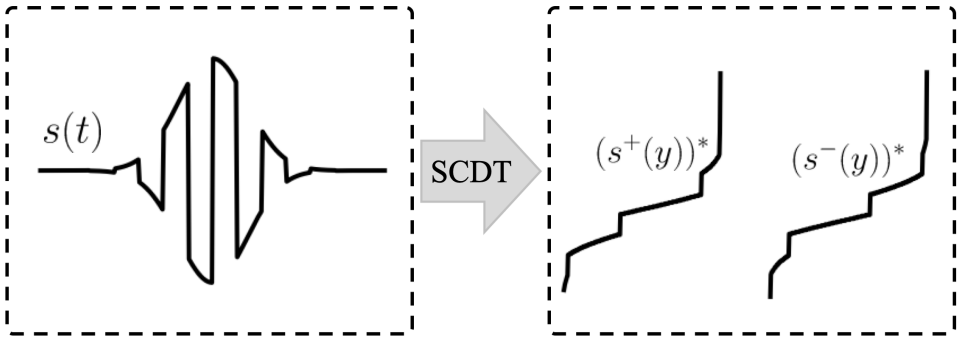


Figure 3.1: SCDT (without the constant terms) of an example signal.

Composition property

If the SCDT of a signed signal s is denoted as $\widehat{s}$, the SCDT of the signal $s_g = g's \circ g$ is given by:

$$\widehat{s}_g = \left(g^{-1} \circ (s^+)^*, \|s^+\|_{L_1}, g^{-1} \circ (s^-)^*, \|s^-\|_{L_1}\right), \tag{3.7}$$

where, $g(t)$ is an invertible and differentiable increasing function, $s \circ g = s(g(t))$, and $g'(t) = dg(t)/dt$ [88]. For example, in case of shift and linear dispersion (i.e., $g(t) = \omega t - \tau$) of a given signal $s(t)$, the SCDT of the signal $s_g(t) = \omega s(\omega t - \mu)$ can be derived from composition property as:

$$\widehat{s}_g = \left(\frac{(s^+)^* + \mu}{\omega}, \|s^+\|_{L_1}, \frac{(s^-)^* + \mu}{\omega}, \|s^-\|_{L_1}\right).$$

The composition property implies that variations along the independent variable caused by $g(t)$ will change only the dependent variable in the transform domain.

Convexity property

Let $\mathbb{S} = \{s_j | s_j = g_j' \varphi \circ g_j, \forall g_j \in \mathcal{G}\}$ be a set of signals, where φ is a given signal and $\mathcal{G} \subset \mathcal{T}$ denotes a set of 1D temporal deformations of a specific kind (e.g., translation, dilation, etc.). The convexity property of the SCDT [88] states that the set $\widehat{\mathbb{S}} = \{\widehat{s}_j : s_j \in \mathbb{S}\}$ is convex for every φ if and only if $\mathcal{G}^{-1} = \{g_j^{-1} : g_j \in \mathcal{G}\}$ is convex.

The set $\mathbb{S}$ defined above can be interpreted as a transport generative model for a signal class while φ being the template signal corresponding to that class.

Numerical Implementation of the SCDT

The SCDT described above is defined for the continuous-time signals. In this section, we describe the numerical method for approximating the SCDT given discrete signals. Let $\mathbf{s} = [s_1, s_2, ..., s_N]^T$ be a N-point discrete-time signal, where $s_n = s[n], \forall n = 1, 2, ..., N$ is the n-th sample of $\mathbf{s}$. The positive and negative parts of the signal after Jordan decomposition are given by $\mathbf{s}^+ = \left[s_1^+, ..., s_N^+\right]^T$ and $\mathbf{s}^- = \left[s_1^-, ..., s_N^-\right]^T$, respectively, where $s_n^+ = \frac{|s_n|+s_n}{2}$, $s_n^- = \frac{|s_n|-s_n}{2}$, and $|s_n|$ is the absolute value of s_n. Next, the CDT is applied numerically to the normalized signals $\frac{\mathbf{s}^+}{\|\mathbf{s}^+\|_{\ell_1}}$ and $\frac{\mathbf{s}^-}{\|\mathbf{s}^-\|_{\ell_1}}$, where $\|\mathbf{s}^+\|_{\ell_1}$ and $\|\mathbf{s}^-\|_{\ell_1}$ are the ℓ_1-norms of the positive and negative parts of the signal $\mathbf{s}$, respectively. As the CDT $(s^{\pm})^*(y)$ is the inverse of the CDF of the signal $s^{\pm}(t)$ for a particular choice of reference signal ($s_0(y) = 1$ for $y \in [0, 1]$), we need to approximate the cumulative function first. The numerical approximation of the cumulative function is given by,

$$S^{\pm}[n] = \sum_{i=1}^{n} \frac{s^{\pm}[n]}{\|\mathbf{s}^{\pm}\|_{\ell_1}}, \qquad n = 1, 2, ..., N,$$

where $\mathbf{S}^+$ and $\mathbf{S}^-$ are the cumulation of the normalized signals $\frac{\mathbf{s}^+}{\|\mathbf{s}^+\|_{\ell_1}}$ and $\frac{\mathbf{s}^-}{\|\mathbf{s}^-\|_{\ell_1}}$, respectively. The CDT is then calculated by taking the generalized inverse of the CDF,

$$(s^{\pm})^*[m] = min\left(\{t[n] : S^{\pm}[n] > y[m]\}\right),$$

where $t \in \Omega_s$, $y \in \Omega_{s_0}$, and $n, m = 1, 2, ..., N$. Here, $(\mathbf{s}^+)^*$ and $(\mathbf{s}^-)^*$ are the CDTs of the normalized discrete signals $\frac{\mathbf{s}^+}{\|\mathbf{s}^+\|_{\ell_1}}$ and $\frac{\mathbf{s}^-}{\|\mathbf{s}^-\|_{\ell_1}}$, respectively. The SCDT of the discrete signal $\mathbf{s}$ is then given by,

$$\widehat{\mathbf{s}} = \left((\mathbf{s}^+)^*, \|\mathbf{s}^+\|_{\ell_1}, (\mathbf{s}^-)^*, \|\mathbf{s}^-\|_{\ell_1}\right).$$

Note that the computational complexity of calculating the SCDT is $O(N \log N)$ for an N-point discrete-time signal.

3.3 Proposed Method

In this section we describe a transport generative model-based problem formulation for time series event classification, and then show how the composition and convexity properties of the SCDT help facilitate signal classification.

3.3.1 Transport generative model and problem statement

In [87], a transport generative model-based problem statement was proposed for signal classification problems where each signal class can be modeled as instances of a certain template observed under some unknown deformations, where the inverse of the deformation set can be modeled as a linear combination of increasing functions. (e.g., translation, scaling, nonrigid deformations described by the combination of polynomials or some other basis functions, etc.). Such models are common in data collected from a broad and important class of physical systems, e.g., radar, sonar, wave optics, acoustic signals propagating through elastic solids, etc. [99]. Signal classes of such type can be described with the following transport generative model:

Transport generative model (single template)

Let $\mathcal{G}^{(c)} \subset \mathcal{T}$ denote a set (usually infinite) of increasing 1D deformations of a specific kind, where $\mathcal{T}$ is a set of all possible increasing diffeomorphisms from $\mathbb{R}$ to $\mathbb{R}$. The 1D mass preserving transport generative model for class c is then defined to be the set:

$$\mathbb{S}^{(c)} = \{s_j^{(c)} | s_j^{(c)} = g_j' \varphi^{(c)} \circ g_j, g_j \in \mathcal{G}^{(c)}, g_j' > 0\}, \tag{3.8}$$

where $g_j' = dg_j/dt$, $s_j^{(c)}$ is the j-th signal from class c, and $\varphi^{(c)}$ is the template pattern corresponding to that class. However, in many applications it is difficult to find a signal class that can be represented using the single template-based transport generative model defined above. In this work, we use a multiple template-based transport generative model to represent such signal classes.

Transport generative model (multiple templates)

For a set of increasing 1D temporal deformations denoted as $\mathcal{G}_m^{(c)} \subset \mathcal{T}$, the 1D mass preserving transport generative model for class c is defined to be the set:

$$\begin{aligned}
\mathbb{S}^{(c)} &= \bigcup_{m=1}^{M_c} \mathbb{S}_{\varphi_m^{(c)}, \mathcal{G}_m^{(c)}}, \\
\mathbb{S}_{\varphi_m^{(c)}, \mathcal{G}_m^{(c)}} &= \left\{ s_{j,m}^{(c)} | s_{j,m}^{(c)} = g_j' \varphi_m^{(c)} \circ g_j, g_j' > 0, g_j \in \mathcal{G}_m^{(c)} \right\}, \\
\left(\mathcal{G}_m^{(c)} \right)^{-1} &= \left\{ \sum_{i=1}^{k} \alpha_i f_{i,m}^{(c)}, \alpha_i \geq 0 \right\},
\end{aligned} \tag{3.9}$$

where $\left\{ f_{1,m}^{(c)}, f_{2,m}^{(c)}, ..., f_{k,m}^{(c)} \right\}$ denotes a set of linearly independent and strictly increasing (within the domain of the signals) functions, and k is a positive integer. Here we assume that the signal classes are

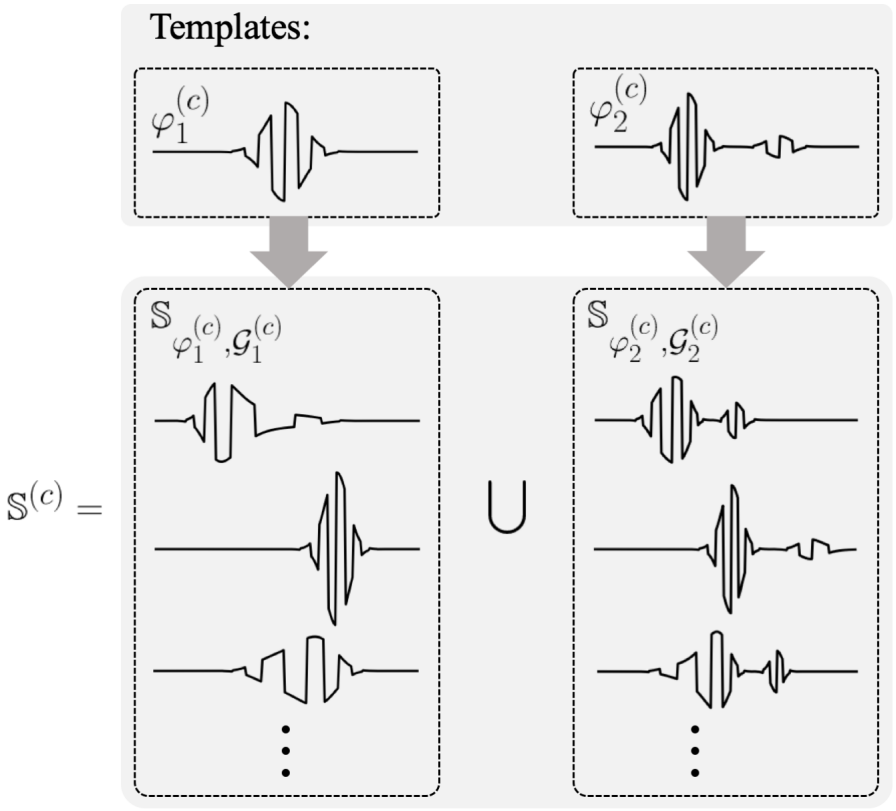


Figure 3.2: Transport generative model example for a signal class 'c'. The set of all signals from class c is denoted as $\mathbb{S}^{(c)}$, which is modeled as the union of subsets $\mathbb{S}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}}$ (for $m = 1, 2$) containing data generated from a template $\varphi_m^{(c)}$ under deformation $\mathcal{G}_m^{(c)}$.

non-overlapping, i.e., $\mathbb{S}^{(c)} \cap \mathbb{S}^{(p)} = \varnothing$ for $c \neq p$. Eq. (3.9) states that the transport generative model for class c is modeled as the union of M_c subsets, where each subset ($\mathbb{S}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}}$) corresponds to data generated from a particular template ($\varphi_m^{(c)}$) under some time deformations ($\mathcal{G}_m^{(c)}$). Here, M_c is the total number of templates used to represent class c, $\varphi_m^{(c)}$ is the m-th template signal from class c, and $s_{j,m}^{(c)}$ is the j-th signal generated from m-th template under deformation defined by g_j. Fig. 3.2 illustrates a few examples of such deformations. In eq. (3.9), by taking the finite sum, we hypothesize that the space where the inverse of the deformation set $\mathcal{G}_m^{(c)}$ lies must be of finite dimension (k-dimensional). Therefore, the set it generates, though infinite, is not a set of all possible diffeomorphisms, which would violate the assumption that the classes are non-overlapping. We assume that with reasonable k (determined through a validation process), the set of deformations $\mathcal{G}_m^{(c)}$ is flexible enough to model the deformations within a data cluster well. Moreover, the introduction of multiple templates ($\varphi_m^{(c)}$) in the proposed transport generative model adds extra flexibility. Note that the template $\varphi_m^{(c)}$ is different from the reference s_0 used to calculate the SCDT. Considering the transport generative model, the classification problem can be defined as follows:

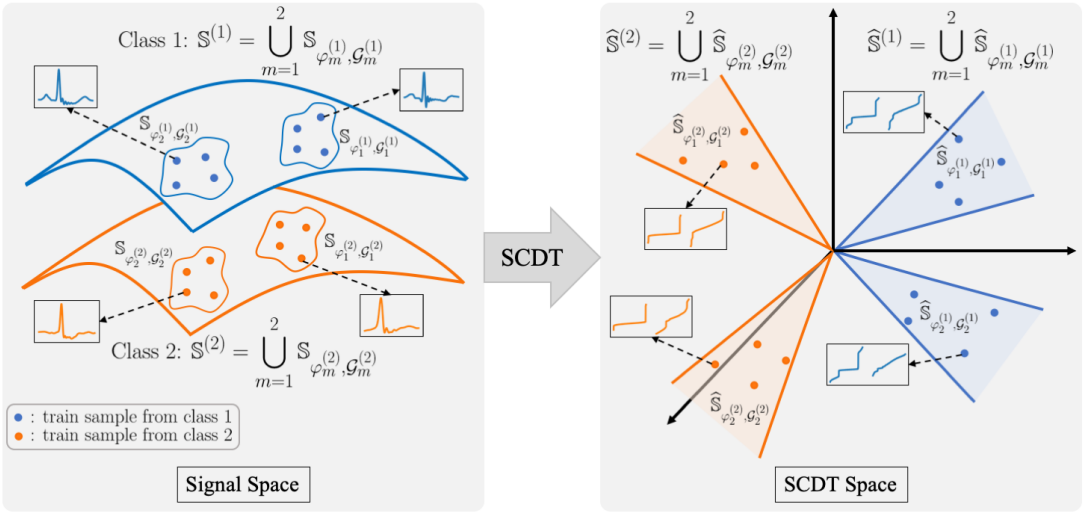


Figure 3.3: Geometric interpretation of data following the proposed transport generative model defined in equations (3.9) and (3.10). On the left panel, two classes ($\mathbb{S}^{(1)}$and $\mathbb{S}^{(2)}$) are depicted in signal space. The set $\mathbb{S}^{(c)}$ for class c is modeled as the union of two subsets: $\mathbb{S}_{\varphi_1^{(c)},\mathcal{G}_1^{(c)}}$ and $\mathbb{S}_{\varphi_2^{(c)},\mathcal{G}_2^{(c)}}$, $c = 1, 2$, where both the subsets are non-convex. The right panel shows the geometry of the signal classes in SCDT domain modeled as the union of convex (for convex $(\mathcal{G}_1^{(c)})^{-1}$ and $(\mathcal{G}_2^{(c)})^{-1}$ as defined in (3.9)) subsets: $\widehat{\mathbb{S}}_{\varphi_1^{(c)},\mathcal{G}_1^{(c)}}$ and $\widehat{\mathbb{S}}_{\varphi_2^{(c)},\mathcal{G}_2^{(c)}}$ for $c = 1, 2$.

Classification problem: *Let $\mathcal{G}_m^{(c)} \subset \mathcal{T}$ be a set of temporal deformations and $\mathbb{S}^{(c)}$ be defined as in eq. (3.9), for classes $c = 1, \cdots, N_c$. Given a set of training samples $\{s_1^{(c)}, s_2^{(c)}, ...\} \subset \mathbb{S}^{(c)}$ for class c, determine the class label of an unknown signal s.*

3.3.2 Proposed solution

We propose a solution to the classification problem defined above using the SCDT in combination with the nearest local subspace method. For a given test sample, the algorithm first searches for k closest training samples (from a particular class) to the test signal in SCDT domain based on a distance definition specified later. Next, a local subspace is spanned by these samples for each class. The unknown class label of the test sample is then estimated based on the shortest distance from the SCDT of the test signal to these local subspaces.

As stated in [87] and [101], the transport generative model described in eq. (3.9) generally yields non-convex signal classes, causing the above classification problem to be difficult to solve. As specified above in section 3.2.3 (see [88] for more details), under certain assumptions, the geometry of the signal class can be simplified. Hence, the proposed solution begins with applying the SCDT defined in eq. (3.5) on the input

signals. The transport generative model in the transform domain is then given by,

$$\begin{aligned}
\widehat{\mathbb{S}}^{(c)} &= \bigcup_{m=1}^{M_c} \widehat{\mathbb{S}}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}}, \\
\widehat{\mathbb{S}}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}} &= \left\{ \widehat{s}_{j,m}^{(c)} | \widehat{s}_{j,m}^{(c)} = g_j^{-1} \circ \widehat{\varphi}_m^{(c)}, g_j \in \mathcal{G}_m^{(c)} \right\}, \\
\left(\mathcal{G}_m^{(c)}\right)^{-1} &= \left\{ \sum_{i=1}^{k} \alpha_i f_{i,m}^{(c)}, \alpha_i \geq 0 \right\},
\end{aligned} \tag{3.10}$$

where $g_j^{-1} \circ \widehat{\varphi}_m^{(c)}$ is the SCDT of the signal $g_j' \varphi_m^{(c)} \circ g_j$. Here, the set $\left(\mathcal{G}_m^{(c)}\right)^{-1}$ is convex by definition. Therefore, using the convexity property highlighted earlier, it can be shown that $\widehat{\mathbb{S}}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}}$ given in eq. (3.10) forms a convex set. Moreover, since the SCDT is a one-to-one map, it follows that if $\mathbb{S}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}} \cap \mathbb{S}_{\varphi_w^{(p)},\mathcal{G}_w^{(p)}} = \varnothing$ for $c \neq p$, then $\widehat{\mathbb{S}}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}} \cap \widehat{\mathbb{S}}_{\varphi_w^{(p)},\mathcal{G}_w^{(p)}} = \varnothing$. Fig. 3.3 illustrates the geometry of signal classes that follow the proposed transport generative model defined in equations (3.9) and (3.10) corresponding to signal and SCDT domains, respectively.

To formulate the solution of the problem defined above, we adapt the subspace-based technique proposed in [87] for the multiple template-based transport generative model. First, Let us define a subspace generated by the convex set $\widehat{\mathbb{S}}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}}$ as:

$$\widehat{\mathbb{V}}_m^{(c)} = \text{span}\left(\widehat{\mathbb{S}}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}}\right). \tag{3.11}$$

Since $\widehat{\mathbb{S}}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}} \cap \widehat{\mathbb{S}}_{\varphi_w^{(p)},\mathcal{G}_w^{(p)}} = \varnothing$ (when $c \neq p$) and $\left(\mathcal{G}_m^{(c)}\right)^{-1}$ is convex by definition, it is reasonable to assume that $\widehat{\mathbb{S}}^{(c)} \cap \widehat{\mathbb{V}}_w^{(p)} = \varnothing$ for any $w = 1, ..., M_p$ (see Appendix B.1). Now, if a test sample s is generated according to the transport generative model defined in eq. (3.9), then there exist a certain class 'c' and a certain template $\varphi_m^{(c)}$ for which $d^2(\widehat{s}, \widehat{\mathbb{V}}_m^{(c)}) = 0$. Here, $\widehat{s}$ is the SCDT of the test sample s, and $d^2(\widehat{s}, \widehat{\mathbb{V}}_m^{(c)})$ is the Euclidean distance between $\widehat{s}$ and the nearest point in subspace $\widehat{\mathbb{V}}_m^{(c)}$. It also follows, $d^2(\widehat{s}, \widehat{\mathbb{V}}_w^{(p)}) > 0$ when $p \neq c$. Therefore, under the assumption that the test sample s is generated according to the transport generative model for one of the (unknown) classes, the unknown class label can be uniquely predicted by solving,

$$\underset{c}{\text{argmin}} \min_{m} \; d^2\left(\widehat{s}, \widehat{\mathbb{V}}_m^{(c)}\right), \tag{3.12}$$

where $\widehat{\mathbb{V}}_m^{(c)}$ is given by eq. (3.11). The proposed algorithm to solve the classification problem is outlined below.

3.3.3 Algorithm: nearest local subspace in SCDT domain

In the signal classification tasks considered below, as usually the case, the template $\varphi_m^{(c)}$ and the deformation set $\mathcal{G}_m^{(c)}$ are usually unknown. Therefore the subset $\widehat{\mathbb{S}}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}}$ is also unknown; hence, we can not readily estimate $\widehat{\mathbb{V}}_m^{(c)}$ using eq. (3.11). Here we devise an algorithm to approximate $\widehat{\mathbb{V}}_m^{(c)}$ using training samples from class c. Let us assume that the test sample s is generated according to the transport generative model $\mathbb{S}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}}$. From eq. (3.10), $\mathbb{S}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}}$ can be defined in the SCDT domain as:

$$\widehat{\mathbb{S}}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}} = \left\{ \left(\sum_{i=1}^{k} \alpha_i f_{i,m}^{(c)} \right) \circ \widehat{\varphi}_m^{(c)}, \alpha_i \geq 0 \right\},$$

where $\left\{ f_{1,m}^{(c)}, f_{2,m}^{(c)}, ..., f_{k,m}^{(c)} \right\}$ is a set of k linearly independent increasing functions. From this definition, it is evident that $\widehat{\mathbb{V}}_m^{(c)}$ (as defined in eq. (3.11)) is a k dimensional space. Therefore, if we were to estimate this span, we would need at least k linearly independent elements from the set $\widehat{\mathbb{S}}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}}$ to model it. Since we do not have the knowledge of $\widehat{\mathbb{S}}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}}$, we employ a nearest local subspace (NLS) search algorithm in SCDT domain to approximate $\widehat{\mathbb{V}}_m^{(c)}$. Let us denote the estimated local subspace for class c as $\widetilde{\widehat{\mathbb{V}}}_m^{(c)}$. The solution to the problem defined in eq.(3.12) is then estimated by solving,

$$\underset{c}{\operatorname{argmin}} \; d^2 \left(\widehat{s}, \widetilde{\widehat{\mathbb{V}}}_m^{(c)} \right). \tag{3.13}$$

Consider a set of training samples $\left\{ s_1^{(c)}, ..., s_j^{(c)}, ..., s_{L_c}^{(c)} \right\} \subset \mathbb{S}^{(c)}$ for class c, where L_c is the total number of training samples given for class c and $s_j^{(c)}$ is the j-th sample. The unknown class of a test sample s is estimated in two steps:

Step 1: We search for the k closest training samples to $\widehat{s}$ from class c based on the distance between $\widehat{s}$ and the span of each training sample. First, we sort the elements from the set $\{\widehat{s}_1^{(c)}, ..., \widehat{s}_{L_c}^{(c)}\}$ into $\{\widehat{z}_1^{(c)}, ..., \widehat{z}_{L_c}^{(c)}\}$ such that

$$d^2(\widehat{s}, \widehat{\mathbb{V}}_{z_1}^{(c)}) \leq \cdots \leq d^2(\widehat{s}, \widehat{\mathbb{V}}_{z_l}^{(c)}) \leq \cdots, \tag{3.14}$$

where $\widehat{\mathbb{V}}_{z_l}^{(c)} = \text{span}\left(\{\widehat{z}_l^{(c)}\} \right)$. Then we pick the first k elements from the sorted set to form $\{\widehat{z}_1^{(c)}, \cdots, \widehat{z}_k^{(c)}\}$ for $k \leq L_c$, which gives the set of k closest training samples to $\widehat{s}$ from class c in the above sense (Fig. 3.4b-3.4c). We repeat this step for all other classes (Fig. 3.4d).

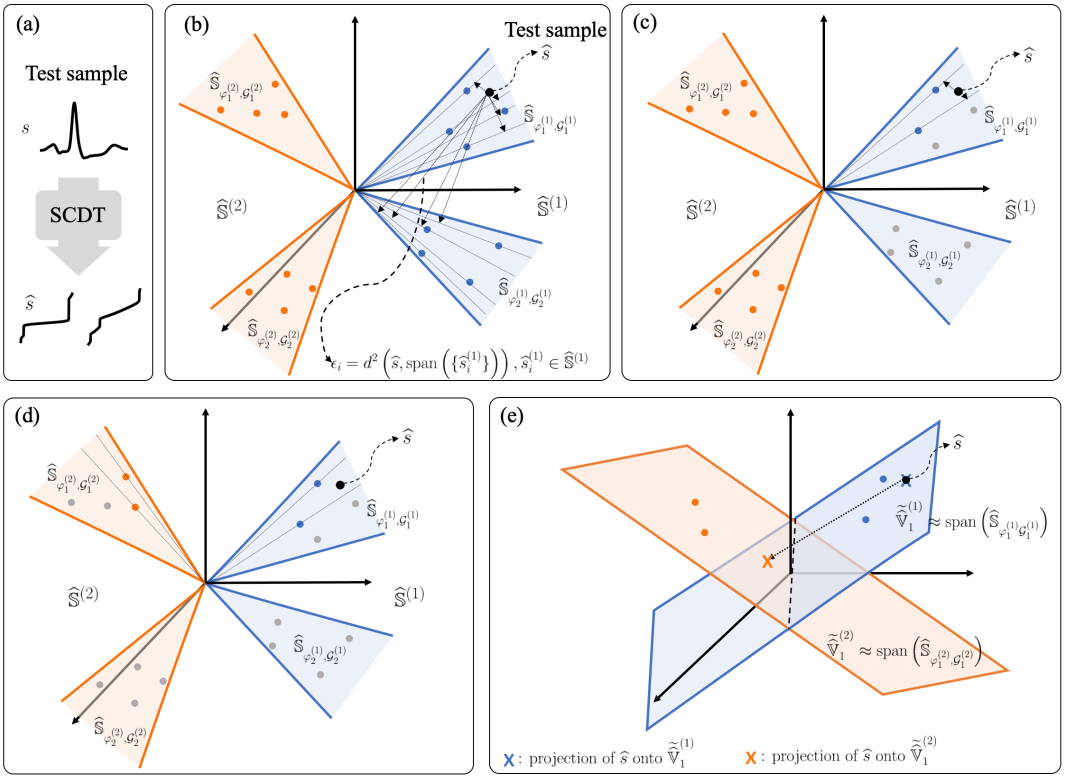


Figure 3.4: Outline of proposed algorithm: (a) apply SCDT on the test signal s to obtain $\widehat{s}$, (b) measure distance between $\widehat{s}$ and subspace corresponding to each training sample from a particular class (class-1 in this example), (c) find k closest training samples ($k = 2$ in this example) to $\widehat{s}$ from class-1, (d) repeat previous steps for other classes (class-2 in this example), (e) build local subspace for each class using the k samples found in previous steps and search for the nearest local subspace to predict the class of s.

Step 2: Compute $\widetilde{\widehat{\mathbb{V}}}_m^{(c)}$ by:

$$\widetilde{\widehat{\mathbb{V}}}_m^{(c)} = \text{span}\left(\{\widehat{z}_1^{(c)}, \cdots, \widehat{z}_k^{(c)}\}\right), \tag{3.15}$$

which approximates the nearest local subspace from class c with respect to $\widehat{s}$. We then predict the unknown class of the test sample s by solving eq. (3.13). Fig. 3.4e illustrates the second step. Note that similar local subspace classification techniques can be found in the literature [102] [103]. Here, we employ the nearest local subspace search technique in the SCDT domain to exploit the properties of the SCDT that simplify the classification problem described above. Next, we show that the performance of the classifier can further be improved by utilizing the composition property of the SCDT through analytical enrichment of the subspace method.

Subspace enrichment

The algorithm outlined above searches for the k closest training samples to a given test sample in SCDT domain, and then forms a local subspace using these k samples. Inspired by [101, 104], here we adapt the technique to mathematically enrich the ensuing subspace with certain prescribed deformations. Here we enrich $\widetilde{\widehat{\mathbb{V}}}_m^{(c)}$ in such a way that it will automatically include the samples undergoing certain time deformations. The spanning sets corresponding to certain specific deformations are derived below:

- Translation: In case of translation, $g(t)$ is given by $t-\mu$, where $\mu \in \mathbb{R}$ is the translation parameter. Using the composition property of the SCDT, the transform of the translated signal $s_g(t) = s(t-\mu)$ is given by $\widehat{s}_g = ((s^+)^* + \mu, \|s^+\|_{L_1}, (s^-)^* + \mu, \|s^-\|_{L_1})$. It implies that the translation applied to the signal along the independent axis results in translation along the dependent axis in SCDT domain. Hence, the spanning set for translation is defined as $\mathbb{U}_T = \{u(t)\}$, where $u(t) = 1$.

- Dilation: A time-dilated (scaled) version of a signal $s(t)$ is defined as: $s_g(t) = \alpha s(\alpha t), \alpha \in \mathbb{R}^+$. The transform of the signal $s_g(t)$ is given by: $\widehat{s}_g = \left(\frac{(s^+)^*}{\alpha}, \|s^+\|_{L_1}, \frac{(s^-)^*}{\alpha}, \|s^-\|_{L_1}, \right)$. An additional spanning set is not required for dilation, as it is inherent in the modeling of the subspace.

- Time-warpings other than translation and dilation are also observed in certain classification problems. To include those deformations in the SCDT domain, we approximate the increasing function $g^{-1} \circ \widehat{z}_l^{(c)}$ as:

$$g^{-1} \circ \widehat{z}_l^{(c)} = \begin{cases} \sum_{n=-N}^{N} c_n \zeta_n(\widehat{z}_l^{(c)}) & \text{for } n \neq 0 \\ c_0 \widehat{z}_l^{(c)} & \text{for } n = 0 \end{cases} \tag{3.16}$$

  where, $c_n > 0, \quad \sum_{n=-N}^{N} c_n = 1$, and

$$\zeta_n(\widehat{z}_l^{(c)}) = \left[\widehat{z}_l^{(c)} - \frac{\sin n\pi \widehat{z}_l^{(c)}}{|n|\pi}\right].$$

  For the set $\{\widehat{z}_1^{(c)}, \cdots, \widehat{z}_k^{(c)}\}$ from eq. (3.15), the spanning set is given by the set $\mathbb{U}_H = \{\zeta_n(\widehat{z}_1^{(c)}), ..., \zeta_n(\widehat{z}_k^{(c)})\}$, for $n = -N, ..., -1, 1, ..., N$. These non-linear increasing deformations can be viewed as small perturbations of the identity function and can be used as the approximation of some time-warpings present in the data [105].

In light of the discussion above, the subspace $\widetilde{\widehat{\mathbb{V}}}_m^{(c)}$ in eq. (3.15) can be enriched as follows:

$$\widetilde{\widehat{\mathbb{V}}}_m^{(c)} = \text{span}\left(\{\widehat{z}_1^{(c)}, \cdots, \widehat{z}_k^{(c)}\} \cup \mathbb{U}_T \cup \mathbb{U}_H\right). \tag{3.17}$$

In a nutshell, The subspace enrichment process consists of adding pre-determined dimensions to the span of the training data. Each of these pre-determined dimensions is simply a vector that becomes another training vector. The PCA algorithm is then employed to obtain an orthonormal basis spanning this set. The enrichment vectors added tend to be few ($2k+1$ per class on average, where k corresponds to the k nearest samples) and thus the computation of the PCA step is not altered much. Note that the subspace $\widehat{\mathbb{V}}_{z_l}^{(c)}$ in (3.14) can also be enriched in similar manner, i.e. $\widehat{\mathbb{V}}_{z_l}^{(c)} = \text{span}\left(\{\widehat{z}_l^{(c)}\} \cup \mathbb{U}_T \cup \mathbb{U}_H\right)$, where $\mathbb{U}_H = \{\zeta_n(\widehat{z}_l^{(c)})\}$, for $n = -N, ..., -1, 1, ..., N$.

Training phase

In the training phase of the algorithm, the subspace corresponding to each of training sample is calculated. The first step is to compute SCDTs for all training samples from class c. Then, we take a training sample $\widehat{s}_l^{(c)}$ and orthogonalize $\{\widehat{s}_l^{(c)}\} \cup \mathbb{U}_T \cup \{\zeta_n(\widehat{s}_l^{(c)})\}$ (where $n = -N, ..., -1, 1, ..., N$) to obtain the basis vectors that span the enriched subspace corresponding to that sample. Let $B_l^{(c)} = \left[b_{l,1}^{(c)}, b_{l,2}^{(c)}, ...\right]$ be a matrix that contains the basis vectors in its column. We repeat these calculations for all the training samples to form $B_l^{(c)}$ for $l = 1, 2, ..., L_c$ and $c = 1, 2, ...,$ etc.

Testing phase

The testing algorithm begins with taking SCDT of the test sample s to obtain $\widehat{s}$ followed by the nearest local subspace search in SCDT domain. In the first step of the algorithm, we estimate the distance of the subspace corresponding to each of the training samples from $\widehat{s}$ by:

$$\epsilon_l = \|\widehat{s} - B_l^{(c)} B_l^{(c)^T} \widehat{s}\|^2, \qquad l = 1, 2, ..., L_c,$$

where $\|.\|$ denotes L_2 norm. As all the signals (and the corresponding SCDTs) are discrete, the L_2 norm (norm of functions on the real line) is replaced by the ℓ_2 norm (norm of sequences) in the calculations. Note that $B_l^{(c)} B_l^{(c)^T}$ is the orthogonal projection matrix onto the space generated by the span of the columns of $B_l^{(c)}$ (computed in the training phase). We then find $\{\widehat{z}_1^{(c)}, \cdots, \widehat{z}_k^{(c)}\}$, a set of k closest training samples to the test sample $\widehat{s}$ from class c, based on the distances $\epsilon_1, ..., \epsilon_{L_c}$. In the next step, we orthogonalize $\{\widehat{z}_1^{(c)}, \cdots, \widehat{z}_k^{(c)}\} \cup \mathbb{U}_T \cup \mathbb{U}_H$ to obtain the basis vectors $\{b_1^{(c)}, b_2^{(c)}, ...\}$ spanning the local subspace from class

c with respect to $\widehat{s}$. Let $B^{(c)} = \left[b_1^{(c)}, b_2^{(c)}, ...\right]$ for $c = 1, 2, ...$, etc. The unknown class of s is then estimated by:

$$\arg\min_{c} \ \|\widehat{s} - B^{(c)} {B^{(c)}}^T \widehat{s}\|^2. \tag{3.18}$$

The most computationally expensive step of the testing phase is the orthogonalization of the set $\{\widehat{z}_1^{(c)}, \cdots, \widehat{z}_k^{(c)}\} \cup \mathbb{U}_T \cup \mathbb{U}_H$. If the set has m-elements (signals) of n-length (discrete samples), the complexity of the orthogonalization process is given by $O(m^2 n)$ [106]. Therefore, the computational complexity of the proposed classifier is given by $O(cm^2 n)$, where c is the number of classes.

Note that the proposed algorithm requires two parameters k and N (see eq. (3.16)) to be tuned prior to the testing phase. We use a validation set split from the training set to estimate the optimum values for these parameters. We then follow the steps of the proposed algorithm outlined above with the validation set for varying k and N. Parameter values corresponding to the best validation accuracy are chosen to be used in the testing phase. This step is done during the training phase.

3.3.4 Proof-of-concept simulation

To demonstrate the efficacy of the proposed algorithm in solving the classification problem stated in section 3.3.1 we performed a simulated experiment. We took six prototype signals shown in Fig. 3.5 as the templates ($\varphi_m^{(c)}$) corresponding to three different classes ($c = 1, 2, 3$), i.e., each class has two templates ($m = 1, 2$). We then generated a synthetic dataset by applying specific time deformations on the prototype signals as follows:

$$\begin{aligned} s_j^{(c)}(t) &= g_j'(t)\varphi_m^{(c)}(g_j(t)), \\ g_j(t) &= \omega\xi(t) + \tau, \quad g_j'(t) > 0, \\ \xi'(t) &= \frac{\partial}{\partial t}\xi(t) = \sum_{n=1}^{N} \alpha_n \frac{1}{\sqrt{2\pi} w_n} e^{-\frac{1}{2}\left(\frac{t-\mu_n}{w_n}\right)^2} \end{aligned} \tag{3.19}$$

where $\alpha_n > 0, \quad \sum_n \alpha_n = 1$ and the parameter values used to calculate $g_j(t)$ are randomly chosen from fixed intervals. The dataset was equally split into training and testing sets. The proposed classification method was then trained with a varying number of training samples per class randomly chosen from the training set and evaluated on the testing set. Note that the samples were chosen in such a way that the training set contains an equal number of samples generated from each template. Fig. 3.5 shows the test accuracy plots with respect to the number of training samples per class. It demonstrates that the proposed method achieves the perfect classification accuracy with few training samples (obtained 99.97% test accuracy with only 16 training samples per class), while some alternative end-to-end systems (discussed in next section)

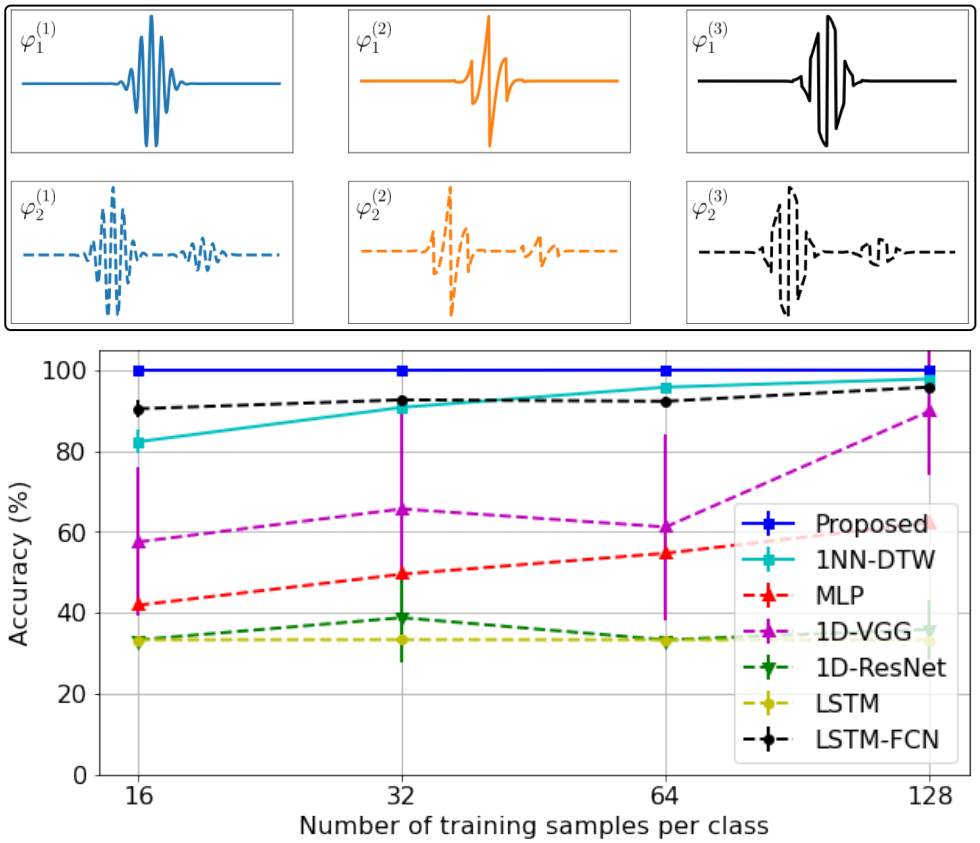


Figure 3.5: A simulated experiment to demonstrate the efficacy of the proposed method in classifying signal classes that follow the transport generative model defined in eq. (3.9).

fail to achieve such performance even with higher number of training data. That means, if the signal classes follow the transport generative model defined in eq. (3.9), the proposed method solves the signal classification problem stated in section 3.3.1.

3.4 Experiments and Results

3.4.1 Experimental setup

The goal is to evaluate the performance of the proposed generic end-to-end classifier with respect to selected state-of-the-art end-to-end time series classification techniques. The classification performance of the different methods were studied in terms of test accuracy, data efficiency, computational efficiency, and robustness to the out-of-distribution samples. We conducted experiments on several time series data and compared the results against several methods: Multilayer Perceptrons (MLP) [107], 1D Visual Geometry Group (VGG) [107], 1D Residual Network (ResNet) [108] [107], Long Short Term Memory (LSTM) [109] [107], Long Short Term Memory Fully Convolutional Network (LSTM-FCN) [40] [107], and 1-nearest neighbor DTW

(1NN-DTW) [35]. For the neural network methods, we used the implementations outlined in [107] with data augmentation. Jittering, scaling, and time-warping were used as the data augmentation methods to generate augmented time series samples, which lead the training set used to train the neural networks to be six times the original training set after data augmentation. During the training process, 10% of the training samples were used as the validation set, and the test performance was reported based on the model that had the best validation performance. In case of the 1NN-DTW method, the validation set was used to tune the parameter corresponding to the warping window size. To show the efficacy of the proposed multiple template-based transport generative model (defined in eq. (3.9)) over the single template-based model (eq. (3.8)), we also compared the results against the SCDT-NS classifier [87] proposed earlier.

We followed the training and testing procedures outlined in the previous section for the proposed method. The orthogonalization operations were performed using singular value decomposition (SVD). The left singular vectors obtained by the SVDs were used to construct the matrices $B_l^{(c)}$ and $B^{(c)}$. Following [101], the number of the basis vectors was chosen in such a way that the sum of variances explained by the selected basis vectors captures at least 99% of the total variance explained by all the samples in the sorted set $\{\widehat{z}_1^{(c)}, \cdots, \widehat{z}_k^{(c)}\}$. The SCDTs were computed with respect to a 1D uniform probability density function. Note that the properties of the CDT/SCDT are independent of the choice of the reference signal s_0. Therefore, in principle, the performance of the proposed classifier will not change if a positive reference signal other than the uniform density function is used to calculate the SCDT of the signals.

3.4.2 Datasets

To evaluate the comparative performance of the proposed method with respect to other end-to-end classifiers, we selected multiple datasets with signal classes representing well-defined time series events. For example, the accelerometer data plotted in Fig. 3.6(a) represent particular hand gestures. Similarly, signals shown in Fig. 3.6(i) represent either normal or abnormal heartbeats. With a focus on this condition, we identified 10 different time series datasets, 8 of which were downloaded from the UCR time series classification archive [110]. Some example signals from these datasets are shown in Fig. 3.6 (more examples have been provided in Appendix B.2). Details about the datasets are given below:

- *GesturePebbleZ2* [111]: Accelerometer data collected using Pebble smart watch from 4 different persons performing 6 hand gestures. (classes: 6, train samples: $22 \sim 25$ per class, test samples: $25 \sim 32$ per class).

- *InsectEPGRegularTrain* [112]: Contains electrical penetration graph (EPG) data which capture voltage changes of the electrical circuit that connects insects and their food source. (classes: 2, train samples: $22 \sim 30$ per class, test samples: $89 \sim 118$ per class).
- *PLAID*: Plug Load Appliance Identification Dataset [113]. The data are intended for load identification research using transient voltage/current measurements from 11 different appliance types. (classes: 11, train samples: $13 \sim 88$ per class, test samples: $13 \sim 87$ per class).
- *UWaveGestureLibraryAll* [114]: A set of eight simple gestures generated from accelerometers using Wii remote. (classes: 8, train samples: $100 \sim 127$ per class, test samples: $433 \sim 460$ per class).
- *Wafer* [115]: A collection of inline process control measurements recorded from various sensors during the processing of silicon wafers. The two classes are normal and abnormal, with a significant class imbalance. Hence, a subset of the original dataset is used to ensure the class balance. (classes: 2, train samples: $97 \sim 100$ per class, test samples: $665 \sim 700$ per class).
- *StarLightCurves* [116]: Collection of time series signals representing the brightness of celestial objects as a function of time. (classes: 3, train samples: 150 per class, test samples: 500 per class).
- *TwoPatterns* [117]: A simulated dataset (classes: 4, train samples: $237 \sim 271$ per class, test samples: $959 \sim 1035$ per class).
- *ECG5000* [110]: A subset of BIDMC Congestive Heart Failure Database (CHFDB) downloaded from PhysioNet. With a purpose of evaluating the methods trained with a large set, we interchanged the train and test sets of the original dataset. (classes: 2, train samples: $1873 \sim 2627$ per class, test samples: $208 \sim 292$ per class).
- *ECG (MLII)* [118]: A subset of a publicly available dataset reported in [119]. The ECG signals were collected from the PhysioNet MIT-BIH Arrhythmia database [120]. A method described in [121] was used to segment the heartbeats from the ECG fragments. (classes: 3, train samples: 200 per class, test samples: 200 per class).
- *Connectionist Bench (Sonar, Mines vs. Rocks)* [122] [123]: This dataset contains energy patterns (with respect to frequency bands) of the signals obtained by bouncing sonar signals off a metal cylinder and some rocks at various angles and under various conditions. (classes: 2, train samples: $49 \sim 55$ per class, test samples: $48 \sim 56$).

In addition to the datasets listed above, we have used another dataset where time series events are not well-defined as an example where the proposed method is not expected to work well. We have used the gearbox

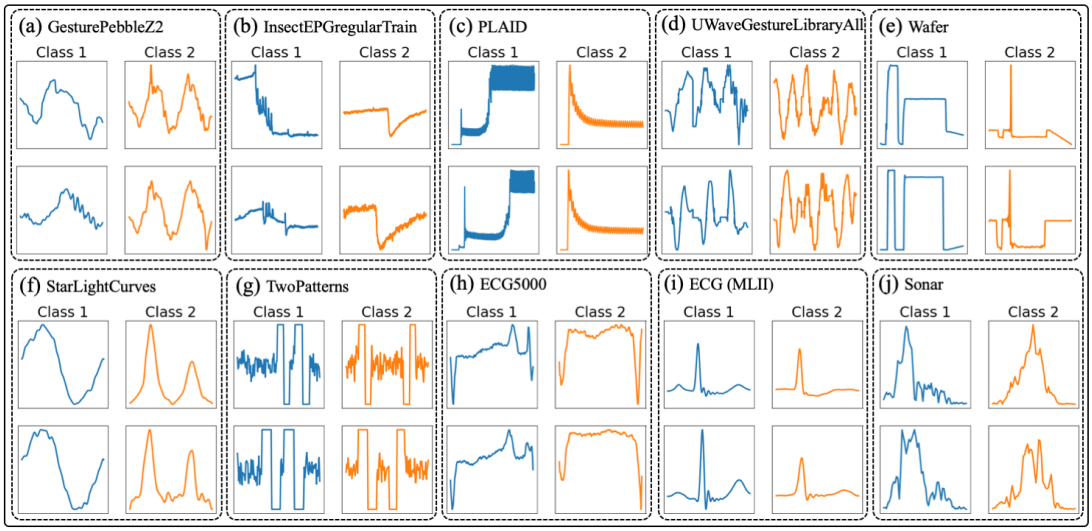


Figure 3.6: Some example signals from the datasets used to evaluate the proposed method.

fault diagnosis data [124], a collection of vibration signals recorded by using SpectraQuest's Gearbox Fault Diagnostics Simulator. The dataset was recorded in two different scenarios: 1) healthy and 2) broken tooth conditions.

3.4.3 Test accuracy

Dataset	MLP	1D-VGG	1D-ResNet	LSTM	LSTM-FCN	SCDT-NS	1NN-DTW	**Proposed**
GesturePebbleZ2	88.54	56.20	53.98	45.82	92.47	81.64	**94.94**	94.30
InsectEPGRegularTrain	80.39	87.63	49.66	60.58	50.63	73.43	82.61	**90.82**
PLAID	17.69	19.39	42.42	29.61	40.04	58.29	**73.56**	70.95
UWaveGestureLibraryAll	92.79	90.23	45.42	39.80	94.16	90.4	94.22	**94.95**
Wafer	96.77	95.32	90.53	51.62	94.02	92.75	**98.16**	95.60
StarLightCurves	84.05	**89.41**	68.98	77.23	81.31	77.4	83.80	84.2
TwoPatterns	56.09	99.67	51.40	41.54	79.6	95.15	99.90	**99.92**
ECG5000	98.40	99.04	98.68	99.16	98.76	93.4	**99.20**	97.6
ECG (MLII)	33.33	33.33	53.50	56.87	64.10	56.67	45.67	**68.83**
Sonar	75.38	78.65	52.31	53.85	73.65	61.54	76.92	**78.85**
Win	0	1	0	0	0	0	4	**5**
AVG arithmetic ranking	4.7	4.1	6.6	6.3	4.7	5.1	2.4	**2.1**
AVG geometric ranking	4.33	3.45	6.44	5.8	4.47	4.92	1.97	**1.67**
MPCE	0.073	0.061	0.112	0.127	0.073	0.071	0.049	**0.038**

Table 3.2: Test accuracy (%), rank-based statistics, and MPCE calculated for the classifiers across different datasets.

To demonstrate the efficacy of the proposed method as a generic classifier, we applied it to the datasets listed above and compared the test accuracies against the aforementioned end-to-end classification methods. Table 3.2 shows the results and a comprehensive comparison with five neural network-based classifiers, 1NN-DTW, and SCDT-NS. The results reported in the table show that for 5 out of 10 datasets, the proposed method outperformed the existing state-of-the-art techniques and provided competitive test accuracy for the remaining

five datasets. The average arithmetic and geometric rankings also demonstrate the efficacy of the proposed method as a generic end-to-end technique to classify segmentable time series events.

Besides test accuracy and rank-based statistics, we also calculated mean per class error (MPCE) for each method. This metric was proposed in [38] to evaluate the performance of a generic classifier on multiple datasets. MPCE is defined as the arithmetic mean of the per class error (PCE) which is calculated for i-th model on j-th dataset as: $\text{PCE}_{i,j} = \frac{e_{i,j}}{c_j}$. Here $e_{i,j}$ is the error rate of i-th model on j-th dataset, and c_j is the total number of classes present in j-th dataset. MPCE for the corresponding model is given by,

$$\text{MPCE}_i = \frac{1}{J}\sum_{j=1}^{J}\text{PCE}_{i,j},$$

where J is the total number of datasets used in the experiment. The MPCE values reported in Table 3.2 indicate that the proposed method generates the least expected error rate per class across all the datasets in comparison to other classifiers.

3.4.4 Data efficiency

To show the data efficiency of the proposed method, we set up an experiment where we trained the models with a varying number of training samples per class. For a training split of a particular size, its samples were randomly drawn from the original training set, and the experiments for this particular size were repeated 10 times. Fig. 3.7 shows average accuracies with respect to the number of training samples per class for two datasets: *UWaveGestureLibraryAll* and *TwoPatterns*. The standard deviation for each split is also shown using the error bar. The plots illustrate that the proposed method achieves higher accuracy than the deep learning methods with fewer training samples. Similar results can be seen in other datasets as well (see Appendix B.3).

3.4.5 Computational efficiency

The proposed classification technique is not only effective in classifying time series events but also computationally very efficient. Fig. 3.8 shows the computational complexity plots of the proposed method and 1NN-DTW as a function of number of training samples per class for *TwoPatterns* dataset. It demonstrates that the proposed classifier requires less CPU operations in comparison to 1NN-DTW. Fig. 3.9 (upper panel) plots the average time (in seconds) required during the training phase as a function of accuracy for each method. It illustrates that all the alternative end-to-end solutions require greater number of computations to train the models with respect to the proposed solution to achieve same level of accuracy. Despite the fact

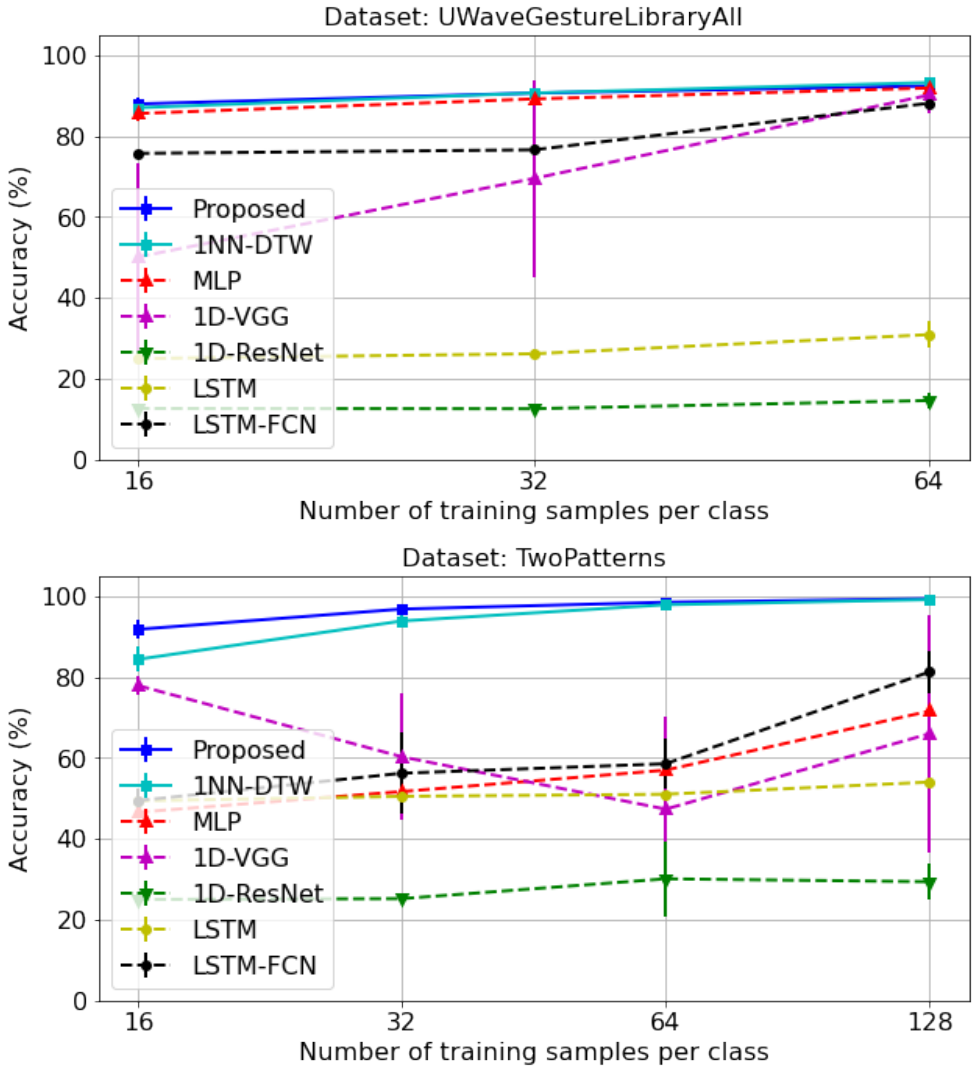


Figure 3.7: Accuracy as a function of number of training samples per class for different classification methods.

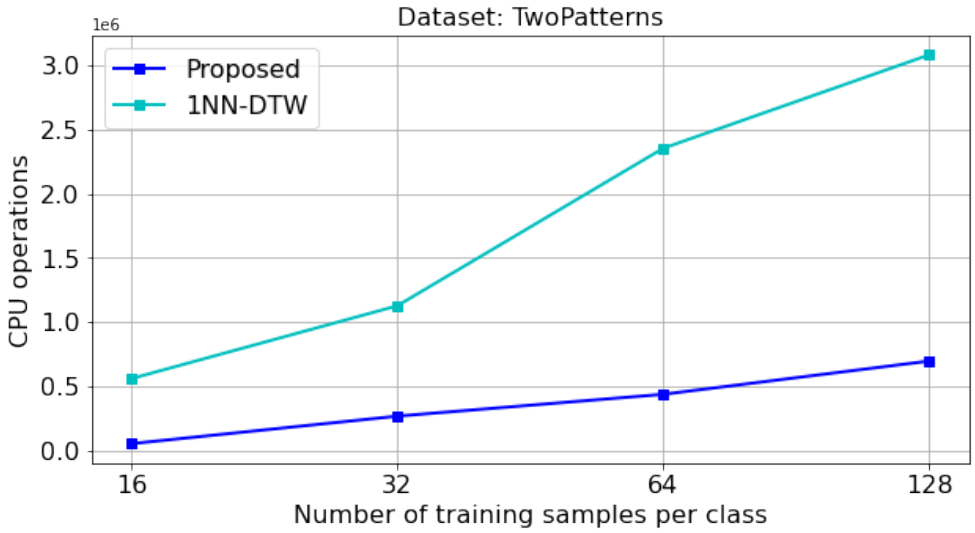


Figure 3.8: Computational complexity plots of 1NN-DTW and the proposed method as a function of number of training samples. Complexities of the proposed method and 1NN-DTW are given by $O(cm^2n)$ (discussed in previous section) and $O(pnw)$. Here p is the total number of training samples, n is the signal length, w is the window length to calculate DTW.

that our method searches for k-closest training samples for a given test signal, it still provides competitive performance as shown in Fig. 3.9 (lower panel) in terms of test time in comparison to deep learning-based methods. The plots of the average time (in seconds) taken by the classification methods to test a signal from *TwoPatterns* dataset show that the test time of the proposed method is close to the neural network-based classifiers, while 1NN-DTW is highly expensive in terms of computation during testing phase. Note that the SCDT is currently implemented with *for* loops in python, which is less than ideal in terms of execution time. A compiled language (e.g. C, C++) would execute *for* loops much faster.

3.4.6 Robust to out-of-distribution samples

To demonstrate the robustness to out-of-distribution examples, we adopted a similar concept used in [87]. We generated a synthetic dataset by applying time deformations defined in eq. (3.19) on three prototype signals: a Gabor wave, an apodized sawtooth wave, and an apodized square wave (shown in the top row of Fig. 3.5). Note that we used a single template per class to generate data for maintaining a simple experimental setup. We varied the magnitude of the confounding factors (i.e., the parameters used to calculate $g_j(t)$) to generate different distributions for training and testing sets. The 'in-distribution' set used during the training process consisted of signals with parameter values chosen randomly from smaller intervals with respect to the 'out-distribution' (testing) set. Table 3.3 shows the list of the parameters of interest for the 'out-of-distribution' experiment. The intervals (from which the parameter values were chosen) corresponding to the training

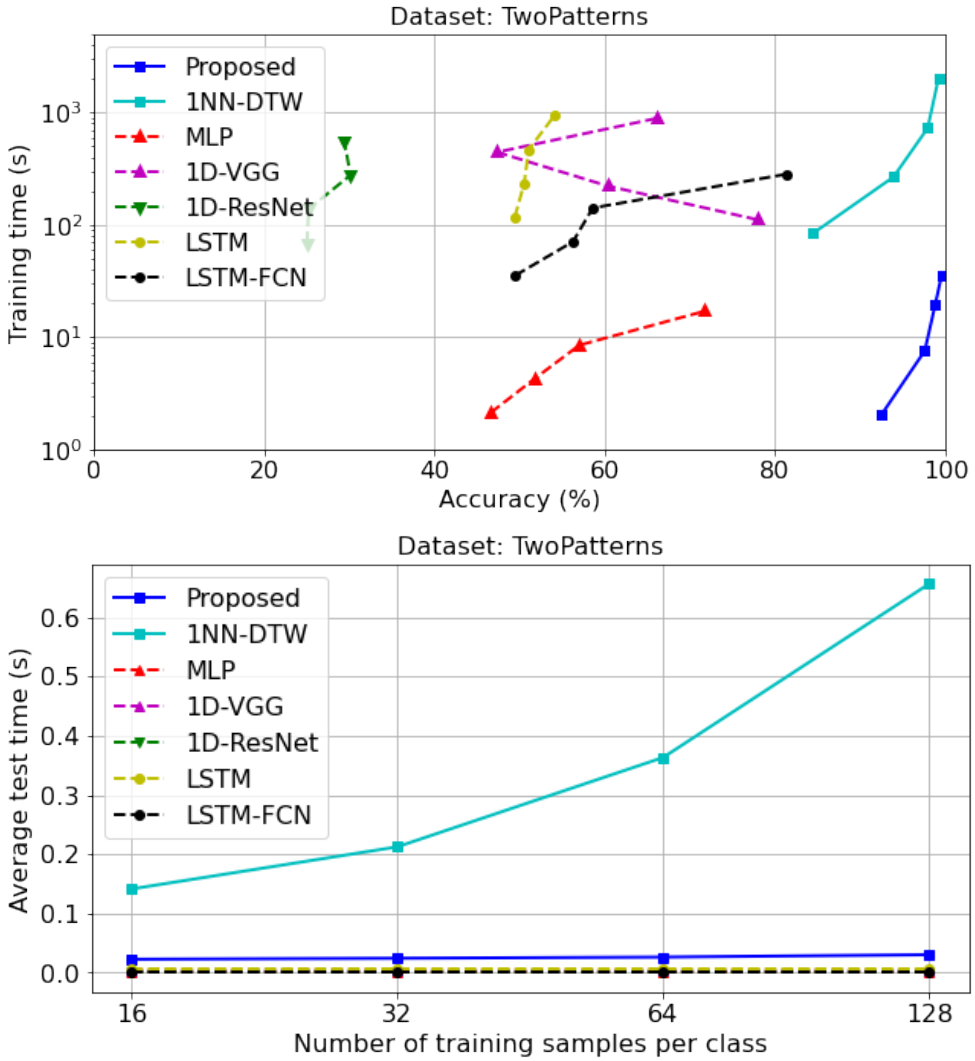


Figure 3.9: (Upper) Training time in seconds vs test accuracy for different classification methods. (Lower) Average time taken by the classification methods to predict the class of a test signal. Experiments were run using Python version 3.6.9 on a computer with an Intel(R) Xeon(R) CPU E5-2630 v4 processor running at 2.20GHz using 62 GB of RAM.

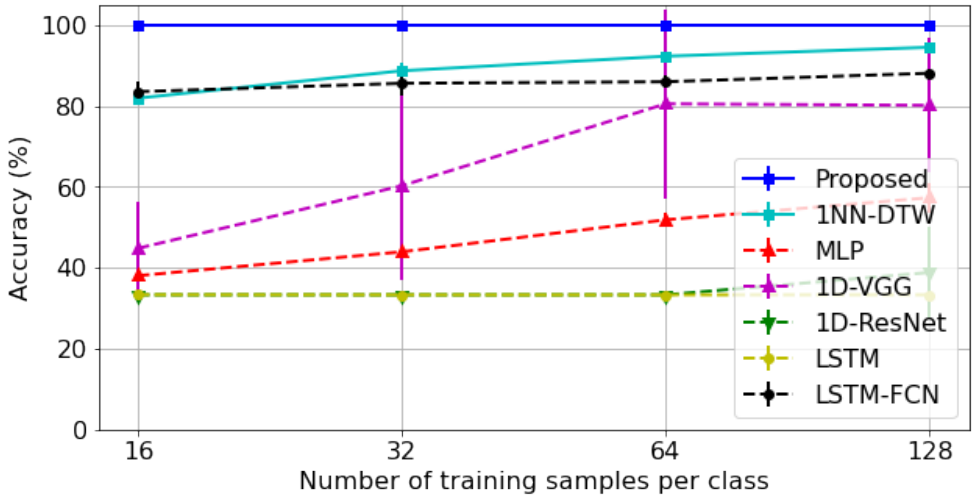


Figure 3.10: Accuracy as a function of number of training samples per class under the out-of-distribution setup.

and testing sets are also given in the table. Fig. 3.10 shows test accuracies with respect to the number of training samples per class for the comparing methods. The plot shows that the proposed method significantly outperforms other end-to-end solutions with very few training samples. Meaning, the proposed classification technique provides the best classification result if the test signal belongs to the 'out-distribution' set in the above sense but follows the transport generative model discussed in the previous section.

Parameter	Train	Test
N	$[2, 5]$	$[2, 10]$
μ_n	$\mathcal{N}(0.5, 0.2)$	$\mathcal{N}(0.5, 0.3)$
ω	$[0.9, 1.1]$	$[0.75, 1.25]$
τ	$[-0.05, 0.05]$	$[-0.1, 0.1]$

Table 3.3: Intervals used in the out-of-distribution setup.

To evaluate the performance of the proposed method on test signals that have undergone a different set of deformations from the training data, we performed a modified simulated experiment. We generated synthetic test signals by applying different types of time-warpings from the training set on the prototype signals shown in Fig. 3.5. The training set was generated according to the following model:

$$
\begin{aligned}
s_j^{(c)}(t) &= g_j'(t)\varphi_m^{(c)}(g_j(t)), \\
g_j(t) &= \omega\zeta(t) + \tau, \quad g_j'(t) > 0, \\
\zeta(t) &= \sum_{n=-N_t}^{N_t} \beta_n \left(t - \frac{\sin(\pi n t)}{|n|\pi} \right),
\end{aligned}
\tag{3.20}
$$

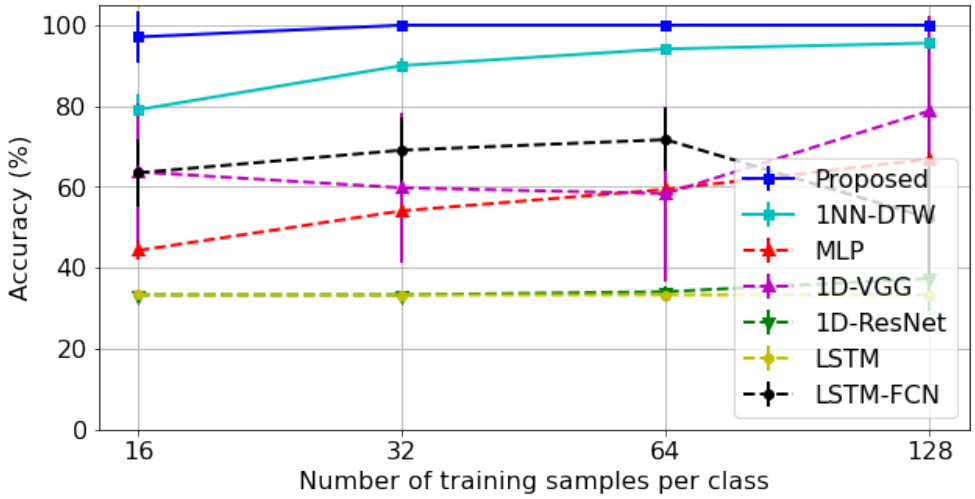
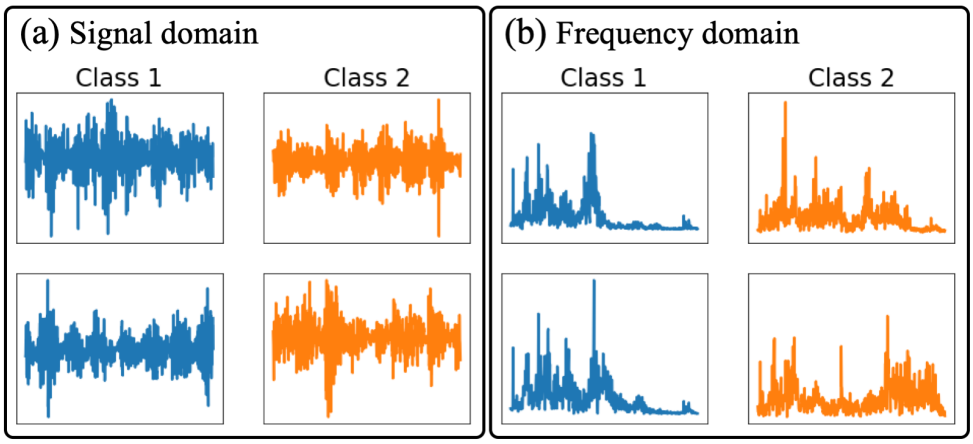


Figure 3.11: Performance of the classification methods when there is mismatch of deformation between training and testing sets.

where $\beta_n > 0, \sum_n \beta_n = 1$. The parameters to calculate $g_j(t)$ were randomly selected from fixed intervals. However, to keep the deformations in the test set different from the training signals, we followed the model described in eq. (3.19) to generate the test signals. The classification methods of interest were then trained with a varying number of training samples per class randomly chosen from the training set and evaluated on the test set. Fig. 3.11 shows the test accuracies as a function of training samples per class for the comparing methods. The plots show that the proposed method provides perfect test accuracy with only 32 training samples per class, while the other methods struggle to perform well under this experimental setup. The results demonstrate that the proposed method is not sensitive to the mismatch of deformation between training and testing sets as long as the deformations follow the definition in eq. (3.9).

3.4.7 A dataset that does not follow transport generative model

As described in 3.4.2, the proposed method performs well if the time series data is well segmented, i.e. contains events with well-defined start and end points. There are examples of signal classification problems where data do not follow these conditions. One such example is the gearbox fault diagnosis dataset [124], where vibration signals are used to detect a faulty gearbox. Fig. 3.12(a) shows that the vibration signals do not possess segmentable time series events. Hence, the proposed method is not suitable for classifying raw gearbox vibration data, as shown in the table in Fig. 3.12. However, we can transform each time series to frequency domain using the Fourier transform [125], where the support of the signals is relatively more compact than in the original signal domain. As expected the proposed method performs better when applied to the frequency domain signals shown in Fig. 3.12(b).



Method	Time domain	Frequency domain
MLP	55.67	90.45
1D-VGG	50.77	56.02
1D-ResNet	71.5	50.0
LSTM	57.32	72.35
LSTM-FCN	**76.5**	90.92
SCDT-NS	52.75	91.5
1NN-DTW	75.75	**93.5**
Proposed	57.25	89.5

Figure 3.12: Plots in (a) show few examples from the gearbox fault diagnosis dataset that do not follow the transport generative model. Plots in (b) show the signals in frequency domain corresponding to the signals shown in (a), which seem to fit the transport generative model; hence, the proposed method performs better in classifying these signals. Table in lower panel shows the test accuracy (%) of the classifiers on both time and frequency domain data.

3.5 Discussion

Classification test accuracies, rank-based statistics, and MPCE reported in Table 3.2 across 10 different time series datasets suggest that the proposed method is a very good generic end-to-end signal classification model for time series containing segmented events. Moreover, Fig. 3.7 shows that the proposed method can achieve high accuracy with few training samples. The computational efficiency of the proposed classifier is also demonstrated in figures 3.8 and 3.9 in terms of CPU operations, and average training and testing time with respect to other end-to-end classifiers. It is also evident from Table 3.2 that the proposed method outperforms the SCDT-NS classifier [87] which uses a single template-based transport generative model.

Another compelling property of the proposed method is the robustness to out-of-distribution samples since it generalizes to samples outside the known distribution when the signal classes conform to the specific transport generative model. Plots in Fig. 3.10 show that other methods fail to achieve good performance under the out-of-distribution setup, whereas the proposed method achieves perfect test accuracy with very few training samples ($\sim$ 16 per class). The reason behind the robustness to the out-of-distribution setup

is that the proposed method is capable of learning the underlying data model, more specifically, the types of unknown deformations present in the signals. It can then successfully classify an unknown signal in the presence of such deformations but with different magnitudes. Moreover, Fig. 3.11 demonstrates that the proposed method is less sensitive to the mismatch of deformations between training and testing sets with respect to other methods as long as the deformations follow the definition in eq. (3.9).

The main assumption of the proposed method is that the dataset needs to conform to the underlying transport generative model proposed earlier. Specifically, we showed above that the method works best when the time series (signal) being classified contains segmented events with finite duration. Fig. 3.12 shows an example where raw signals (from a gearbox vibration experimental setup) do not possess well-defined time series events of finite duration. Hence, the proposed method performs poorly in classifying those signals. However, the same signals in frequency domain seem to fit the transport generative model better and the proposed method performs better in classifying the gearbox data in frequency domain.

To summarize, this chapter presents a new idea of representing segmented signal data using a transport generative model such that signals from a particular class can be considered as observations of a set of unknown templates under some unknown deformations. Under this assumption, we formulated a classification problem for segmented signal classes. Then we showed that if the data follow the proposed transport generative model, a simple solution can be devised by searching nearest local subspace in SCDT domain. Through extensive experiments, we demonstrated that the proposed solution is effective in classifying unknown signals, computationally very cheap, data efficient, and robust to out-of-distribution samples.

3.6 Conclusion

This chapter introduced a new end-to-end signal classification method based on a recently developed signal transform. First, we formulated the problem statement based on a multiple template-based transport generative model observed under unknown deformations. Then, we proposed an end-to-end solution to the problem by employing a nearest local subspace search algorithm in SCDT domain. Although the problem statement and solution are based on the assumption that signals are observations of templates under confounding deformations, knowledge of these templates or confound deformations is not required. The model was demonstrated to achieve high test accuracy across multiple time series datasets. Several experiments show that the proposed method not only outperforms the state-of-the-art deep learning based end-to-end methods but also is data efficient and robust to out-of-distribution examples. Moreover, it provides competitive performance in classifying segmented time series events with respect to 1NN-DTW with very cheap computational complexity. We note that the approach assumes that the data being classified follows a certain transport generative

model. Namely, each time series should contain an event with finite support, that is, with beginning and end within the recorded time series. We showed that for signal classes that do not contain events of finite duration the approach is less effective.

Acknowledgements

This work is done in collaboration with Dr. Shiying Li, Dr. Xuwang Yin, Mohammad Shifat E. Rabbi, Dr. Yan Zhuang, and Dr. Gustavo K. Rohde, to whom I express my deep gratitude for their contributions to this chapter.

Chapter 4

Nonlinear System Identification Using the Signed Cumulative Distribution Transform In Structural Health Monitoring Applications

4.1 Introduction

Earlier chapters of this thesis presented techniques that utilize generative model-based approaches to solve non-linear signal estimation and classification problems. Our findings indicate that when data adheres to a specific generative model, some signal processing problems can effectively be solved with significantly high computational and data efficiency. This chapter aims to explore the feasibility of applying generative model-based solutions to structural health monitoring (SHM) applications. SHM involves the collection and analysis of sensor data to assess the condition of a structure [126]. In several SHM applications, parametric identification of a dynamical system is a critical aspect, wherein techniques for signal classification and estimation are utilized to understand the system's dynamics [48]. In this chapter, we utilize a generative model-based approach to solve the problem of nonlinear system identification in the field of structural dynamics.

In numerous SHM applications, the dynamics of a system is mathematically represented as a Partial Differential Equation (PDE). The task of identifying such a system involves retrieving the PDE coefficients. Consider a model dynamical system (depicted in Fig. 4.1) describing a 1D wave propagation phenomenon

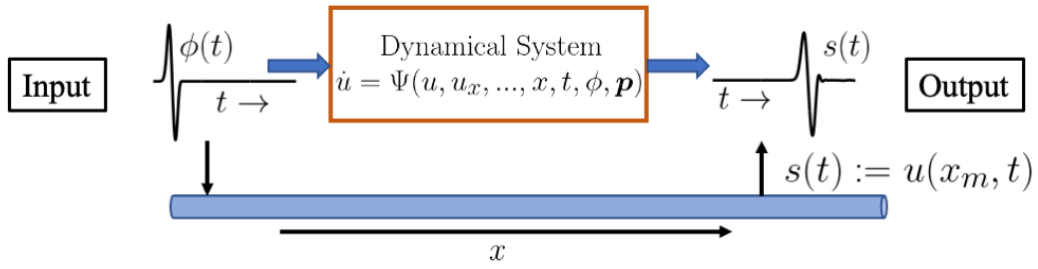


Figure 4.1: A model dynamical system describing 1D wave propagation through an elastic medium.

through an elastic medium. Typically, the dynamics of such systems are expressed as partial differential equations (PDEs) [60] in the following form:

$$\dot{u} = \Psi(u, u_x, u_{xx}, ..., x, t, \phi, \boldsymbol{\xi}), \tag{4.1}$$

where $\dot{u} = \frac{\partial u}{\partial t}$, u_x and u_{xx} are the first and second order derivatives of u with respect to x, respectively. Ψ represents a partial differential equation-based model for the dynamical system that is parameterized with $\boldsymbol{\xi}$. $\phi(t)$ denotes the source signal that initiates the propagation and $s(t)$ is the sensor measurement measured at a specific location as a function of time (as depicted in Fig. 4.1). The main objective of system identification problems is to estimate the system coefficients $\boldsymbol{\xi}$ of the governing PDE of the dynamical system using the sensor measurements s. In this work, the focus is primarily on identifying the coefficients associated with damage-induced nonlinearities.

The field of Structural Health Monitoring (SHM) comprises a body of work aimed at the detection, identification, and ultimately forecasting of structural damage. Typically it is presumed that this information is to be inferred from the dynamic response of the structure to ambient or applied excitation [43]. While numerous approaches to the problem exist, they can be loosely categorized in terms of the *a priori* information required for their implementation.

On one hand, we can view the problem as one of "statistical pattern recognition" [127] where by certain properties of the response signal (referred to in the literature as "features") are used to classify the response as coming from a particular damage state. Recent advances in machine learning have furthered research in this general approach [128]. A key challenge with the associated data models lies in capturing the damage-induced aspects of a structural response while ignoring those due to covariates (e.g., a change in the frequency response due to damage vs. a change due to temperature variation) [129]. A system that is unable to distinguish among these sources will produce an unacceptably high number of "false positives", mis-categorizing other influences on the system response as being damage related.

Another class of techniques views the problem as one of model-based system identification [43]. In this setting, a model of the structure, including the damage, is formulated and various estimation methods are used to identify the parameters related to structural damage. One of the main advantages of such an approach is at least partial immunity to signal fluctuations un-related to damage. The model is explicitly separating the physics of damage from these "other" sources which are collectively modeled as noise. The challenge with this approach, of course, lies in the modeling. Predicting a structures response to a particular excitation is quite challenging, as the exact forcing function is often unknown and even simple structures behave in non-idealized fashion in practice (see e.g., [130] for an example).

These different viewpoints (data driven vs. model based) are not mutually exclusive, of course, and represent endpoints on the continuum of *a priori* information we wish to bring to bear on the problem. The pattern recognition approach presumes very little knowledge and seeks outliers of a data model, while the model-based approach attempts to leverage structural mechanics to refine the search and subsequent estimation of the damage. Both approaches must deal with the uncertainty inherent in measurements (e.g. sensor noise) and other possible sources of "clutter" that can corrupt the structures response.

The method proposed in this work lies somewhere in between these aforementioned extremes. We presume some basic knowledge of the structure, namely that it supports traveling wave solutions and that damage results in the presence of nonlinearity in the underlying model. This latter assumption has been widely applied and forms the basis for a number of techniques that focus on damage-induced nonlinearities [131, 132]. However, we do not presume to know or be able to measure the excitation signal. The underlying assumption is that the sensor data collected at a particular location can be considered as an instance of a template pattern under some time-warpings. This assumption is suitable for analyzing solutions to various classical PDE systems, such as wave equation and convection-diffusion equation at a given location. Given the premise, the proposed approach employs the SCDT nearest local subspace classifier (SCDT-NLS) [133] to detect the presence and severity of the nonlinearity induced by structural damage. Fig. 4.2 demonstrates the outline of the proposed system identification approach. Note that the proposed method does not necessitate prior knowledge of the pattern template or time warpings. Instead, it utilizes the sensor measurements with known system coefficients as training samples to recover the system parameters for an unknown test sample.

The remainder of this chapter is structured as follows: Section 4.2 provides a brief overview of the definitions and explanations of CDT, SCDT, and SCDT-NLS. In Section 4.3, we present the generative model-based identification problem and propose a solution. Section 4.4 presents the analytical derivation of the proposed generative model for some classical PDEs, while extensive experimentation is carried out in Section 4.5 to evaluate the proposed solution. Section 4.6 summarizes the knowledge learned from the experiments, and finally, Section 4.7 concludes this chapter.

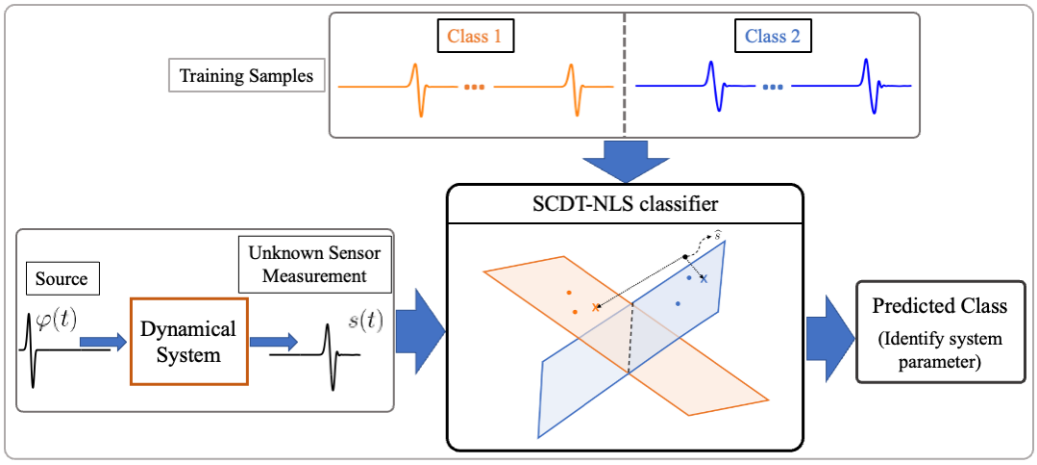


Figure 4.2: Outline of the proposed system identification approach.

Table 4.1: Description of symbols

Symbols	Description
$\phi(t)$	Source signal that excites the propagation medium
$s(t)$	Sensor measurement as a function of time
$s_0(y)$	Reference signal to calculate the transform
$\widehat{s}(y)$	SCDT of signal $s(t)$
$\varphi(t)$	Template pattern corresponding to a class of signals
$g(t)$	Strictly increasing and differentiable function
$s \circ g$	$s(g(t))$: composition of $s(t)$ with $g(t)$
$\mathcal{T}$	Set of all possible increasing diffeomorphisms
$\mathbb{S}/\widehat{\mathbb{S}}$	Set of signals/SCDT of the signals

4.2 Preliminaries

4.2.1 Notation

In the following sub-sections, we introduce CDT, SCDT and SCDT-NLS for signals $s \in L_1(\Omega_s)$ on the signal domain $\Omega_s \subseteq \mathbb{R}$. We denote $s_j^{(c)}$ as a signal generated by a warping map g_j acting on template $\varphi^{(c)}$ of class c. Some other notations used throughout this chapter are listed in Table 4.1.

4.2.2 The Cumulative Distribution Transform

CDT [134] of positive smooth normalized functions is an invertible nonlinear 1D signal transform from the space of smooth positive probability densities to the space of diffeomorphisms. Given, a signal $s(t), t \in \Omega_s$ and a reference signal $s_0(y), y \in \Omega_{s_0} \subseteq \mathbb{R}$ such that $\int_{\Omega_s} s(u)du = \int_{\Omega_{s_0}} s_0(u)du = 1$ and $s_0(y), s(t) > 0$ in

their respective domains. The CDT of the signal $s(t)$ is the function $s^*(y)$ computed as:

$$\int_{\inf(\Omega_s)}^{s^*(y)} s(u)du = \int_{\inf(\Omega_{s_0})}^{y} s_0(u)du\,, \tag{4.2}$$

whose inverse is defined in differential form as:

$$s(t) = \left(s^{*^{-1}}(t)\right)' s_0(s^{*^{-1}}(t)). \tag{4.3}$$

Note, the CDT $s^*(y)$ is alternatively defined as:

$$s^*(y) = S^{-1}(S_0(y))\,, \tag{4.4}$$

where $S(t) = \int_{-\infty}^{t} s(u)du$ and $S_0(y) = \int_{-\infty}^{y} s_0(u)du$. If the reference signal is uniform, i.e. $s_0(y) = 1$ in $\Omega_{s_0} = [0, 1]$, we have $S_0(y) = y$ and $s^*(y) = S^{-1}(y)$. It means the CDT is the inverse of the cumulative distribution function of the given signal $s(t)$.

Although the CDT can widely be used in classification [134] and estimation [135] problems, the CDT framework is defined only for positive density functions. Aldroubi et al. [88] proposed the signed cumulative distribution transform (SCDT) as an extension of the CDT to general finite signed signals.

4.2.3 The Signed Cumulative Distribution Transform

SCDT [88] is an extension of the CDT for general finite signed signals without requirements on the total mass. Given a signal $s(t)$, Jordan decomposition of a signed signal $s(t)$ is given by $s(t) = s^+(t) - s^-(t)$, where $s^+(t)$ and $s^-(t)$ are the absolute values of the positive and negative parts of the signal $s(t)$. The SCDT of $s(t)$ with respect to $s_0(y)$ is then defined as:

$$s(t) \overset{\text{SCDT}(s_0)}{\longleftrightarrow} \widehat{s}(y) = \left(\widehat{s^+}(y), \widehat{s^-}(y)\right), \tag{4.5}$$

where $\widehat{s^+}(y)$ and $\widehat{s^-}(y)$ are the transforms for the signals $s^+(t)$ and $s^-(t)$ as:

$$\widehat{s^\pm}(y) = \begin{cases} \left((s^\pm)^*(y), \|s^\pm\|_{L_1}\right), & \text{if } s \neq 0 \\ (0, 0), & \text{if } s = 0, \end{cases} \tag{4.6}$$

with L_1 norm $\|\cdot\|_{L_1}$ and $(s^\pm)^*$ as the CDT of the normalized signal $\frac{s^\pm}{\|s^\pm\|_{L_1}}$ with respect to a strictly positive reference signal s_0. Fig. 4.3 demonstrates the SCDT calculation of an example signal. The inverse SCDT is

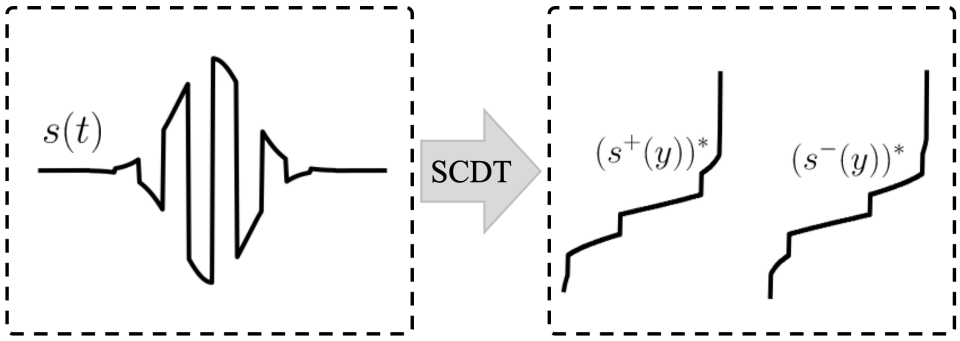


Figure 4.3: SCDT (without the constant terms) of an example signal.

defined as:

$$s(t) = \|s^+\|_{L_1} \left((s^+)^{*^{-1}}(t)\right)' s_0((s^+)^{*^{-1}}(t)) - \|s^-\|_{L_1} \left((s^-)^{*^{-1}}(t)\right)' s_0((s^-)^{*^{-1}}(t)). \tag{4.7}$$

The SCDT has a number of properties that are useful for the signal classification problems.

Composition property

Composition property states that the SCDT of the signal $s_g = g's \circ g$, is defined as:

$$\widehat{s}_g = \left(g^{-1} \circ (s^+)^*, \|s^+\|_{L_1}, g^{-1} \circ (s^-)^*, \|s^-\|_{L_1}\right), \tag{4.8}$$

where $g(t)$ is an invertible smooth warping map, $s \circ g = s(g(t))$ and $g'(t) = dg(t)/dt$ [88]. For example, a shift and linear dispersion (i.e., $g(t) = \omega t - \tau$) of a given signal $s(t)$ is $s_g(t) = \omega s(\omega t - \mu)$ having SCDT:

$$\widehat{s}_g = \left(\frac{(s^+)^* + \mu}{\omega}, \|s^+\|_{L_1}, \frac{(s^-)^* + \mu}{\omega}, \|s^-\|_{L_1}\right).$$

The composition property implies that variations along the independent variable caused by $g(t)$ will change only the dependent variable in the transform domain.

Convexity property

Given a signal φ and a set of 1D temporal deformations $\mathcal{G}$ (e.g., translation, dilation, etc), the set of the SCDTs of the signals from $\mathbb{S} = \{s_j : s_j = g_j' \varphi \circ g_j, \forall g_j \in \mathcal{G}\}$ is given by $\widehat{\mathbb{S}} = \{\widehat{s}_j : \widehat{s}_j = g_j^{-1} \circ \widehat{\varphi}, s_j \in \mathbb{S}\}$ (using composition property). The convexity property of the SCDT [88] states that the set $\widehat{\mathbb{S}}$ is convex for every φ if and only if $\mathcal{G}^{-1} = \{g_j^{-1} : g_j \in \mathcal{G}\}$ is convex. The set $\mathbb{S}$ defined above can be interpreted as a transport generative model for a signal class while φ being the template signal corresponding to that class.

Numerical Implementation of the SCDT

The SCDT described above is defined for the continuous-time signals. In this section, we describe the numerical method for approximating the SCDT given discrete signals. Let $\mathbf{s} = [s_1, s_2, ..., s_N]^T$ be a N-point discrete-time signal, where $s_n = s[n], \forall n = 1, 2, ..., N$ is the n-th sample of $\mathbf{s}$. The positive and negative parts of the signal after Jordan decomposition are given by $\mathbf{s}^{\pm} = \left[s_1^{\pm}, ..., s_N^{\pm}\right]^T$, respectively, where $s_n^{\pm} = \frac{|s_n| \pm s_n}{2}$, and $|s_n|$ is the absolute value of s_n. Next, the CDT is applied numerically to the normalized signals $\frac{\mathbf{s}^+}{\|\mathbf{s}^+\|_{\ell_1}}$ and $\frac{\mathbf{s}^-}{\|\mathbf{s}^-\|_{\ell_1}}$, where $\|\cdot\|_{\ell_1}$ is the ℓ_1-norms. As the CDT $(s^{\pm})^*(y)$ is the inverse of the CDF of the signal $s^{\pm}(t)$ for a particular choice of reference signal ($s_0(y) = 1$ for $y \in [0, 1]$), we need to approximate the cumulative function first. The numerical approximation of the cumulative function is given by,

$$S^{\pm}[n] = \sum_{i=1}^{n} \frac{s^{\pm}[n]}{\|\mathbf{s}^{\pm}\|_{\ell_1}}, \qquad n = 1, 2, ..., N,$$

where $\mathbf{S}^+$ and $\mathbf{S}^-$ are the cumulation of the normalized signals $\frac{\mathbf{s}^+}{\|\mathbf{s}^+\|_{\ell_1}}$ and $\frac{\mathbf{s}^-}{\|\mathbf{s}^-\|_{\ell_1}}$, respectively. The CDT is then calculated by taking the generalized inverse of the CDF,

$$(s^{\pm})^*[m] = min\left(\{t[n] : S^{\pm}[n] > y[m]\}\right),$$

where $t \in \Omega_s$, $y \in \Omega_{s_0}$, and $n, m = 1, 2, ..., N$. Here, $(\mathbf{s}^+)^*$ and $(\mathbf{s}^-)^*$ are the CDTs of the normalized discrete signals $\frac{\mathbf{s}^+}{\|\mathbf{s}^+\|_{\ell_1}}$ and $\frac{\mathbf{s}^-}{\|\mathbf{s}^-\|_{\ell_1}}$, respectively. The SCDT of the discrete signal $\mathbf{s}$ is then given by,

$$\widehat{\mathbf{s}} = \left((\mathbf{s}^+)^*, \|\mathbf{s}^+\|_{\ell_1}, (\mathbf{s}^-)^*, \|\mathbf{s}^-\|_{\ell_1}\right).$$

Note that the computational complexity of calculating the SCDT is $O(N \log N)$ for an N-point discrete-time signal.

4.2.4 Nearest Local Subspace Classifier in SCDT Domain

SCDT-NLS [133] was proposed as an end-to-end classification technique to classify segmented signal classes that can be seen as observations of a set of template patterns that have undergone some temporal deformations.

Such class of signals can be modeled by the generative model defined as:

$$\begin{aligned}
&\mathbb{S}^{(c)} = \bigcup_{m=1}^{M_c} \mathbb{S}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}}, \\
&\mathbb{S}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}} = \left\{ s_{j,m}^{(c)} | s_{j,m}^{(c)} = g_j' \varphi_m^{(c)} \circ g_j, g_j' > 0, g_j \in \mathcal{G}_m^{(c)} \right\}, \\
&\left(\mathcal{G}_m^{(c)}\right)^{-1} = \left\{ \sum_{i=1}^{k} \alpha_i f_{i,m}^{(c)}, \alpha_i \geq 0 \right\},
\end{aligned} \tag{4.9}$$

where c denotes a particular class, $\left\{ f_{1,m}^{(c)}, f_{2,m}^{(c)}, ..., f_{k,m}^{(c)} \right\}$ denotes a set of linearly independent and strictly increasing (within the domain of the signals) functions, and $\mathcal{G}_m^{(c)}$ denotes a set of increasing 1D temporal deformations. It states that the generative model for a given class c can be expressed as the collection of M_c subsets, where each subset ($\mathbb{S}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}}$) represents data generated by a specific template signal ($\varphi_m^{(c)}$) under various temporal deformations ($\mathcal{G}_m^{(c)}$). Here, M_c refers to the total number of templates utilized to represent class c, while $\varphi_m^{(c)}$ denotes the m-th template signal from class c, and $s_{j,m}^{(c)}$ corresponds to the j-th signal generated from the m-th template signal, subjected to the deformation defined by g_j.

Since $\left(\mathcal{G}_m^{(c)}\right)^{-1}$ is convex by definition, the set $\mathbb{S}^{(c)}$ can be represented as a collection of convex sets in the SCDT space, as per the convexity property. the generative model in the transform space is given by:

$$\begin{aligned}
&\widehat{\mathbb{S}}^{(c)} = \bigcup_{m=1}^{M_c} \widehat{\mathbb{S}}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}}, \\
&\widehat{\mathbb{S}}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}} = \left\{ \widehat{s}_{j,m}^{(c)} | \widehat{s}_{j,m}^{(c)} = g_j^{-1} \circ \widehat{\varphi}_m^{(c)}, g_j \in \mathcal{G}_m^{(c)} \right\},
\end{aligned} \tag{4.10}$$

where $g_j^{-1} \circ \widehat{\varphi}_m^{(c)}$ is the SCDT of the signal $g_j' \varphi_m^{(c)} \circ g_j$. Under the assumption that $\mathbb{S}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}} \cap \mathbb{S}_{\varphi_w^{(p)},\mathcal{G}_w^{(p)}} = \varnothing$ for $c \neq p$, and an unknown sample s is generated according to the generative model, the unknown class label can be uniquely predicted by solving,

$$\underset{c}{\operatorname{argmin}} \min_{m} \; d^2 \left(\widehat{s}, \widehat{\mathbb{V}}_m^{(c)} \right), \tag{4.11}$$

where $\widehat{\mathbb{V}}_m^{(c)} = \text{span}\left(\widehat{\mathbb{S}}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}} \right)$. Given a set of training samples $\left\{ s_1^{(c)}, ..., s_j^{(c)}, ..., s_{L_c}^{(c)} \right\} \subset \mathbb{S}^{(c)}$ for class c, the unknown class of a test sample s is estimated in two steps:

Step 1: A set of k closest training samples to $\widehat{s}$ from class c are chosen based on the distance between $\widehat{s}$ and the span of each training sample. First, the elements from the set $\{\widehat{s}_1^{(c)}, ..., \widehat{s}_{L_c}^{(c)}\}$ are sorted into $\{\widehat{z}_1^{(c)}, ..., \widehat{z}_{L_c}^{(c)}\}$

such that

$$d^2(\widehat{s}, \widehat{\mathbb{V}}_{z_1}^{(c)}) \leq \cdots \leq d^2(\widehat{s}, \widehat{\mathbb{V}}_{z_l}^{(c)}) \leq \cdots, \tag{4.12}$$

where $\widehat{\mathbb{V}}_{z_l}^{(c)} = \text{span}\left(\{\widehat{z}_l^{(c)}\} \cup \mathbb{U}_T \cup \mathbb{U}_H\right)$, $\mathbb{U}_T$ and $\mathbb{U}_H$ are spanning sets (defined in [133]) used to enrich the subspace. First k elements from the sorted set are chosen to form $\{\widehat{z}_1^{(c)}, \cdots, \widehat{z}_k^{(c)}\}$ for $k \leq L_c$, which gives the set of k closest training samples to $\widehat{s}$ from class c in the above sense. This step is repeated for all the signal classes, i.e., $c = 1, 2, ...,$ etc.

Step 2: The set $\{\widehat{z}_1^{(c)}, \cdots, \widehat{z}_k^{(c)}\} \cup \mathbb{U}_T \cup \mathbb{U}_H$ is orthogonalized to obtain the basis vectors $\{b_1^{(c)}, b_2^{(c)}, ...\}$. These basis vectors are used to construct a matrix $B^{(c)}$ for each signal class c, i.e., $B^{(c)} = \left[b_1^{(c)}, b_2^{(c)}, ...\right]$ for $c = 1, 2, ...,$ etc. The unknown class of s is then estimated by:

$$\arg\min_c \ \|\widehat{s} - B^{(c)} {B^{(c)}}^T \widehat{s}\|^2, \tag{4.13}$$

where $\|.\|$ denotes L_2 norm.

4.3 Proposed Method

In this work, we explore a transport transform-based data-driven approach to propose an effective and data-efficient solution for the nonlinear system identification problem.

4.3.1 Nonlinear System Identification

Consider a dynamical system that describes the propagation of 1D waves through a damaged elastic medium. The corresponding PDE equation for this system can be expressed as follows:

$$\rho\ddot{u} - Eu_{xx} + \eta\dot{u}_{xx} - \rho M\ddot{u}_{xx} - Fu_{xxxx} + \beta u_x u_{xx} = 0, \tag{4.14}$$

where ρ is the mass density of the medium and E is Young's modulus. Here, η denotes the dissipation parameter, and the model parameters M and F account for inertial and nonlocal elastic effects, respectively. The nonlinearity coefficient is represented by β. The third term in equation (4.14) corresponds to dissipation, while the fourth and fifth terms are related to dispersion, and the sixth term models nonlinearity. Let us assume that a source signal $\phi(t)$ initiates wave propagation (with positive speed ν) through the elastic medium, and a sensor placed at a fixed location $x = x_m$ measures a signal $s(t)$. The PDE in eq. (4.14) can alternatively be

represented as:

$$\xi_0\ddot{u} - \xi_1 u_{xx} + \xi_2\dot{u}_{xx} - \xi_3\ddot{u}_{xx} - \xi_4 u_{xxxx} + \beta u_x u_{xx} = 0, \tag{4.15}$$

where $\boldsymbol{\xi} = [\xi_0, \xi_1, \xi_2, \xi_3, \xi_4]^T$ are the coefficients of the linear terms of the PDE. The objective of this work is to recover the nonlinearity parameter β using the sensor data $s(t)$ measured at a particular location $x = x_m$.

4.3.2 Generative Model and Problem Statement

This work aims to propose a generative model-based problem statement to identify the nonlinearity parameter β using the sensor measurement $s(t)$ where each measured signal can be modeled as an instance of a particular template observed under some unknown deformation such as translation, dispersion, or time warping. Such a family of signals can be described with the following generative model:

Generative Model

Let the coefficients in $\boldsymbol{\xi}$ vary within a given range, such that $\xi_i \in [\mathcal{E}_i - \epsilon, \mathcal{E}_i + \epsilon]$ for $i = 0, 1, ..., 4$ and $\mathcal{E}_i \in \mathbb{R}$. Let $\mathcal{G}_{\boldsymbol{\xi}}^{(\beta)} \subset \mathcal{T}$ denote a set of increasing 1D deformations of a specific kind, where $\mathcal{T} : \mathbb{R} \to \mathbb{R}$ is a set of all possible increasing diffeomorphisms. Given a β, there exists a template pattern $\varphi_{\boldsymbol{\xi}}^{(\beta)}(x, t)$ and a warping function $g_j(x, t) \in \mathcal{G}_{\boldsymbol{\xi}}^{(\beta)}$ for small ϵ such that the family of the sensor measurement $s(t)$ at a fixed location $x = x_m$ can be modeled as:

$$\mathbb{S}_{\varphi_{\boldsymbol{\xi}}^{(\beta)}, \mathcal{G}_{\boldsymbol{\xi}}^{(\beta)}} = \{s_j^{(\beta)} | s_j^{(\beta)} = \dot{g}_j \varphi_{\boldsymbol{\xi}}^{(\beta)} \circ g_j, g_j \in \mathcal{G}_{\boldsymbol{\xi}}^{(\beta)}, \dot{g}_j > 0\}, \tag{4.16}$$

where $s_j^{(\beta)}$ is the j-th signal under a given nonlinearity (β), and $\dot{g}_j$ represents first order derivative of g_j with respect to t. Note that since the sensor location x is fixed at x_m, the signals s_j, $\varphi_{\boldsymbol{\xi}}^{(\beta)}$, and the warping functions g_j are represented as a function of time t only, e.g., $\varphi_{\boldsymbol{\xi}}^{(\beta)}(t) := \varphi_{\boldsymbol{\xi}}^{(\beta)}(x = x_m, t)$, $g_j(t) := g_j(x = x_m, t)$, etc.

The generative model described in eq. (4.16) is simply a Lagrangian restatement of the continuity equation and expresses the conservation of signal intensity at a particular sensor location. Such models are common in many wave propagation phenomena, e.g., wave equation, diffusion equation, convection-diffusion equation, etc. The warping function $g_j(\cdot)$ transforms the independent variable (time, t in this problem) according to the structural dynamics (including damage), and the effect of this function is to morph the template signal $\varphi_{\boldsymbol{\xi}}^{(\beta)}$ into the observed signal $s_j^{(\beta)}$.

Problem Statement

Considering the generative model stated above, the nonlinear identification problem for the dynamical system described by the PDE shown in eq. (4.15) is defined as follows:

Identification problem: *Let* $\mathcal{G}_{\boldsymbol{\xi}}^{(\beta)} \subset \mathcal{T}$ *be a set of increasing temporal deformations, and* $\mathbb{S}_{\varphi_{\boldsymbol{\xi}}^{(\beta)},\mathcal{G}_{\boldsymbol{\xi}}^{(\beta)}}$ *be defined as in eq. (4.16). Given a set of sensor measurements* $\{s_1^{(\beta)}, s_2^{(\beta)}, ...\} \subset \mathbb{S}_{\varphi_{\boldsymbol{\xi}}^{(\beta)},\mathcal{G}_{\boldsymbol{\xi}}^{(\beta)}}$ *with known* $\beta \in \mathbb{R}$*, identify the nonlinearity parameter* β *for an unknown measurement* s*.*

In this work, we further split this nonlinear identification problem into two sub-problems:

- **Detect nonlinearity:** The problem of detecting nonlinearity can be framed as a binary classification problem where classes 1 and 2 represent the sensor data corresponding to $\beta = 0$ and $\beta \neq 0$, respectively. Signals from class c (for $c = 1, 2$) can be modeled as:

$$\begin{aligned} \mathbb{S}^{(c)} &= \bigcup_{m=1}^{M_c} \mathbb{S}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}}, \\ \mathbb{S}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}} &= \{s_{j,m}^{(c)} | s_{j,m}^{(c)} = \dot{g}_j \varphi_m^{(c)} \circ g_j, g_j \in \mathcal{G}_m^{(c)}, \dot{g}_j > 0\}, \end{aligned} \tag{4.17}$$

  where $\varphi_m^{(c)}$ is the template pattern corresponding to $\beta_m^{(c)}$ from class c, and $s_{j,m}^{(c)}$ is the j-th signal generated from $\varphi_m^{(c)}$ under the deformation defined by g_j. Given the generative model for the signals from class c, the nonlinearity detection problem can be defined as follows:

  Let $\mathcal{G}_m^{(c)} \subset \mathcal{T}$ *be a set of increasing temporal deformations, and* $\mathbb{S}^{(c)}$ *be defined as in eq. (4.17), for classes* $c = 1, 2$*. Given a set of sensor measurements* $\{s_1^{(c)}, s_2^{(c)}, ...\} \subset \mathbb{S}^{(c)}$ *for class* c*, determine the class label of an unknown signal* s*, meaning, detect if the nonlinearity parameter* β *is zero or not.*

- **Estimate nonlinearity parameter:** The objective is to estimate the value of the nonlinearity parameter β for an unknown measurement s, given a set of sensor measurements $\{s_1^{(\beta)}, s_2^{(\beta)}, ...\} \subset \mathbb{S}_{\varphi_{\boldsymbol{\xi}}^{(\beta)},\mathcal{G}_{\boldsymbol{\xi}}^{(\beta)}}$ with known β values. However, obtaining a precise estimation of β requires a large number of training samples from the set $\mathbb{S}_{\varphi_{\boldsymbol{\xi}}^{(\beta)},\mathcal{G}_{\boldsymbol{\xi}}^{(\beta)}}$, which may not be feasible in a real-world scenario. In this work, the estimation of the nonlinearity parameter is framed as a coarse regression problem where a range $[l_\beta, h_\beta]$ is predicted for the unknown β. Furthermore, this regression problem can be viewed as a multi-class classification problem, where each class represents signals corresponding to β values ranging from l_β to h_β. Signals from class c (for $c = 1, 2, ..., N_c$) can be modeled by the generative model defined in eq. (4.17), where the set $\mathbb{S}_{\varphi_m^{(c)},\mathcal{G}_m^{(c)}}$ corresponds to $\beta_m^{(c)} \in [l_{\beta_m^{(c)}}, h_{\beta_m^{(c)}}]$. Given the generative model for class c, the problem of estimating nonlinearity parameter is defined as follows:

Let $\mathcal{G}_m^{(c)} \subset \mathcal{T}$ be a set of increasing temporal deformations, and $\mathbb{S}^{(c)}$ be defined as in eq. (4.17), for classes $c = 1, 2, ..., N_c$. Given a set of sensor measurements $\{s_1^{(c)}, s_2^{(c)}, ...\} \subset \mathbb{S}^{(c)}$ for class c, determine the class label of an unknown signal s, meaning, estimate a range $[l_\beta, h_\beta]$ for the nonlinearity parameter β.

4.3.3 Proposed Solution

The proposed solution to the classification problems defined earlier involves using the SCDT-NLS method [133] described in section 4.2.4. The generative model defined in equation (4.17) typically results in nonconvex signal classes, making the classification problems challenging to solve. However, as explained in section 4.2.3, we can simplify the geometry of the signal classes under certain assumptions by using the SCDT. Hence, the proposed solution begins with applying the SCDT on the sensor measurements. The generative model in SCDT domain is then given by,

$$\begin{aligned}
\widehat{\mathbb{S}}^{(c)} &= \bigcup_{m=1}^{M_c} \widehat{\mathbb{S}}_{\varphi_m^{(c)}, \mathcal{G}_m^{(c)}}, \\
\widehat{\mathbb{S}}_{\varphi_m^{(c)}, \mathcal{G}_m^{(c)}} &= \{\hat{s}_{j,m}^{(c)} | \hat{s}_{j,m}^{(c)} = g_j^{-1} \circ \widehat{\varphi}_m^{(c)}, g_j^{-1} \in \left(\mathcal{G}_m^{(c)}\right)^{-1}\},
\end{aligned} \tag{4.18}$$

where $g_j^{-1} \circ \widehat{\varphi}_m^{(c)}$ refers to the SCDT of the signal $\dot{g}_j \varphi_m^{(c)} \circ g_j$. In many wave propagation phenomena, such as those described by the wave equation, convection-diffusion equation, etc., the set $\left(\mathcal{G}_m^{(c)}\right)^{-1}$ can be proven to be convex (shown in the next section). As a result, by utilizing the convexity property of the SCDT outlined in section 4.2.3, we can demonstrate that the set $\widehat{\mathbb{S}}_{\varphi_m^{(c)}, \mathcal{G}_m^{(c)}}$ defined in equation (4.18) forms a convex set. Moreover, since the SCDT is a one-to-one mapping, if $\mathbb{S}_{\varphi_m^{(c)}, \mathcal{G}_m^{(c)}} \cap \mathbb{S}_{\varphi_w^{(p)}, \mathcal{G}_w^{(p)}} = \varnothing$ for $c \neq p$, then $\widehat{\mathbb{S}}_{\varphi_m^{(c)}, \mathcal{G}_m^{(c)}} \cap \widehat{\mathbb{S}}_{\varphi_w^{(p)}, \mathcal{G}_w^{(p)}} = \varnothing$.

To formulate the solution to the nonlinearity detection and estimation problems defined as binary and multi-class classification problems, respectively, in section 4.3.2, we adopt the local subspace-based technique proposed in [133]. Under the assumption that an unknown test sample s is generated according to the generative model for one of the classes, the unknown class label can be predicted by solving,

$$\underset{c}{\operatorname{argmin}} \min_{m} \; d^2\left(\hat{s}, \widehat{\mathbb{V}}_m^{(c)}\right), \tag{4.19}$$

where $d(\hat{s}, \widehat{\mathbb{V}}_m^{(c)})$ is the Euclidean distance between $\hat{s}$ and the nearest point in subspace $\widehat{\mathbb{V}}_m^{(c)}$, and $\widehat{\mathbb{V}}_m^{(c)}$ is given by,

$$\widehat{\mathbb{V}}_m^{(c)} = \operatorname{span}\left(\widehat{\mathbb{S}}_{\varphi_m^{(c)}, \mathcal{G}_m^{(c)}}\right). \tag{4.20}$$

We then exploit the nearest local subspace search algorithm (outlined in [133]) in SCDT domain to solve this classification problem.

Consider a set of training samples $\left\{s_1^{(c)}, ..., s_j^{(c)}, ..., s_{L_c}^{(c)}\right\} \subset \mathbb{S}^{(c)}$ for class c, where L_c is the total number of training samples given for class c, and $s_j^{(c)}$ is the j-th sample. The training and testing phases of the algorithm are described below:

Training Phase

During the training phase of the algorithm, the subspace that corresponds to each training sample is computed. Firstly, the SCDTs are computed for all training samples from class c. Then, we take a training sample $\widehat{s}_l^{(c)}$ and orthogonalize $\{\widehat{s}_l^{(c)}\} \cup \mathbb{U}_T \cup \mathbb{U}_H$ to obtain the basis vectors that span the enriched subspace corresponding to that sample. Here, $\mathbb{U}_T$ and $\mathbb{U}_H$ are the spanning sets (defined in section 3.3.3) to enrich the subspace with translation and other time-warpings, respectively. Let $B_l^{(c)} = \left[b_{l,1}^{(c)}, b_{l,2}^{(c)}, ...\right]$ be a matrix that contains the basis vectors in its column. These calculations are repeated for all the training samples to form $B_l^{(c)}$ for $l = 1, 2, ..., L_c$ and $c = 1, 2, ...$, etc. Note that we employ the same subspace enrichment technique as described in [133].

Testing Phase

The testing phase begins with taking SCDT of the test sample s to obtain $\widehat{s}$ followed by the nearest local subspace search in SCDT domain. In the first step of the algorithm, we estimate the distance of the subspace corresponding to each of the training samples from $\widehat{s}$ by:

$$\epsilon_l = \|\widehat{s} - B_l^{(c)} {B_l^{(c)}}^T \widehat{s}\|^2, \qquad l = 1, 2, ..., L_c,$$

where $\|.\|$ denotes L_2 norm. As all signals (and corresponding SCDTs) are discrete, the L_2 norm, which is the norm of functions on the real line, is replaced by the ℓ_2 norm, the norm of sequences, in the calculations. It should be noted that $B_l^{(c)} {B_l^{(c)}}^T$ is the orthogonal projection matrix onto the space generated by the span of the columns of $B_l^{(c)}$ that was computed during the training phase. Subsequently, a set of k closest training samples to the test sample $\widehat{s}$ from class c, denoted by $\{\widehat{s}_1^{(c)}, \cdots, \widehat{s}_k^{(c)}\}$, is found based on the distances $\epsilon_1, \epsilon_2, ...$, etc. Next, $\{\widehat{s}_1^{(c)}, \cdots, \widehat{s}_k^{(c)}\} \cup \mathbb{U}_T \cup \mathbb{U}_H$ is orthogonalized to obtain the basis vectors $\{b_1^{(c)}, b_2^{(c)}, ...\}$ spanning the local subspace from class c with respect to $\widehat{s}$. Let $B^{(c)} = \left[b_1^{(c)}, b_2^{(c)}, ...\right]$ for $c = 1, 2, ...$, etc. The unknown class of s is then estimated by:

$$\arg\min_c \ \|\widehat{s} - B^{(c)} {B^{(c)}}^T \widehat{s}\|^2. \tag{4.21}$$

The presence and severity of damage-induced nonlinearity within the propagation medium can be determined by the predicted class label c. In a detection problem, it indicates the presence of nonlinearity, while in an estimation problem, the label indicates the extent of nonlinearity.

It is important to note that the proposed solution does not require knowledge of the template patterns or the warping functions present in the data. The approach utilizes a set of training samples to search for the nearest local subspace for a given test sample. In the case of dynamical systems where the PDEs are known, we can numerically simulate the PDE solutions and use the simulated data during the training phase. If the system PDE is unknown, previously acquired data from a controlled experimental setup can be utilized as training samples.

4.4 Generative Models for Some Classical PDE Models

This work proposes a generative model-based approach for the PDE system identification problem. The proposed method is applicable to the PDEs where the solution measured at a particular sensor location conforms the generative model defined in eq. (4.16). Meaning, there exists a template pattern and a warping function, which is invertible within the signal's domain, such that the sensor data can be modeled as the instance of the template under the warping function, and the family of the inverses of the warping functions form a convex set. In this section, we investigate several common PDEs with analytical solutions and demonstrate mathematically that such generative models are prevalent in many dynamical systems.

4.4.1 Standard Wave Equation

We begin with the 1D linear wave equation in its standard form, which is expressed as the partial differential equation (PDE):

$$\rho \ddot{u}(x,t) - E u_{xx}(x,t) = 0, \tag{4.22}$$

where ρ denotes the mass density, and E represents Young's modulus. The wave speed, denoted as ν, can be calculated using the formula $\nu = \sqrt{\frac{E}{\rho}}$. Consider the initial condition: $u(x, t = 0) = e^{-x^2}$. The solution to the eq. (4.22) along the positive x axis can be represented as:

$$u(x,t) = e^{-(x-\nu t)^2}. \tag{4.23}$$

From the solution, we can obtain the analytical expression for the sensor measurement $s(t)$ at the specific location $x = x_m$ by deriving it as follows:

$$s(t) := u(x = x_m, t) = e^{-(\nu t - x_m)^2}, \tag{4.24}$$

which can be modeled by the generative model with the following specifications:

$$\begin{aligned} &s(t) = \dot{g}_\nu(t)\varphi_\nu(g_\nu(t)), \quad g_\nu \in \mathcal{G}_\nu \\ &\text{template: } \varphi_\nu(t) = \frac{1}{\nu} e^{-t^2} \\ &\text{time-warping: } g_\nu(t) = \nu t - x_m. \end{aligned} \tag{4.25}$$

The inverse of the warping function can be calculated as follows:

$$g_\nu^{-1}(t) = \frac{1}{\nu}(t + x_m). \tag{4.26}$$

We can represent the set of inverses of the warping functions as $\mathcal{G}_\nu^{-1}$, i.e., $g_\nu^{-1} \in \mathcal{G}_\nu^{-1}$. It can be observed from eq. (4.26) that the set $\mathcal{G}_\nu^{-1}$ is convex with respect to $\frac{1}{\nu}$. Therefore, we can conclude that the solution to the classical linear wave equation, as measured at a particular location, conforms to the proposed generative model.

4.4.2 Diffusion Equation

The proposed generative model can also be applied to the diffusion equation. The 1D diffusion equation is given by,

$$\dot{u}(x,t) - D u_{xx}(x,t) = 0, \tag{4.27}$$

where D represents the diffusion coefficient. Given the initial condition $u(x,0) = \frac{1}{\sqrt{4\pi}} e^{-\frac{x^2}{4}}$, we can obtain the solution to eq. (4.27) as:

$$u(x,t) = \frac{1}{\sqrt{4\pi D t}} e^{-\frac{x^2}{4Dt}}, \tag{4.28}$$

from which, we can derive the expression for the sensor measurement at $x = x_m$, i.e., $s(t) = \frac{1}{\sqrt{4\pi D t}} e^{-\frac{x_m^2}{4Dt}}$. Similar to the wave equation, we can find a template φ_D and a warping function $g_D \in \mathcal{G}_D$, both parameterized

with D, for the diffusion equation so that $s(t)$ can be represented using the generative model given as follows:

$$\begin{aligned} &s(t) = \dot{g}_D(t)\varphi_D(g_D(t)), \quad g_D \in \mathcal{G}_D \\ &\text{template: } \varphi_D(t) = \frac{x_m}{D\sqrt{4\pi t}} e^{-\frac{1}{4t}} \\ &\text{time-warping: } g_D(t) = \frac{Dt}{x_m^2}. \end{aligned} \tag{4.29}$$

The inverse of the warping function is given by $g_D^{-1}(t) = \frac{x_m^2 t}{D}$, which generates the set $\mathcal{G}_D^{-1}$ that is convex with respect to $\frac{1}{D}$. Hence, the sensor measurements for the diffusion equation also adhere to the proposed generative model.

4.4.3 Convection-Diffusion Equation

Given the 1D convection-diffusion equation (PDE) initialized with an initial condition $u(x, 0)$ as:

$$\begin{aligned} &\dot{u}(x,t) = \nu u_x(x,t) + D u_{xx}(x,t), \quad x \in \Omega_x \,, t \in \mathbb{R}_+, \\ &\text{Initial condition: } u(x,0) = \frac{1}{\sqrt{4\pi}} e^{-\frac{x^2}{4}}, \end{aligned} \tag{4.30}$$

where ν and D denote the wave speed and the diffusion coefficient, respectively. Here, we assume that D takes random values from a uniform distribution $U(D_0 - \epsilon, D_0 + \epsilon)$, $D > 0, \forall \epsilon$, and ν is a fixed positive value which needs to be estimated. A solution to the PDE defined in eq. (4.30) can be derived as:

$$u(x,t) = \frac{1}{\sqrt{4\pi D t}} e^{-\frac{(x-\nu t)^2}{4Dt}}. \tag{4.31}$$

The expression for the sensor data $s(t)$ at location $x = x_m$ is given by:

$$s(t) = \frac{1}{\sqrt{4\pi D t}} e^{-\frac{(x_m-\nu t)^2}{4Dt}}, \tag{4.32}$$

which can be represented using the transport generative model as:

$$s(t) = \dot{g}_{\nu,D}(t)\varphi_{\nu,D}(g_{\nu,D}(t)), \quad g_{\nu,D} \in \mathcal{G}_{\nu,D}, \tag{4.33}$$

$$x_m, \nu, D > 0, \quad t > \frac{x_m}{\nu}. \tag{4.34}$$

The template $\varphi_{\nu,D}(t)$ and the warping function $g_{\nu,D}(t)$ are derived as following:

$$\varphi_{\nu,D}(t) = \frac{1}{\nu}\sqrt{\frac{4D}{\pi}}\frac{\left((x_m\nu + 2Dt) + \sqrt{(x_m\nu + 2Dt)^2 - \nu^2 x_m^2}\right)^{\frac{3}{2}}}{\left((x_m\nu + 2Dt) + \sqrt{(x_m\nu + 2Dt)^2 - \nu^2 x_m^2}\right)^2 - \nu^2 x_m^2} e^{-t}\,, \tag{4.35}$$

$$g_{\nu,D}(t) = \frac{(x_m - \nu t)^2}{4Dt}.$$

For a detailed derivation, please refer to the appendix C.1. Since, $g_{\nu,D}(t)$ is a quadratic polynomial function of t, it does not have a unique inverse warping map. However, under the condition that $g^{-1}_{\nu,D}(t)$ is monotonically increasing, we choose the inverse of the warping map to be,

$$g^{-1}_{\nu,D}(t) = \frac{1}{\nu^2}\left(x_m\nu + 2Dt + \sqrt{(x_m\nu + 2Dt)^2 - \nu^2 x_m^2}\right). \tag{4.36}$$

Convexify set of inverses of warping maps: Let, the parameter D is uniformly distributed between $D_0 - \epsilon$ and $D_0 + \epsilon$, i.e., $D \sim U(D_0 - \epsilon, D_0 + \epsilon)$. From eq. (4.36) it is evident that the set $\mathcal{G}^{-1}_{\nu,D}$ formed by the family of functions $g^{-1}_{\nu,D}(t)$ is non-convex with respect to D. We aim to convexify the set by taking 1st order Taylor expansion of the function $g^{-1}_{\nu,D}$ with respect to D at D_0, since we assume that $D \in [D_0 - \epsilon\,, D_0 + \epsilon]$. The resulting expression is given by:

$$\begin{aligned}
g^{-1}_{\nu,D}(t) \approx & g^{-1}_{\nu,D_0}(t) + \frac{\partial g^{-1}_{\nu,D}(t)}{\partial D}\Big|_{D=D_0}(D - D_0) \\
= & \frac{1}{\nu^2}\left(x_m\nu + 2D_0t + \sqrt{(x_m\nu + 2D_0t)^2 - \nu^2x_m^2}\right) \\
& + \frac{2t}{\nu^2}\left(1 + \frac{x_m\nu + 2D_0t}{\sqrt{(x_m\nu + 2D_0t)^2 - \nu^2x_m^2}}\right)(D - D_0) \\
:= & \tilde{g}^{-1}_{\nu,D}(t)\,.
\end{aligned}$$

Since $D \in [D_0 - \epsilon\,, D_0 + \epsilon]$, we evaluate the approximate inverse warping map $\tilde{g}^{-1}_{\nu,D}(t)$ at the boundary values of D, i.e., $D = D_0 \pm \epsilon$ to obtain:

$$\begin{aligned}
& g^{-1}_{\nu,D_0\pm\epsilon}(t) \approx \tilde{g}^{-1}_{\nu,D_0\pm\epsilon}(t) \\
& = \frac{1}{\nu^2}\left(x_m\nu + 2D_0t + \sqrt{(x_m\nu + 2D_0t)^2 - \nu^2x_m^2}\right) \pm \frac{2t}{\nu^2}\left(1 + \frac{x_m\nu + 2D_0t}{\sqrt{(x_m\nu + 2D_0t)^2 - \nu^2x_m^2}}\right)\epsilon \\
& = g^{-1}_{\nu,D_0}(t) \pm \frac{2t}{\nu^2}\left(1 + \frac{x_m\nu + 2D_0t}{\sqrt{(x_m\nu + 2D_0t)^2 - \nu^2x_m^2}}\right)\epsilon.
\end{aligned}$$

The expression for the approximated inverse warping map above suggests that the set $\mathcal{G}_{\nu,D}^{-1}$ can be approximated as convex with respect to the parameter D. Additionally, it can also be shown that the family of inverse warping maps defined in eq. (4.36) can approximately form a convex set with respect to ν (see appendix C.1). Therefore, the proposed generative model-based problem formulation can be utilized to estimate parameter D while allowing ν to take a value from a uniform distribution $U(\nu_0 - \epsilon, \nu_0 + \epsilon)$ or vice versa.

In this section, we have demonstrated that for certain PDEs with analytical solutions, the sensor measurements at a particular location indeed conform the proposed generative model. Consequently, the proposed method is capable of solving the system identification problem for these types of PDEs. However, for PDEs like the one described in equation (4.14), deriving an analytical solution is challenging, making it difficult to verify if the solution satisfies the generative model. Hence, in this work, a data-driven nonlinearity identification approach is employed for such PDE models using the proposed method under the assumption that the sensor measurements follow the generative model defined in eq. (4.16).

4.5 Numerical Experiments and Results

4.5.1 Experimental Setup

Since obtaining an analytical solution to the PDE defined in eq. (4.14) is difficult, we resorted to numerically simulating the PDE solution using the spectral method[136]. For this purpose, we utilized a fast Fourier transform-based implementation[137] of the PDE simulator, which mandates the use of periodic boundary conditions. The initial conditions for the displacement and the velocity are given by:

$$\begin{aligned} u(x,0) &= e^{-\frac{(x-x_0)^2}{2\sigma^2}} \\ \dot{u}(x,0) &= \frac{\nu(x-x_0)}{2\sigma^2} u(x,0), \end{aligned} \tag{4.37}$$

where $\dot{u}(x,0) := \frac{\partial u(x,t)}{\partial t}|_{t=0}$ and ν is the wave speed. The selection of the initial velocity $\dot{u}(x,0)$ was made to produce a pure traveling wave in $+x$ direction for the standard wave equation. A spatial grid of 600 points was used, mean location x_0 and standard deviation σ were set to 50 and 7, separately. The wave speed, which is given by $\nu = \sqrt{\frac{E}{\rho}}$, was computed with the mass density of the medium ρ set at 1, and Young's modulus E chosen from a uniform distribution $U(0.95, 1.05)$. In addition, the parameter F was fixed at 0.01, while M and η were both selected from the distributions $U(0.2, 0.3)$ and $U(0.1, 0.2)$, respectively. In this particular experiment, the velocity $\dot{u}(x,t)$ was considered as the primary quantity of interest and was measured as a

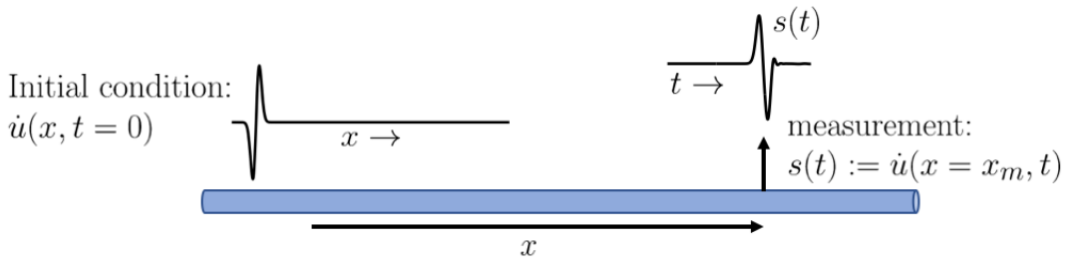


Figure 4.4: Simulation setup for 1D wave propagation through an elastic medium.

function of time at a fixed sensor location $x = 300$, i.e., the sensor measurement $s(t) = \dot{u}(x = 300, t)$. The simulation setup for this experiment is depicted in Fig. 4.4.

Class	β
1	0
2	$U(0.01, 0.6)$

Table 4.2: Binary class classification problem setup for nonlineairty detection. $U(.,.)$ represents a uniform distribution.

Class	$\beta \sim U(l_\beta, h_\beta)$
1	$U(0.01, 0.2)$
2	$U(0.21, 0.4)$
3	$U(0.41, 0.6)$

Table 4.3: Coarse regression (3-class) problem setup for nonlineairty level estimation.

The nonlinearity detection problem was set up as a binary class classification problem, where class 1 denotes signals corresponding to $\beta = 0$ (no nonlinearity) and class 2 consists of signals produced by randomly changing β. Table 4.2 presents the distribution used to select β. To assess the degree of nonlinearity, a coarse regression problem was formulated, in which the proposed method predicts a range of β values for an unknown signal. In the initial experiment, the sensor measurements were grouped into three categories based on their corresponding β values (distributions shown in Table 4.3). Subsequently, this experiment was converted into a relatively finer setup with ten classes, where a narrower range of β values was predicted for an unknown sensor measurement. Table 4.4 displays the ten distributions that correspond to the ten classes. The classification method described in section 4.3 was then utilized to identify the nonlinearity parameter β from the sensor measurements. Fig. 4.5 displays a number of sensor measurements, characterized by the absence ($\beta = 0$) or presence ($\beta > 0$) of nonlinearity. Furthermore, Fig. 4.6 presents a set of example signals featuring three distinct levels of nonlinearity.

Class	$\beta \sim U(l_\beta, h_\beta)$
1	$U(0.01, 0.06)$
2	$U(0.07, 0.12)$
3	$U(0.13, 0.18)$
4	$U(0.19, 0.24)$
5	$U(0.25, 0.30)$
6	$U(0.31, 0.36)$
7	$U(0.37, 0.42)$
8	$U(0.43, 0.48)$
9	$U(0.49, 0.54)$
10	$U(0.55, 0.6)$

Table 4.4: Coarse regression (10-class) problem setup for nonlineairty level estimation.

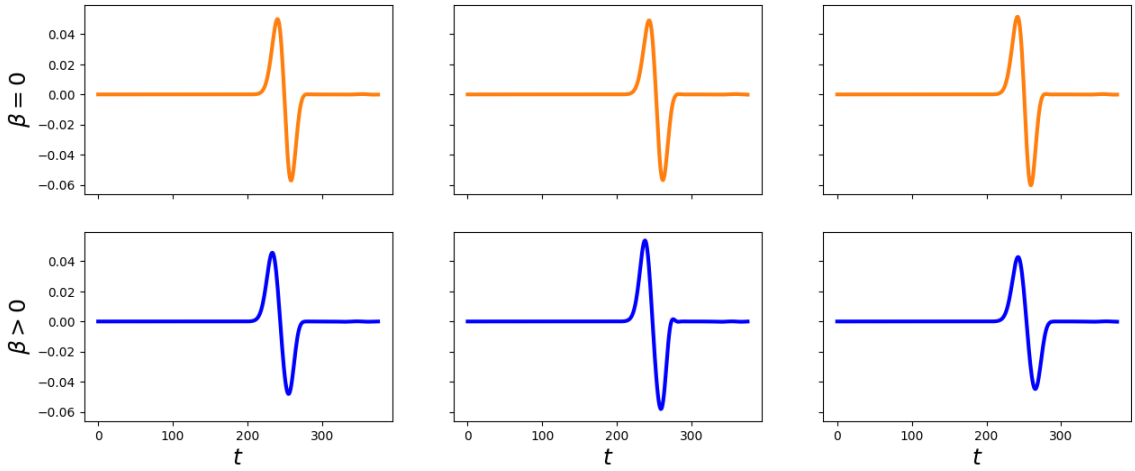


Figure 4.5: Sensor measurements at a particular location (top) without nonlinearity, and (bottom) with nonlinearity.

4.5.2 Evaluation

We conducted a comparative analysis of the proposed method with linear-support vector machine (SVM) [138] classifier and some state-of-the-art deep neural network techniques, including Multilayer Perceptrons (MLP) [107], 1D Visual Geometry Group (VGG) [107], 1D Residual Network (ResNet) [108] [107], and Long Short Term Memory with Fully Convolutional Network (LSTM-FCN) [40] [107]. We also implemented a Fourier transform-based method as a traditional approach, which employs a linear-SVM classifier to classify the sensor measurements in the Fourier domain. Moreover, we compared the performance against SCDT-NS [87], which models each signal class as a single subspace in the SCDT domain.

Detect nonlinearity

The PDE simulation procedure described in section 4.5.1 was executed for 2200 times per class with the randomly varying parameters. During the simulation, the sensor signal was measured at a single location,

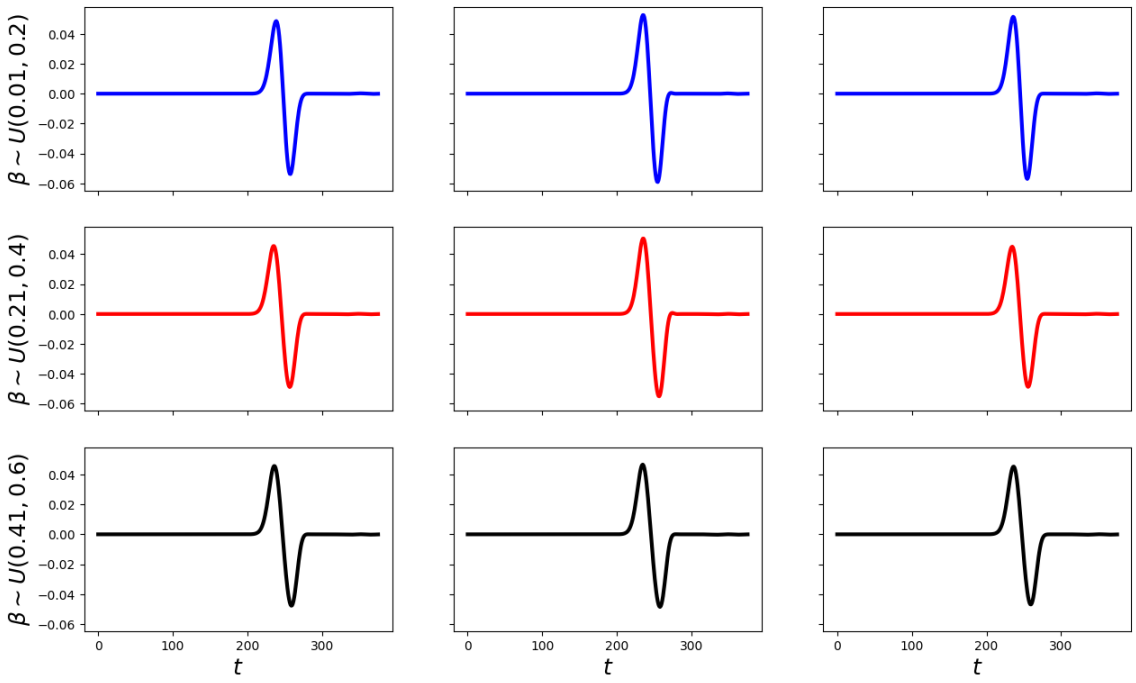


Figure 4.6: Sensor measurements at a particular location under the presence of three different levels of nonlinearity.

resulting in 2200 sensor measurements per class. The collected data was then split into two sets, with 2000 signals per class for training and 200 signals per class for testing. The classification models listed above were trained using the training set, and the testing set was used to evaluate the methods. The performance of detecting nonliearity was reported as classification accuracy. Table 4.5 shows the comparative results of the classification methods. The table shows that the proposed solution achieved a nonlinearity detection accuracy of 98.0%, which is higher than that of the comparative methods. These results suggest that the proposed approach has the potential to detect nonlinearity induced by damage in the propagation medium.

Methods	Accuracy (%)
MLP	73.8
1D-VGG	91.9
1D-ResNet	63.2
LSTM-FCN	83.6
SVM	79.8
FT-SVM	95.5
SCDT-NS	91.5
SCDT-NLS	**98.0**

Table 4.5: Nonlinearity detection accuracy ($\%$) for different classification methods. All the models were trained using 2000 samples per class, and tested on 200 samples per class.

In addition to its effectiveness, the proposed solution is also data-efficient, allowing it to achieve superior detection accuracy using fewer training samples than other methods. In order to demonstrate the data

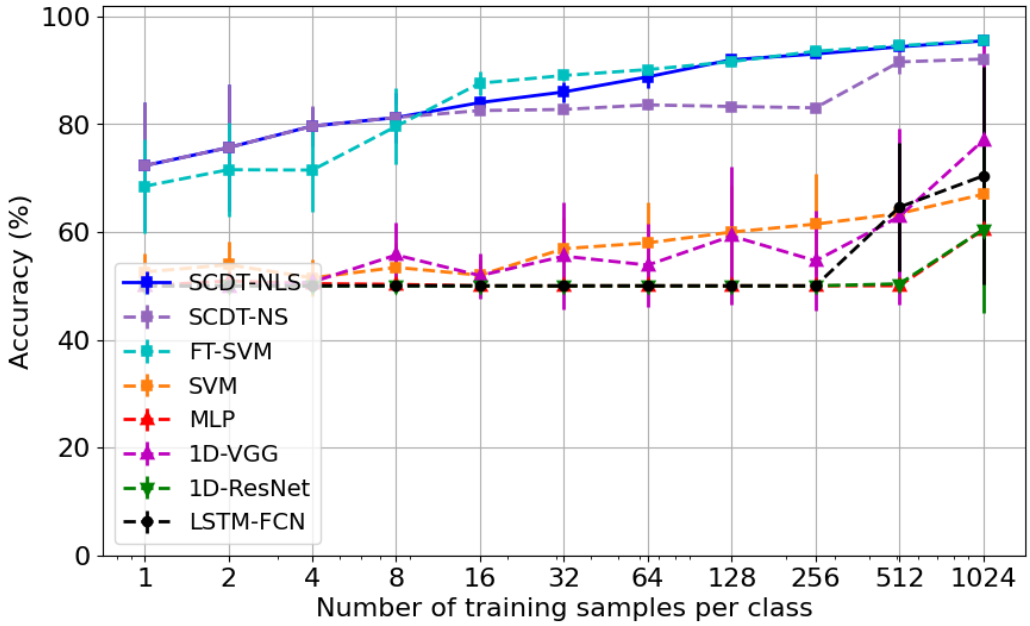


Figure 4.7: Nonlinearity detection accuracy as a function of number of training samples per class for different classification methods.

efficiency of the proposed approach, an experiment was conducted where all classifiers were trained with varying numbers of training samples per class. A training split of a certain size was randomly selected from the original training set, and the experiments for this particular size were repeated 10 times. The results are presented in Figure 4.7, which displays the average detection accuracy as a function of the number of training samples per class for the different classification methods. The error bars indicate the standard deviation for each split. These plots demonstrate that the proposed method outperforms the comparative methods in terms of accuracy, even with a smaller number of training samples.

Estimate degree of nonlinearity

As described in section 4.3.2, the estimation of nonlinearity parameter β is stated as a classification problem, where the proposed approach predicts a range for β. Initially, a three-class classification problem was set up based on the values of β, where three uniform distributions shown in Table 4.3 were used for selecting β for each class. Afterwards, a more detailed regression problem was created, where the sensor measurements were classified into ten different classes, and β-values were selected from ten non-overlapping distributions as shown in Table 4.4. In both scenarios, the classification models were trained using 2000 samples per class, collected by simulating the PDE, and then used to predict β-ranges, i.e., $\widetilde{l}_\beta$ and $\widetilde{h}_\beta$, for 200 test samples per class. To evaluate the performance of the methods, the mean squared errors (MSE) were calculated as

follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\beta_i - \frac{\widetilde{l}_{\beta_i} + \widetilde{h}_{\beta_i}}{2}\right)^2, \tag{4.38}$$

where N represents the total number of test samples, β_i is the true nonlinearity parameter value of the i-th test sample, and $\widetilde{l}_{\beta_i}$ and $\widetilde{h}_{\beta_i}$ are the lower and upper limits of the predicted class, respectively. The performance of the classification methods for estimating β under both regression problem setups is presented in Table 4.6. According to the results, the proposed approach yields the least MSE in both cases, suggesting that it outperforms all other methods in accurately estimating the degree of nonlinearity present in the propagation medium. Fig. 4.8 illustrates the nonlinearity estimation performances of different methods as the number of training samples varies. The plots in the figure highlight that SCDT-NLS outperforms the other methods in estimating the degree of nonlinearity using fewer training data.

Methods	MSE (3-class)	MSE (10-class)
MLP	1.87×10^{-2}	3.12×10^{-2}
1D-VGG	5.10×10^{-3}	1.71×10^{-2}
1D-ResNet	3.72×10^{-2}	3.84×10^{-2}
LSTM-FCN	1.25×10^{-2}	9.77×10^{-3}
SVM	1.99×10^{-2}	2.67×10^{-2}
FT-SVM	3.66×10^{-3}	9.84×10^{-4}
SCDT-NS	4.30×10^{-3}	1.41×10^{-3}
SCDT-NLS	$\mathbf{3.14 \times 10^{-3}}$	$\mathbf{3.08 \times 10^{-4}}$

Table 4.6: Nonlinearity estimation error in MSE (eq. 4.38) for different methods. All the models were trained using 2000 samples per class, and tested on 200 samples per class.

4.6 Discussion

Nonlinearity detection and estimation performances reported in Tables 4.5 and 4.6 suggest that the SCDT-NLS classifier-based approach is a good method for solving nonlinear system identification problems in structural health monitoring applications. In this work, we designed the system identification problem as a classification problem. From the results it is evident that the proposed solution outperforms the state-of-the-art neural network-based time series classification techniques and a Fourier transform-based traditional approach.

The proposed solution utilizes sensor data from a specific location to identify nonlinearity. This approach operates under the assumption that the sensor data conforms to the generative model presented in equation (4.16). Meaning, for a given nonlinear parameter β, the signal measured by a sensor at a particular location can be viewed as an instance of a template observed under a smooth invertible warping function. However, the proposed system identification technique does not require knowledge of the template or the warping function

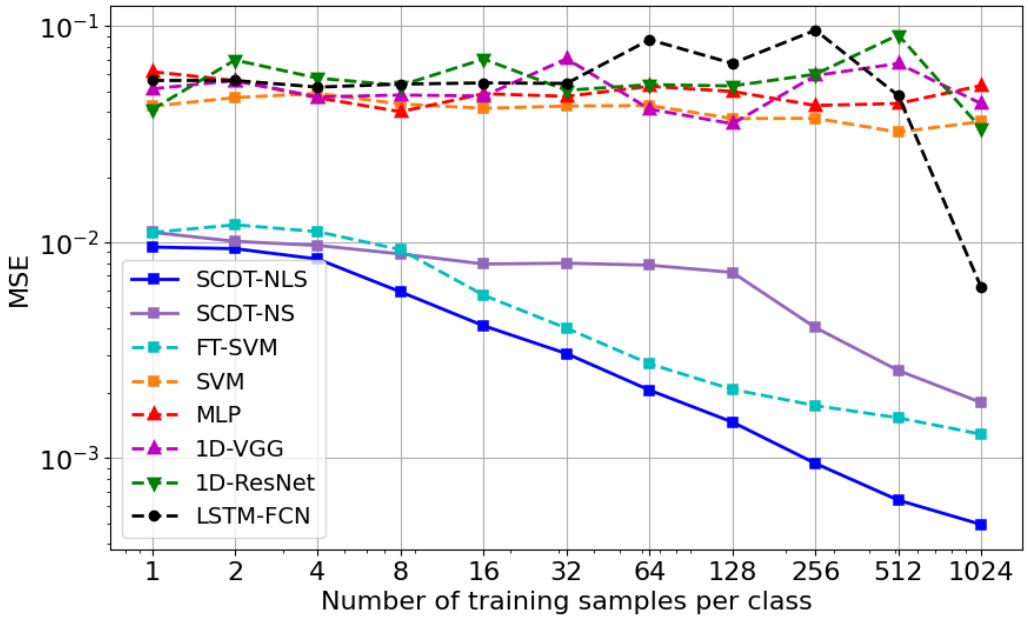


Figure 4.8: Nonlinearity estimation performance (MSE) as a function of number of training samples per class for different methods under coarse regression (10-class) setup.

to recover β. If the generation of an unknown signal satisfies the generative model, the proposed approach can detect and estimate β by searching for the nearest local subspace in the SCDT domain using a set of training samples.

In case of a dynamical system where the PDE model can be analytically formulated, the training samples can be obtained by employing a PDE simulation technique, such as the spectral or finite difference method. Conversely, if the PDE model is unknown, training samples can be generated through a controlled experiment with a known nonlinearity. However, generating a large number of training samples under damage-induced nonlinearity can be challenging in many structural health monitoring applications. As depicted in Fig. 4.8, the proposed approach provides superior nonlinearity estimates with fewer training samples compared to alternative methods. This feature renders the proposed solution suitable for structural health monitoring applications where generating numerous sensor data with known nonlinearity is often impractical.

To summarize, this chapter proposes a novel approach to address nonlinear system identification problems in structural health monitoring applications by leveraging sensor data. Specifically, the signals measured at a specific sensor location are considered as observations of a set of template patterns subjected to some unknown time-warpings. While certain PDE systems, like the classical wave equation or convection-diffusion equation, have analytical formulations for sensor measurement that align with this generative model, others pose difficulties in obtaining such solutions. However, through comprehensive experimentation, this study

shows that the proposed generative model is an effective way to model sensor measurements even in complex PDE systems.

4.7 Conclusion

This chapter explored the feasibility of utilizing a generative model-based approach to solve nonlinear system identification problems in structural health monitoring (SHM) applications. It is presumed that the presence of damage in the structural system results in the presence of nonlinearity in the underlying PDE model. Under the assumption that the sensor data adheres to a specific generative model, we formulated the system identification as a coarse regression problem and employed the SCDT nearest local subspace classifier to detect and estimate the nonlinearity parameter of the dynamical system. Extensive experiments were conducted, and the proposed solution was found to provide a significantly better estimate of the nonlinearity when compared to state-of-the-art data-driven pattern recognition methods.

Acknowledgements

This work is done in collaboration with Dr. Duy Hoang Thai, Dr. Jonathan M. Nichols, and Dr. Gustavo K. Rohde, to whom I express my deep gratitude for their contributions to this chapter.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The tasks of solving signal estimation and classification problems lie at the core of a wide range of signal processing applications. Most existing methods fail to model the underlying nonlinearities of the systems, resulting in suboptimal solutions. This thesis introduces a new approach to address the problems of signal estimation and classification. The approach involves using a generative model that connects the signal data to the physical processes that produce it. The generative model represents the signal data such that the signal can be seen as an observation of a template under some temporal deformations. Next, the estimation and classification problems are formulated for the family of signals that follow the proposed generative model. The proposed solutions to these problems utilize transport-based signal transformation techniques such as the cumulative distribution transform (CDT) and the signed cumulative distribution transform (SCDT) to solve the problems effectively and efficiently.

Chapter 2 of this thesis introduced a novel approach to estimate the parametric change of a signal by minimizing the Wasserstein distance between the measured and model signals. The proposed approach, aided by the use of the cumulative distribution transform, was demonstrated to produce a generic closed-form solution to the estimation problem. We also discussed the estimator's properties in the presence of noise and presented a novel technique to mitigate the impact of noise on the estimates. Through numerical experiments, we showed that our proposed approach performs well in comparison to existing methods and is significantly more computationally efficient. We further applied our approach to a source localization problem where a crack was localized on a metal structure. In summary, our method provides a robust and accurate means

to estimate parameters governing signal energy modification during propagation by leveraging the CDT and Wasserstein cost.

In Chapter 3, we presented a novel end-to-end method for classifying segmented signals using the signed cumulative distribution transform (SCDT), a newly developed signal transformation technique. We first formulated a multiple template-based problem statement for the signal classes, where the signals can be represented as observations of a set of template patterns under some unknown deformations. Our proposed solution to this problem utilized a nearest local subspace search technique in the SCDT domain to detect the class of an unknown signal. We conducted extensive experiments to demonstrate that our method outperforms state-of-the-art end-to-end signal classification methods in classifying segmented signal data. Moreover, we showed that our method is data-efficient, computationally inexpensive, and robust to out-of-distribution examples. In short, our proposed method provides an accurate and efficient means to classify segmented signals when the signal classes follow a particular generative model.

In Chapter 4, we applied the generative model-based problem formulation approach introduced earlier to develop a novel system identification technique for structural health monitoring (SHM) applications. In many SHM scenarios, it is believed that the presence of damage in the structural system leads to nonlinearity in the underlying PDE model of the dynamical system. Assuming that the sensor data conforms to a specific generative model, we formulated the task of system identification as a coarse regression problem and utilized the SCDT nearest local subspace classifier to detect and estimate the nonlinearity parameter. Initially, we demonstrated analytically that the proposed generative model is prevalent in many classical PDE systems where analytical solutions can be obtained. However, in the case of complex PDE systems where deriving analytical solutions is difficult, we conducted extensive experiments and showed that our proposed solution provides a significantly more accurate estimate of the nonlinearity compared to state-of-the-art data-driven pattern recognition methods.

To summarize, this thesis put forth novel mathematical modeling-based approaches to solve certain estimation and classification problems for segmented signals that follow a particular generative model. The proposed solutions have potential applications in various signal processing domains. This thesis primarily focused on structural health monitoring (SHM) applications and showcased the effectiveness of the proposed solutions in solving source localization and system identification problems.

5.2 Future Work

In Chapter 2, we studied the properties and performances of the CDT-based estimator in the presence of noise. To gain a deeper understanding of the statistical properties of the CDT and develop a maximum

likelihood estimator, further efforts will be directed towards statistical analysis. Similar analysis will be conducted for the SCDT-based classification technique (proposed in Chapter 3) as well. Future work will further include exploring ways to extend the classification method to classify time series events that are not readily segmentable. Chapter 4 introduced a novel system identification technique for identifying damage-induced nonlinearity in dynamical systems. In future work, this technique will be applied to experiments aimed at identifying properties of the system (propagation medium).

References

[1] Vishnuvardhan Janapati, Fotis Kopsaftopoulos, Frank Li, Sang Jun Lee, and Fu-Kuo Chang. Damage detection sensitivity characterization of acousto-ultrasound-based structural health monitoring techniques. *Structural Health Monitoring*, 15(2):143–161, 2016.

[2] R. N. McDonough and A. D. Whalen. *Detection of Signals in Noise*. Academic Press, San Diego, second edition, 1995.

[3] J. C. Hassab. On the weighting of time delay measurements for improved direction finding. *Journal of the Acoustical Society of America*, 78(5):1658–1663, 1985.

[4] B. Engquist, B. D. Froese, and Y. Yang. Optimal transport for seismic full waveform inversion. *arXiv preprint arXiv:1602.01540*, 2016.

[5] J. Chen, Y. Chen, H. Wu, and D. Yang. The quadratic wasserstein metric for earthquake location. *Journal of Computational Physics*, 373:188–209, 2018.

[6] J. G. Proakis. Digital communications, 1983.

[7] Alon Amar, Geert Leus, and Benjamin Friedlander. Emitter localization given time delay and frequency shift measurements. *IEEE Transactions on Aerospace and Electronic Systems*, 48(2):1826–1837, 2012.

[8] J. M. Nichols, M. N. Hutchinson, N. Menkart, G. A. Cranch, and G. K. Rohde. Time delay estimation via Wasserstein distance minimization. *Signal Processing Letters*, 26(6):908–912, 2019.

[9] O. D. Lara and M. A. Labrador. A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials*, 15(3):1192–1209, 2012.

[10] S. K. Berkaya, A. K. Uysal, E. S. Gunal, S. Ergin, Serkan G., and M. B. Gulmezoglu. A survey on ECG analysis. *Biomedical Signal Processing and Control*, 43:216–235, 2018.

[11] A. Subasi and M. I. Gursoy. EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert systems with applications*, 37(12):8659–8666, 2010.

[12] A. Fehske, J. Gaeddert, and Jeffrey H. Reed. A new approach to signal classification using spectral correlation and neural networks. In *First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, 2005. DySPAN 2005.*, pages 144–150. IEEE, 2005.

[13] Osama Abdeljaber, Onur Avci, Mustafa Serkan Kiranyaz, Boualem Boashash, Henry Sodano, and Daniel J Inman. 1-D CNNs for structural damage detection: Verification on a structural health monitoring benchmark data. *Neurocomputing*, 275:1308–1317, 2018.

[14] R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao. Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, 115:213–237, 2019.

[15] C. Lu, T. Lee, and C. Chiu. Financial time series forecasting using independent component analysis and support vector regression. *Decision support systems*, 47(2):115–125, 2009.

[16] G. Jacovitti and G. Scarano. Discrete time techniques for time delay estimation. *IEEE Transactions on Signal Processing*, 41(2):525–533, 1993.

[17] X. Jiang, W. J. Zeng, H. C. So, S. Rajan, and T. Kirubarajan. Robust matched filtering in ℓ_p-space. *IEEE Transactions on Signal Processing*, 63(23):6184–6199, 2015.

[18] J. Chen, J. Benesty, and Y. Huang. Performance of GCC- and AMDF-based time-delay estimation in practical reverberant environments. *EURASIP Journal on Applied Signal Processing*, 1:25–36, 2005.

[19] J. Benesty, Y. Huang, and J. Chen. Time delay estimation via minimum entropy. *IEEE Signal Processing Letters*, 14(3):157–160, 2007.

[20] G. K. Rohde, F. Bucholtz, and J. M. Nichols. Maximum empirical likelihood estimation of time delay in independently and identically distributed noise. *IET Signal Processing*, 8(7):720–728, 2014.

[21] N. J. I. Mars and G. W. Van Arragon. Time delay estimation in nonlinear systems. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):619–621, 1981.

[22] Q. Jin, K. M. Wong, and Z. Q. Luo. The estimation of time delay and doppler stretch of wideband signals. *IEEE Transactions on Signal Processing*, 43(4):904–916, 1995.

[23] X. X. Niu, P. C. Ching, and Y. T. Chan. Wavelet based approach for joint time delay and doppler stretch measurements. *IEEE Transactions on Aerospace and Electronic Systems*, 35(3):1111–1119, 1999.

[24] R. Tao, W. Q. Zhang, and E. Q. Chen. Two-stage method for joint time delay and doppler shift estimation. *IET Radar, Sonar & Navigation*, 2(1):71–77, 2008.

[25] Andreas Jakobsson, A Lee Swindlehurst, and Petre Stoica. Subspace-based estimation of time delays and Doppler shifts. *IEEE Transactions on Signal Processing*, 46(9):2472–2483, 1998.

[26] Benjamin Friedlander. High resolution Doppler and delay estimation. In *2013 Asilomar Conference on Signals, Systems and Computers*, pages 181–185. IEEE, 2013.

[27] Jian-Feng Gu, Jaber Moghaddasi, and Ke Wu. Delay and Doppler shift estimation for OFDM-based radar-radio (RadCom) system. In *2015 IEEE international wireless symposium (IWS 2015)*, pages 1–4. IEEE, 2015.

[28] Michail K Tsatsanis and Georgios B Giannakis. Subspace methods for blind estimation of time-varying fir channels. *IEEE Transactions on Signal Processing*, 45(12):3084–3093, 1997.

[29] Alex Nanopoulos, Rob Alcock, and Yannis Manolopoulos. Feature-based classification of time-series data. *International Journal of Computer Research*, 10(3):49–61, 2001.

[30] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data mining and knowledge discovery*, 31(3):606–660, 2017.

[31] Ben D Fulcher and Nick S Jones. Highly comparative feature-based time-series classification. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):3026–3037, 2014.

[32] Amaia Abanda, Usue Mori, and Jose A Lozano. A review on distance based time series classification. *Data Mining and Knowledge Discovery*, 33(2):378–412, 2019.

[33] Jason Lines and Anthony Bagnall. Time series classification with ensembles of elastic distance measures. *Data Mining and Knowledge Discovery*, 29(3):565–592, 2015.

[34] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:, 1994.

[35] Xiaopeng Xi, Eamonn Keogh, Christian Shelton, Li Wei, and Chotirat Ann Ratanamahatana. Fast time series classification using numerosity reduction. In *Proceedings of the 23rd international conference on Machine learning*, pages 1033–1040, 2006.

[36] Yi Zheng, Qi Liu, Enhong Chen, Yong Ge, and J Leon Zhao. Time series classification using multi-channels deep convolutional neural networks. In *International conference on web-age information management*, pages 298–310. Springer, 2014.

[37] Bendong Zhao, Huanzhang Lu, Shangfeng Chen, Junliang Liu, and Dongya Wu. Convolutional neural networks for time series classification. *Journal of Systems Engineering and Electronics*, 28(1):162–169, 2017.

[38] Zhiguang Wang, Weizhong Yan, and Tim Oates. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, pages 1578–1585. IEEE, 2017.

[39] Zhicheng Cui, Wenlin Chen, and Yixin Chen. Multi-scale convolutional neural networks for time series classification. *arXiv preprint arXiv:1603.06995*, 2016.

[40] Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Shun Chen. LSTM fully convolutional networks for time series classification. *IEEE access*, 6:1662–1669, 2017.

[41] Anders Rytter. Vibrational based inspection of civil engineering structures. 1993.

[42] Ingolf Müller, Vishnuvardhan Janapati, Sourav Banerjee, K Lonkar, Surajit Roy, and F-K Chang. On the performance quantification of active sensing shm systems using model-assisted pod methods. *Structural Health Monitoring 2011*, 2011.

[43] J. M. Nichols and K. D. Murphy. *Modeling and Estimation of Structural Damage*. Wiley & Sons, West Sussex, UK, 2016.

[44] James P Crutchfield and BS McNamara. Equations of motion from a data series. *Complex systems*, 1:417–452, 1987.

[45] Jer-Nan Juang and Richard S Pappa. An eigensystem realization algorithm for modal parameter identification and model reduction. *Journal of guidance, control, and dynamics*, 8(5):620–627, 1985.

[46] Hayden Schaeffer. Learning partial differential equations via data discovery and sparse optimization. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2197):20160446, 2017.

[47] Jonathan H Tu. *Dynamic mode decomposition: Theory and applications*. PhD thesis, Princeton University, 2013.

[48] Jean-Philippe Noël and Gaëtan Kerschen. Nonlinear system identification in structural dynamics: 10 more years of progress. *Mechanical Systems and Signal Processing*, 83:2–35, 2017.

[49] Sheng Chen and Steve A Billings. Representations of non-linear systems: the narmax model. *International journal of control*, 49(3):1013–1032, 1989.

[50] Raul González-García, Ramiro Rico-Martìnez, and Ioannis G Kevrekidis. Identification of distributed parameter systems: A neural net based approach. *Computers & chemical engineering*, 22:S965–S968, 1998.

[51] Naoya Takeishi, Yoshinobu Kawahara, and Takehisa Yairi. Learning koopman invariant subspaces for dynamic mode decomposition. *Advances in Neural Information Processing Systems*, 30, 2017.

[52] Christoph Wehmeyer and Frank Noé. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *The Journal of chemical physics*, 148(24):241703, 2018.

[53] Enoch Yeung, Soumya Kundu, and Nathan Hodas. Learning deep neural network representations for koopman operators of nonlinear dynamical systems. In *2019 American Control Conference (ACC)*, pages 4832–4839. IEEE, 2019.

[54] Ioannis G Kevrekidis, C William Gear, James M Hyman, Panagiotis G Kevrekidis, Olof Runborg, Constantinos Theodoropoulos, et al. Equation-free, coarse-grained multiscale computation: enabling microscopic simulators to perform system-level analysis. *Commun. Math. Sci*, 1(4):715–762, 2003.

[55] Ioannis G Kevrekidis and Giovanni Samaey. Equation-free multiscale computation: Algorithms and applications. *Annual review of physical chemistry*, 60:321–344, 2009.

[56] Dimitrios Giannakis and Andrew J Majda. Nonlinear laplacian spectral analysis for time series with intermittency and low-frequency variability. *Proceedings of the National Academy of Sciences*, 109(7):2222–2227, 2012.

[57] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

[58] Shizhe Chen, Ali Shojaie, and Daniela M Witten. Network reconstruction from high-dimensional ordinary differential equations. *Journal of the American Statistical Association*, 112(520):1697–1707, 2017.

[59] Samuel H Rudy, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Data-driven discovery of partial differential equations. *Science advances*, 3(4):e1602614, 2017.

[60] Samuel Rudy, Alessandro Alla, Steven L Brunton, and J Nathan Kutz. Data-driven identification of parametric partial differential equations. *SIAM Journal on Applied Dynamical Systems*, 18(2):643–660, 2019.

[61] S. R. Park, S. Kolouri, S. Kundu, and G. K. Rohde. The cumulative distribution transform and linear pattern recognition. *Applied and computational harmonic analysis*, 45:616–641, 2018.

[62] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

[63] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84:375–393, 2000.

[64] J. M. Nichols, T. H. Emerson, and G. K. Rohde. A transport model for broadening of a linearly polarized, coherent beam due to inhomogeneities in a turbulent atmosphere. *Journal of Modern Optics*, 66(8):835–849, 2019.

[65] L. Gosse and O. Runborg. Finite moment problems and applications to multiphase computations in geometric optics. *Communications in Mathematical Sciences*, 3(3):373–392, 2005.

[66] J. Rubinstein and G. Wolansky. Geometrical optics and optimal transport. *Journal of the Optical Society of America – A*, 32(10):1817–1823, 2017.

[67] J. M. Nichols, T. H. Emerson, L. Cattell, S. Park, A. Kanaev, F. Bucholtz, A. Watnik, T. Doster, and G. K. Rohde. A transport-based model for turbulence-corrupted imagery. *Applied Optics*, 57(16):4524–4536, 2018.

[68] J. D. Achenbach. *Wave propagation in elastic solids*. Applied Mathematics and Mechanics Series. North-Holland Publishing Company, Amsterdam, 1973.

[69] Björn Engquist and Yunan Yang. Seismic inversion and the data normalization for optimal transport. *arXiv preprint arXiv:1810.08686*, 2018.

[70] Stefania Colonnese, Stefano Rinauro, and Gaetano Scarano. Generalized method of moments estimation of location parameters: Application to blind phase acquisition. *IEEE Transactions on Signal Processing*, 58(9):4735–4749, 2010.

[71] C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[72] B. Engquist and B. D. Froese. Application of the wasserstein metric to seismic signals. *arXiv preprint arXiv:1311.4581*, 2013.

[73] Y. Yang, B. Engquist, J. Sun, and B. F. Hamfeldt. Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion. *Geophysics*, 83(1):R43–R62, 2018.

[74] Aboulnasr Hassanien, Sergiy A Vorobyov, and Alex B Gershman. Moving target parameters estimation in noncoherent mimo radar systems. *IEEE Transactions on Signal Processing*, 60(5):2354–2361, 2012.

[75] JM Nichols, M Currie, F Bucholtz, and WA Link. Bayesian estimation of weak material dispersion: theory and experiment. *Optics express*, 18(3):2076–2089, 2010.

[76] Vivek Ashok Bohara and See Ho Ting. Theoretical analysis of OFDM signals in nonlinear polynomial models. In *2007 6th International Conference on Information, Communications & Signal Processing*, pages 1–5. IEEE, 2007.

[77] Wikipedia. Stone–Weierstrass theorem. [Accessed: 2020-03-22].

[78] S. M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, New Jersey, 1993.

[79] H. V. Trees, K. Bell, and Z. Tian. Detection, estimation, and modulation theory, part I, 1968.

[80] Wikipedia. Quantile function. [Accessed: 2020-03-22].

[81] Warren Gilchrist. *Statistical modelling with quantile functions*. CRC Press, 2000.

[82] S. R. Bowling, M. T. Khasawneh, S. Kaewkuekool, and B. R. Cho. A logistic approximation to the cumulative normal distribution. *Journal of Industrial Engineering and Management*, 2(1):114–127, 2009.

[83] *Optimization Toolbox, MATLAB version 9.4 (R2018a)*. The MathWorks Inc., Natick, Massachusetts, 2018.

[84] J. D. Hart, M. N. Hutchinson, C. R. S. Williams, J. M. Nichols, G. A. Cranch, A. H. M. Rubaiyat, and G. K. Rohde. Fiber laser acoustic emission sensing based crack detection via cumulative distribution transform. In *Submitted to 2020 Optical Networking and Communication Conference*.

[85] D. G. Aggelis and T. E. Matikas. Effect of plate wave dispersion on the acoustic emission parameters in metals. *Computers and Structures*, 98:17–22, 2012.

[86] Fredrik Gustafsson and Fredrik Gunnarsson. Positioning using time-difference of arrival measurements. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 6, pages VI–553. IEEE, 2003.

[87] Abu Hasnat Mohammad Rubaiyat, Mohammad Shifat-E-Rabbi, Yan Zhuang, Shiying Li, and Gustavo K Rohde. Nearest subspace search in the signed cumulative distribution transform space for 1d signal classification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3508–3512. IEEE, 2022.

[88] A. Aldroubi, R. D. Martin, I. Medri, G. K. Rohde, and S. Thareja. The signed cumulative distribution transform for 1-D signal analysis and classification. *Foundations of Data Science*, 2022.

[89] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4):917–963, 2019.

[90] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119:3–11, 2019.

[91] Mohamad Mahmoud Al Rahhal, Yakoub Bazi, Haikel AlHichri, Naif Alajlan, Farid Melgani, and Ronald R Yager. Deep learning approach for active classification of electrocardiogram signals. *Information Sciences*, 345:340–354, 2016.

[92] Mustafa Gokce Baydogan, George Runger, and Eugene Tuv. A bag-of-features framework to classify time series. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2796–2802, 2013.

[93] Anthony Bagnall, Jason Lines, Jon Hills, and Aaron Bostrom. Time-series classification with COTE: the collective of transformation-based ensembles. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2522–2535, 2015.

[94] Jason Lines, Sarah Taylor, and Anthony Bagnall. Time series classification with HIVE-COTE: The hierarchical vote collective of transformation-based ensembles. *ACM Transactions on Knowledge Discovery from Data*, 12(5), 2018.

[95] Houtao Deng, George Runger, Eugene Tuv, and Martyanov Vladimir. A time series forest for classification and feature extraction. *Information Sciences*, 239:142–153, 2013.

[96] S. R. Park, S. Kolouri, S. Kundu, and G. K. Rohde. The cumulative distribution transform and linear pattern recognition. *Applied and computational harmonic analysis*, 45:616–641, 2018.

[97] Wei Wang, Dejan Slepčev, Saurav Basu, John A Ozolek, and Gustavo K Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101(2):254–269, 2013.

[98] Shiying Li, Abu Hasnat Mohammad Rubaiyat, and Gustavo K. Rohde. Geodesic properties of a generalized wasserstein embedding for time series analysis. In *Proceedings of Topological, Algebraic, and Geometric Learning Workshops 2022*, volume 196 of *Proceedings of Machine Learning Research*, pages 216–225. PMLR, 25 Feb–22 Jul 2022.

[99] A. H. M. Rubaiyat, K. M. Hallam, J. M. Nichols, M. N. Hutchinson, S. Li, and G. K. Rohde. Parametric signal estimation using the cumulative distribution transform. *IEEE Transactions on Signal Processing*, 68:3312–3324, 2020.

[100] H. L. Royden and P. Fitzpatrick. *Real analysis*, volume 32. Macmillan New York, 1988.

[101] M. Shifat-E-Rabbi, X. Yin, A. H. M. Rubaiyat, S. Li, S. Kolouri, A. Aldroubi, J. M. Nichols, and G. K. Rohde. Radon cumulative distribution transform subspace modeling for image classification. *Journal of Mathematical Imaging and Vision*, pages 1–19, 2021.

[102] Jorma Laaksonen. Local subspace classifier. In *International Conference on Artificial Neural Networks*, pages 637–642. Springer, 1997.

[103] Hakan Cevikalp, Diane Larlus, Matthijs Douze, and Frédéric Jurie. Local subspace classifiers: Linear and nonlinear approaches. In *2007 IEEE Workshop on Machine Learning for Signal Processing*, pages 57–62. IEEE, 2007.

[104] Mohammad Shifat E Rabbi, Yan Zhuang, Shiying Li, Abu Hasnat Mohammad Rubaiyat, Xuwang Yin, and Gustavo K Rohde. Invariance encoding in sliced-Wasserstein space for image classification with limited training data. *arXiv preprint arXiv:2201.02980*, 2022.

[105] Victor M Panaretos and Yoav Zemel. *An invitation to statistics in Wasserstein space*. Springer Nature, 2020.

[106] Michael Holmes, Charles Lee, and Alexander Gray. QUIC-SVD: Fast SVD using cosine trees. *Advances in Neural Information Processing Systems*, 21, 2008.

[107] Brian Kenji Iwana and Seiichi Uchida. An empirical survey of data augmentation for time series classification with neural networks. *Plos one*, 16(7):e0254841, 2021.

[108] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Data augmentation using synthetic data for time series classification with deep residual networks. *arXiv preprint arXiv:1808.02455*, 2018.

[109] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[110] Hoang Anh Dau, Eamonn Keogh, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, Yanping, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, Gustavo Batista, and Hexagon-ML. The UCR time series classification archive, October 2018. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.

[111] Antigoni Mezari and Ilias Maglogiannis. Gesture recognition using symbolic aggregate approximation and dynamic time warping on motion data. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 342–347, 2017.

[112] Denis S Willett, Justin George, Nora S Willett, Lukasz L Stelinski, and Stephen L Lapointe. Machine learning for characterization of insect vector feeding. *PLoS computational biology*, 12(11):e1005158, 2016.

[113] Jingkun Gao, Suman Giri, Emre Can Kara, and Mario Bergés. PLAID: a public dataset of high-resoultion electrical appliance measurements for load identification research: demo abstract. In *proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, pages 198–199, 2014.

[114] Jiayang Liu, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. uWave: Accelerometer-based personalized gesture recognition and its applications. *Pervasive and Mobile Computing*, 5(6):657–675, 2009.

[115] Robert Thomas Olszewski. *Generalized feature extraction for structural pattern recognition in time-series data*. Carnegie Mellon University, 2001.

[116] Umaa Rebbapragada, Pavlos Protopapas, Carla E Brodley, and Charles Alcock. Finding anomalous periodic time series. *Machine learning*, 74(3):281–313, 2009.

[117] Pierre Geurts. *Contributions to decision tree induction: bias/variance tradeoff and time series classification*. PhD thesis, ULiège-University of Liège, 2002.

[118] P. Pławiak. ECG signals (744 fragments), 2017.

[119] P. Pławiak. Novel genetic ensembles of classifiers applied to myocardium dysfunction recognition based on ECG signals. *Swarm and evolutionary computation*, 39:192–208, 2018.

[120] Sajad Mousavi and Fatemeh Afghah. Inter-and intra-patient ECG heartbeat classification for arrhythmia detection: a sequence to sequence deep learning approach. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1308–1312. IEEE, 2019.

[121] M. M. Bassiouni, E. A. El-Dahshan, W. Khalefa, and A. M. Salem. Intelligent hybrid approaches for human ECG signals identification. *Signal, Image and Video Processing*, 12(5):941–949, 2018.

[122] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[123] R Paul Gorman and Terrence J Sejnowski. Analysis of hidden units in a layered network trained to classify sonar targets. *Neural networks*, 1(1):75–89, 1988.

[124] National Renewable Energy Laboratory. Gearbox fault diagnosis data [data set], 2018. http://data.openei.org/submissions/623.

[125] Ronald Newbold Bracewell and Ronald N Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-hill New York, 1986.

[126] Alfredo Güemes, Antonio Fernandez-Lopez, Angel Renato Pozo, and Julián Sierra-Pérez. Structural health monitoring for advanced composite structures: a review. *Journal of Composites Science*, 4(1):13, 2020.

[127] K. Worden, C. R. Farrar, G. Manson, and G. Park. The fundamental axioms of structural health monitoring. *Proceedings of the Royal Society of London A*, 463:1639–1664, 2007.

[128] C. R. Farrar and K. Worden. *Structural Health Monitoring: A Machine Learning Perspective*. Wiley, New York, 2012.

[129] H. Sohn, K. Worden, and C. R. Farrar. Statistical damage classification under changing environmental and operational conditions. *Journal of Intelligent Material Systems and Structures*, 13:561–574, 2002.

[130] E. Z. Moore, J. M. Nichols, and K. D. Murphy. Model-based SHM: Demonstration of identification of a crack in a thin plate using free vibration data. *Mechanical Systems and Signal Processing*, 29:284–295, 2012.

[131] C. R. Farrar, K. Worden, M. D. Todd, G. Park, J. M. Nichols, D. E. Adams, M. T. Bement, and K. Farinholt. Nonlinear system identification for damage detection. Technical Report LA-14353-MS, Los Alamos National Laboratory, 2007.

[132] J. M. Nichols, S. T. Trickey, M. Seaver, S. R. Motley, and E. D. Eisner. Using ambient vibrations to detect loosening of a composite-to-metal bolted joint in the presence of strong temperature fluctuations. *Journal of Vibration and Acoustics*, 129:710–717, 2007.

[133] Abu Hasnat Mohammad Rubaiyat, Shiying Li, Xuwang Yin, Mohammad Shifat E Rabbi, Yan Zhuang, and Gustavo K Rohde. End-to-end signal classification in signed cumulative distribution transform space. *arXiv preprint arXiv:2205.00348*, 2022.

[134] Se Rim Park, Soheil Kolouri, Shinjini Kundu, and Gustavo K Rohde. The cumulative distribution transform and linear pattern classification. *Applied and computational harmonic analysis*, 45(3):616–641, 2018.

[135] Abu Hasnat Mohammad Rubaiyat, Kyla M Hallam, Jonathan M Nichols, Meredith N Hutchinson, Shiying Li, and Gustavo K Rohde. Parametric signal estimation using the cumulative distribution transform. *IEEE Transactions on Signal Processing*, 68:3312–3324, 2020.

[136] David A Kopriva. *Implementing spectral methods for partial differential equations: Algorithms for scientists and engineers*. Springer Science & Business Media, 2009.

[137] Steven L Brunton and J Nathan Kutz. *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press, 2022.

[138] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[139] F. Frishman. *On the Arithmetic Means and Variances of Products and Ratios of Random Variables*. Defense Technical Information Center, ARMY RESEARCH OFFICE RESEARCH TRIANGLE PARK NC., 1971.

[140] S. Y. Chun and A. Shapiro. Normal versus noncentral chi-square asymptotics of misspecified models. *Multivariate behavioral research*, 44 6:803–27, 2009.

Appendix A

Supplementary Materials for Chapter 2

A.1 Noise correction in the CDF space

Let the noisy signal, $z_\eta(t) = z_g(t) + \eta(t)$, where $\eta(t) \sim \mathcal{N}(0, \sigma^2)$. The associated positive pdf,

$$\begin{aligned} r(t) = B(z_\eta)(t) &= \frac{z_n^2(t)}{\|z_\eta\|_{\ell_2}^2},\ t = t_1, \cdots, t_N \\ &= \frac{\{z_g(t) + \eta(t)\}^2}{\mathcal{E}_{z_\eta}} \end{aligned} \tag{A.1}$$

where $\mathcal{E}_{z_\eta} \equiv \|z_\eta\|_{\ell_2}^2$ is the total energy of the received (noisy) signal. In terms of the CDF we may expand (A.1) and write

$$\begin{aligned} R(t) &= \int_{t_1}^{t} \frac{\{z_g(u) + \eta(u)\}^2}{\mathcal{E}_{z_\eta}} du \\ \Rightarrow R(t)\mathcal{E}_{z_\eta} &= \int_{t_1}^{t} \{z_g^2(u) + \eta^2(u) + 2z_g(u)\eta(u)\} du \end{aligned} \tag{A.2}$$

Taking the expected value over different realizations of the noise, recognizing $E[\eta^2] \equiv \sigma^2$, one obtains

$$\begin{aligned} E[R(t)\mathcal{E}_{z_\eta}] &= \int_{t_1}^{t} z_g^2(u) du + \int_{t_1}^{t} \sigma^2 du \\ \Rightarrow \mathcal{E}_{z_\eta} E[R(t)] &= S_g(t)\mathcal{E}_z + \sigma^2(t - t_1) \end{aligned} \tag{A.3}$$

where $S_g(t) = \int_{t_1}^{t} \frac{z_g^2(u)}{\mathcal{E}_z} du$ is the CDF of noise free signal, which we seek to estimate ($\mathcal{E}_z = \|z_g\|_{\ell_2}^2$ is the energy of the noise free signal).

In Eqn. (A.3) we have treated the noisy signal energy $\mathcal{E}_{z_\eta}$ as constant and equal to its mean, i.e. $\mathcal{E}_{z_\eta} = E[\mathcal{E}_{z_\eta}] = \mathcal{E}_z + \sigma^2(t_N - t_1)$. This approximation is valid as long as the fluctuations in the energy of the received signal are small with respect to $E[\mathcal{E}_{z_\eta}]$ [139]. This is often the case, especially for applications where the number of digital samples in the signals being processed is large, as the assumption simply says that fluctuations in the total signal mass among different realizations will by typically small relative to the total signal mass. As an example Fig. A.1 plots both $\mathcal{E}_z$ and $\mathcal{E}_{z_\eta}$ for 100 realizations of noise with SNR 5, 10, and 15 dB. As expected the fluctuations are very small with respect to $E[\mathcal{E}_{z_\eta}]$.

Replacing $\mathcal{E}_{z_\eta}$ with its mean, we can re-arrange Eqn. (A.3) to yield

$$S_g(t) = \frac{E[R(t)]\{\mathcal{E}_z + \sigma^2(t_N - t_1)\} - \sigma^2(t - t_1)}{\mathcal{E}_z} \tag{A.4}$$

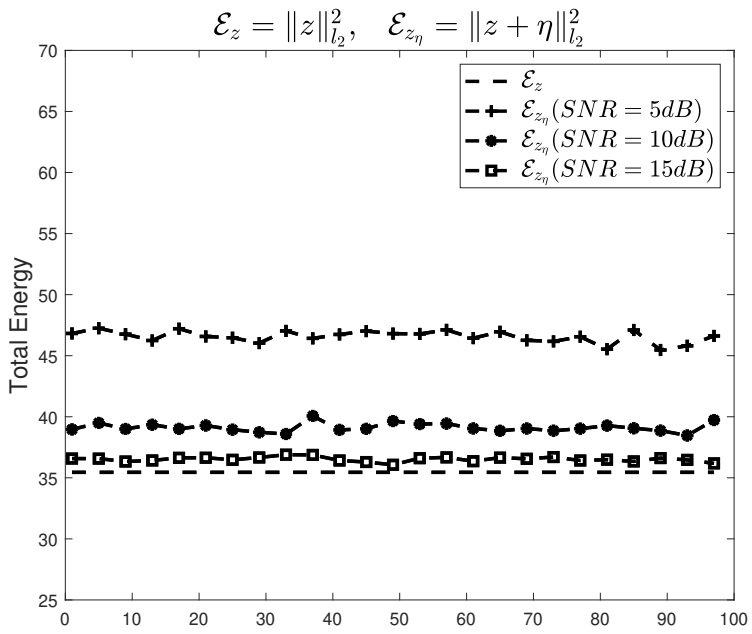


Figure A.1: Plots of total energies for 100 realizations of noise with SNR: 5, 10, 15 dB. The fluctuations are very small relative to the magnitude of the mean energy.

for the noise-corrected CDF, $S_g(t)$. The idea is simply that i.i.d. noise will result, on average, in the addition of a straight line to the CDF. By subtracting this "noise CDF" one can account for the additive contribution. In practice, one replaces $E[R]$ by the estimated CDF of $r(t)$ to get the estimated noise free CDF, $S_g(t)$. This step is performed prior to estimating the CDT.

Alternatively we can define the signal-to-noise ratio $SNR = \mathcal{E}_z/\sigma^2(t_N - t_1)$ in which case (A.4) becomes

$$S_g(t) = \frac{E[R(t)]\,[SNR+1] - \frac{t-t_1}{(t_N-t_1)}}{SNR} \tag{A.5}$$

In the cases where obtaining the SNR is easier as opposed to the energy of noise free signal and the noise variance (as required by Eqn. A.4), the formulation (A.5) may be preferable.

A.2 Distribution of Noise in the CDT Space

In practice, CDF is being calculated using following equation,

$$\begin{aligned} R(t_k) &= \frac{1}{\mathcal{E}_{z_\eta}} \sum_{i=1}^{k} (z(t_i) + \eta(t_i))^2 \\ &= \frac{1}{\mathcal{E}_{z_\eta}} \sum_{i=1}^{k} X(t_i)^2, \; t_1 \le t_k \le t_N \end{aligned}$$

where, $X(n) \sim \mathcal{N}(z(n), \sigma^2)$. Normalizing $X(n)$ with standard deviation σ to make it a normal distribution with unit variance,

$$\begin{aligned} R(t_k) &= \frac{\sigma^2}{\mathcal{E}_{z_\eta}} \sum_{i=1}^{k} \left(\frac{X(t_i)}{\sigma} \right)^2 \\ &= \frac{\sigma^2}{\mathcal{E}_{z_\eta}} Y(t_k) \end{aligned}$$

where, $Y(t_k) \sim \chi^2_{NC}(k, \lambda_k)$ is a random variable with non-central chi-squared distribution. The two parameters of the distribution are: the number of degrees of freedom k and noncentrality parameter $\lambda_k = \sigma^{-2} \sum_{i=1}^{k} z^2(t_i)$.

With the exception of the first few CDF values, k can be considered large enough so that the asymptotic approximation $Y(t_k) \sim \mathcal{N}(k + \lambda_k, 2k + 4\lambda_k)$ can be made yielding

$$\begin{aligned} R(t_k) &\dot{\sim} \mathcal{N}(S(t_k), \Sigma^2(t_k)) \\ &= \mathcal{N}\left(\frac{\sigma^2 (k + \lambda_k)}{\mathcal{E}_{z_\eta}}, \frac{\sigma^4 (2k + 4\lambda_k)}{\mathcal{E}^2_{z_\eta}} \right) \\ & t_1 \leq t_k \leq t_N \end{aligned} \tag{A.6}$$

This approximation is expected to be valid if either k or λ_k are large [140]. Denote as $\Gamma(t_k)$ the random variable used to model $R(t_k)$ and the associated normal PDF of (A.6) as $p_{\Gamma(t_k)}(R(t_k))$.

By the definition of the CDT, $\hat{s}_k \equiv S^{-1}(y_k),\ y_k \in [0, 1]$, so the random variable modelling $\hat{r}_k$ (the result of the CDT) can be written $\hat{R}(y_k) = S^{-1}(\Gamma(\hat{s}_k))$. Thus, we seek the transformation of the random variable $\Gamma(\hat{s}_k)$ by the inverse function $S^{-1}(\cdot)$. Following the rules for the transformation of a random variable we have

$$p_{\hat{R}_k}(\hat{r}_k) = p_{\Gamma(\hat{s}_k)}(r_k)) \left| \frac{\partial \hat{s}(r_k))}{\partial r_k} \right|^{-1} \Big|_{r_k \to S(\hat{r}_k)} \tag{A.7}$$

Given the normal approximation to $p_{\Gamma(t_k)}(R_k)$ given by Eqn. (A.6), noting that $S(\cdot)$ is monotonically increasing, and recalling $\hat{s}(S(x)) = x$ yields

$$\begin{aligned} p_{\hat{R}_k}(\hat{r}_k) = \frac{e^{-\frac{(S(\hat{r}_k) - S(\hat{s}_k))^2}{2\Sigma^2(\hat{s}_k)}}}{\sqrt{2\pi}\Sigma(\hat{s}_k)} \frac{\partial S(\hat{r}_k)}{\partial \hat{r}_k}, \\ 1 \leq k \leq N. \end{aligned} \tag{A.8}$$

Note the independent variable $\hat{r}_k$ is different from the fixed value $\hat{s}_k$. Note also that the mean and variance of (A.6) are evaluated at the clean CDT values (as opposed to times t_k). The mean, $S(\hat{s}_k)$ and variance $\Sigma(\hat{s}_k)$ in (A.8) are found by interpolating $S(t_k),\ \Sigma(t_k)$ respectively.

Appendix B

Supplementary Materials for Chapter 3

B.1 Additional Facts and Remarks

Observation 1: Let $c \neq p$ be two class labels, and $\mathbb{S}^{(c)}$ and $\mathbb{S}^{(p)}$ be defined as in eq. (9). Assume that for any $\varphi_j^{(c)} \in \mathbb{S}^{(c)}$, $f_{j,\varphi_w^{(p)}} \notin \text{span}\Big\{\big(\mathcal{G}_w^{(p)}\big)^{-1}\Big\}$ for any $w = 1, ..., M_p$ where $(f_{j,\varphi_w^{(p)}})'\varphi_j^{(c)}(f_{j,\varphi_w}) = \varphi_w^{(p)}$, it follows that $\widehat{\mathbb{V}}_w^{(p)} \cap \widehat{\mathbb{S}}^{(c)} = \varnothing$ for any $w = 1, ..., M_p$.

Proof. Assume by contradiction that there exists some $\varphi_j^{(c)} \in \mathbb{S}^{(c)}$ such that $\widehat{\varphi}_j^{(c)} \in \widehat{\mathbb{V}}_w^{(p)} \cap \widehat{\mathbb{S}}^{(c)}$. By the composition property and the definition of $f_{j,\varphi_w^{(p)}}$ above, we have that $f^{-1}_{j,\varphi_w^{(p)}} \circ \widehat{\varphi}_j^{(c)} = \widehat{\varphi}_w^{(p)}$ and hence $\widehat{\varphi}_j^{(c)} = f_{j,\varphi_w^{(p)}} \circ \widehat{\varphi}_w^{(p)}$. Since by assumption $\widehat{\varphi}_j^{(c)} \in \widehat{\mathbb{V}}_w^{(p)}$, it follows that $f_{j,\varphi_w^{(p)}} \in \text{span}\Big\{\big(\mathcal{G}_w^{(p)}\big)^{-1}\Big\}$, which is a contradiction.

Remark 1: The last argument above follows from the simple fact that $\widehat{\mathbb{V}}_w^{(p)} = \text{span}\Big(\widehat{\mathbb{S}}_{\varphi_w, \mathcal{G}_w^{(p)}}\Big) = \Big\{ \sum_{i=1}^{k} \beta_i f_{i,w}^{(p)} \circ \widehat{\varphi}_w^{(p)} : \beta_i \in \mathbb{R}\Big\} = \Big\{f \circ \widehat{\varphi}_w^{(p)} : f \in \text{span}\big(\mathcal{G}_w^{(p)}\big)^{-1}\Big\}$.

Remark 2: By the disjointness assumption about different signal classes, i.e., $\mathbb{S}^{(p)} \cap \mathbb{S}^{(c)} = \varnothing$ for $c \neq p$, it follows that $f_{j,\varphi_w^{(p)}} \notin \big(\mathcal{G}_w^{(p)}\big)^{-1}$ for all $p = 1, ..., M_p$. This property intuitively says that signals in $\mathbb{S}^{(c)}$ differs "significantly" from the template $\varphi_w^{(p)}$ in the sense that they can not be recovered using $\varphi_w^{(p)}$ under deformations in $\mathcal{G}_w^{(p)}$. Recall that $\big(\mathcal{G}_w^{(p)}\big)^{-1} = \Big\{ \sum_{i=1}^{k} \alpha_i f_{i,w}^{(p)} : \alpha_i \geq 0\Big\}$, which differs from $\text{span}\Big\{\big(\mathcal{G}_w^{(p)}\big)^{-1}\Big\}$ in the sense that the latter has no restriction on the α_i's. We comment loosely that the assumptions in Observation 1 is not unreasonable when signals in class (c) are "significantly" different than templates in class (p) and the deformation sets are reasonably "small" in the sense that $f_{j,\varphi_w^{(p)}} \notin \text{span}\Big\{\big(\mathcal{G}_w^{(p)}\big)^{-1}\Big\}$.

B.2 Dataset

Fig. 3.6 of Chapter 3 shows some examples of the ten datasets used to evaluate the proposed classification method. We have included few more examples from the datasets in this appendix (Fig. B.1-B.10).

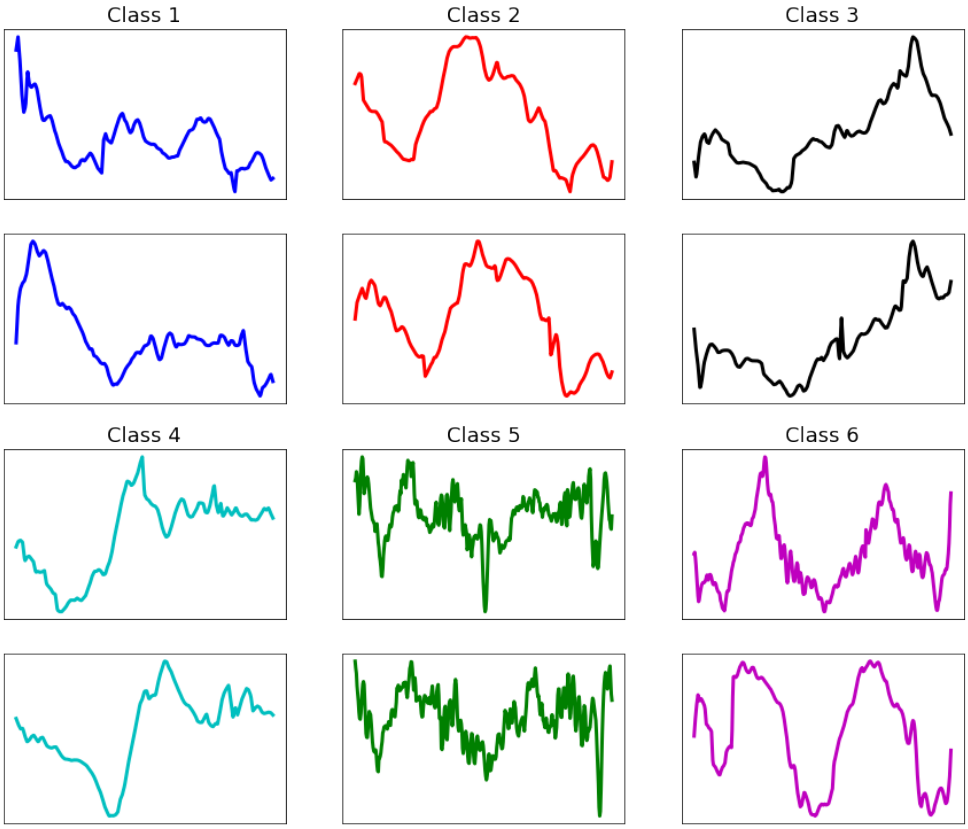


Figure B.1: Dataset: GesturePebbleZ2

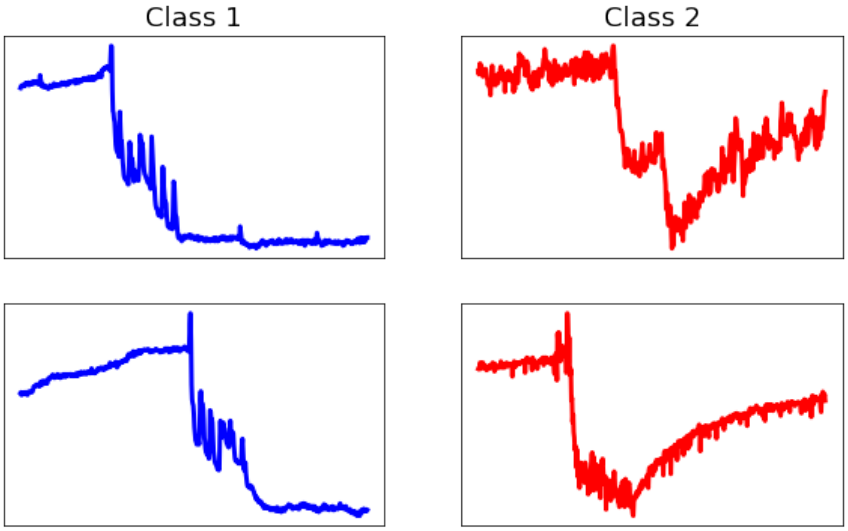


Figure B.2: Dataset: InsectEPGRegularTrain

B.3 Data Efficiency of the Proposed Method

Section 3.4.4 of Chapter 3 shows that the proposed method is data efficient in comparison with the CNNs and provides two examples. Fig. B.11 shows the results from same experiment setup for four other datasets.

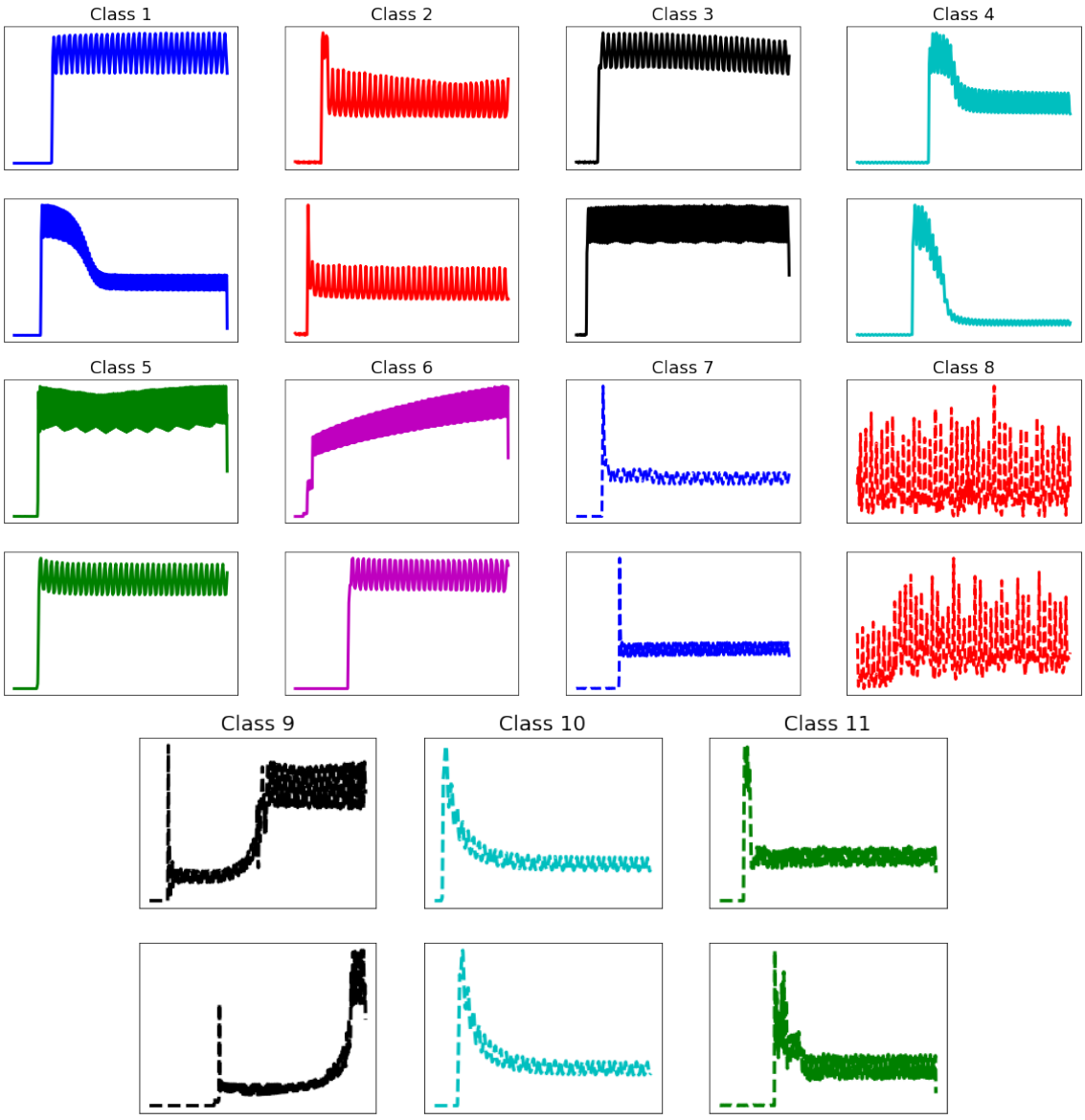


Figure B.3: Dataset: PLAID

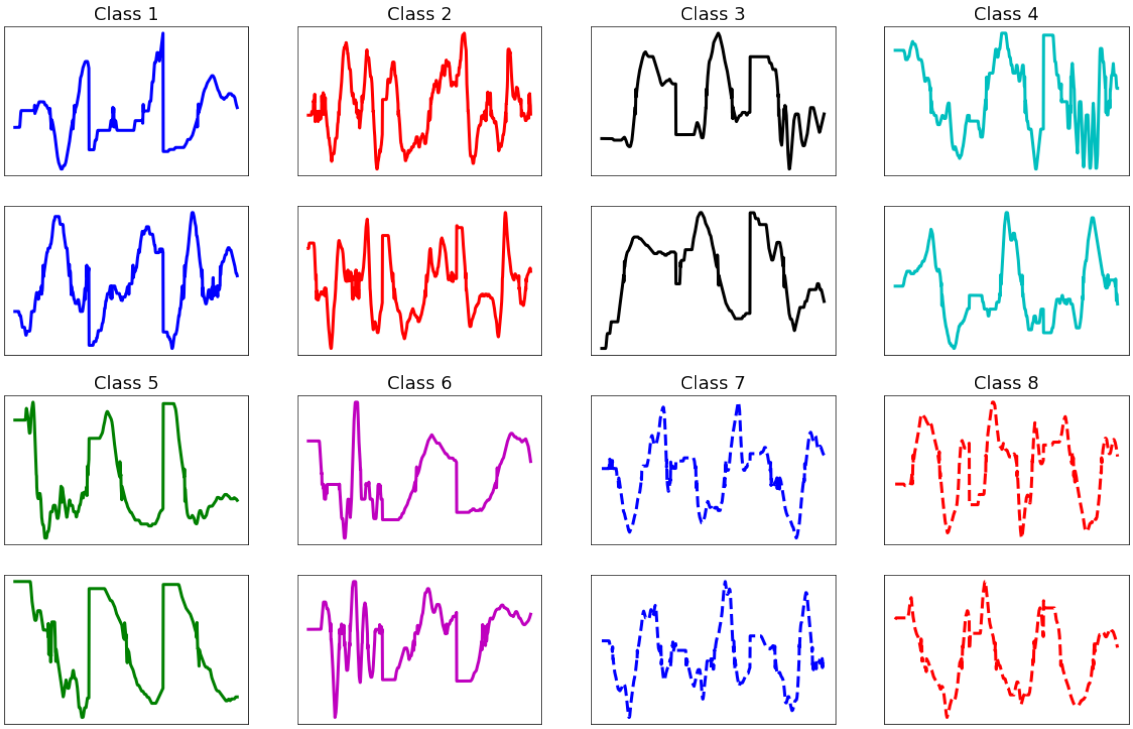


Figure B.4: Dataset: UWaveGestureLibraryAll

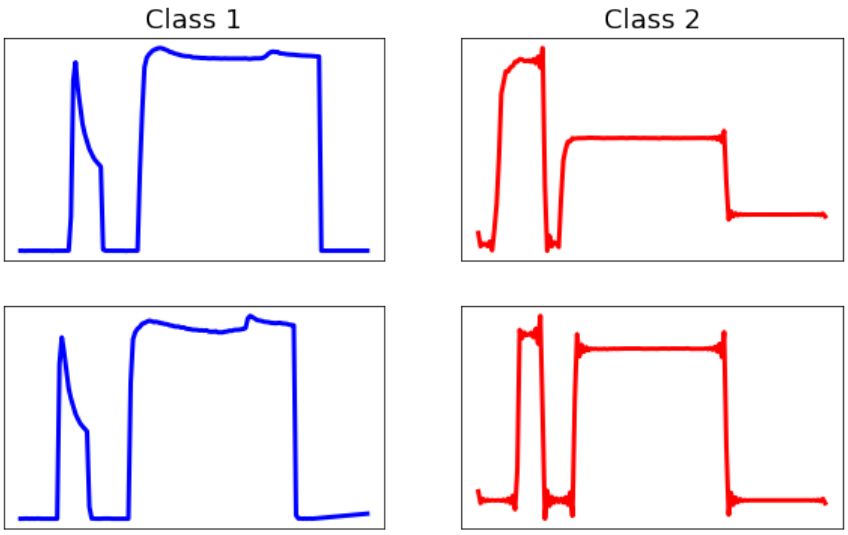


Figure B.5: Dataset: Wafer

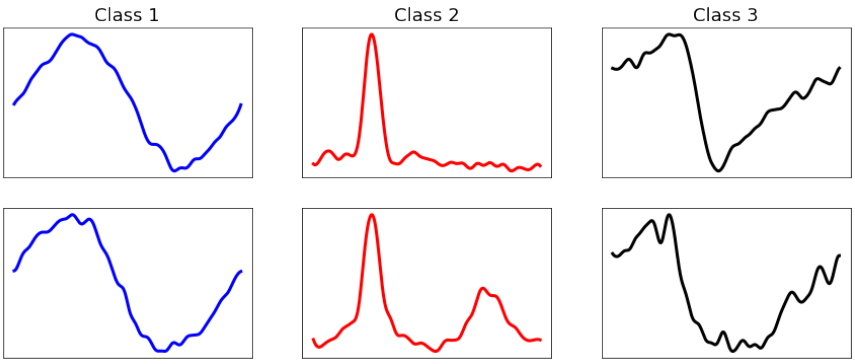


Figure B.6: Dataset: StarLightCurves

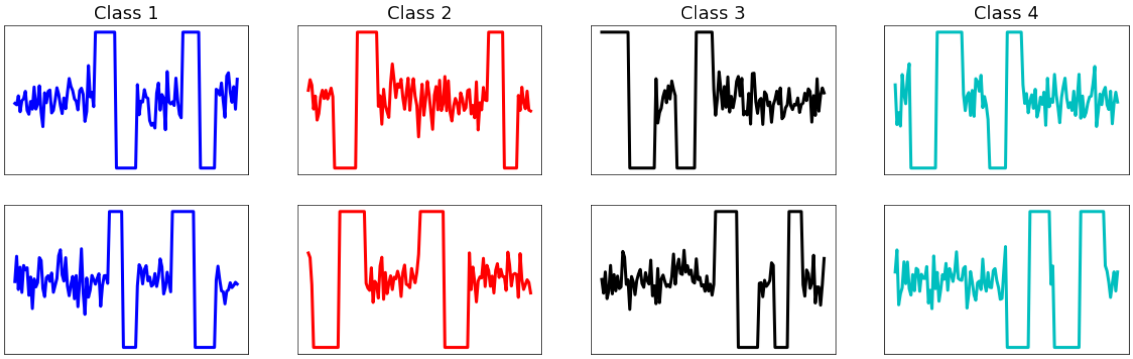


Figure B.7: Dataset: TwoPatterns

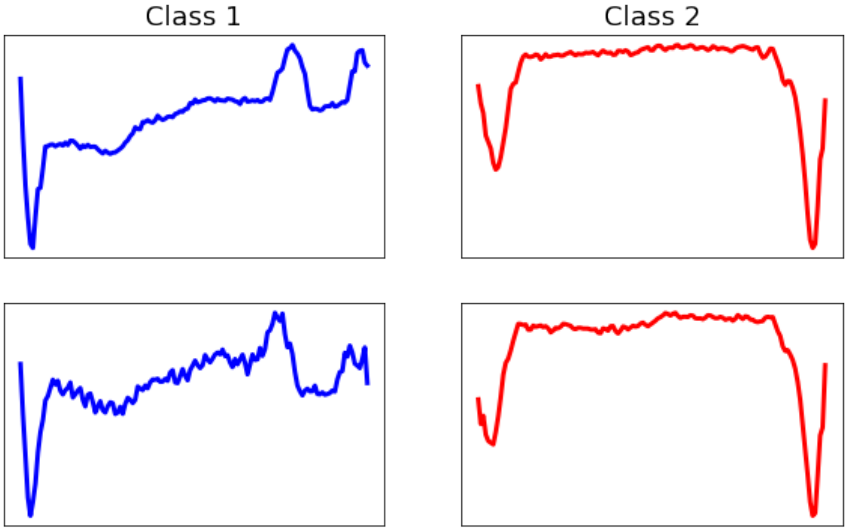


Figure B.8: Dataset: ECG5000

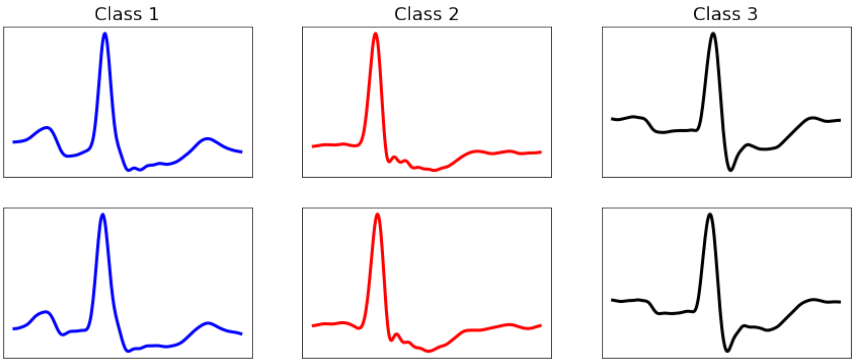


Figure B.9: Dataset: ECG (MLII)

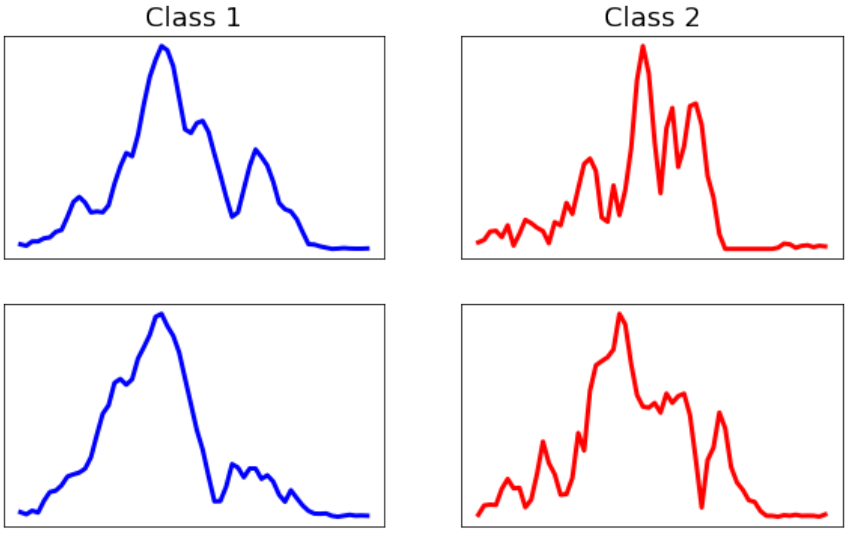


Figure B.10: Dataset: Sonar

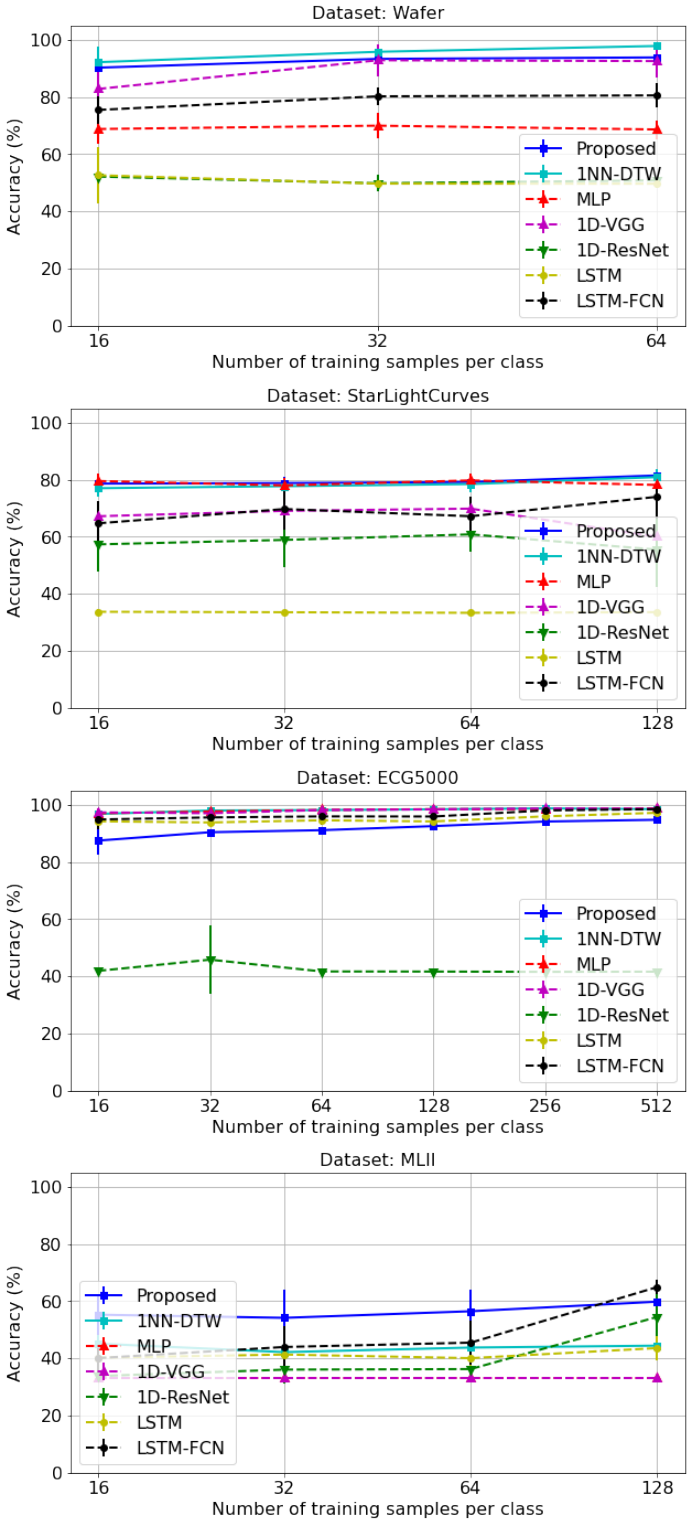


Figure B.11: Accuracy as a function of number of training samples per class for different classification methods.

Appendix C

Supplementary Materials for Chapter 4

C.1 Generative Model for Convection-Diffusion Equation

1D convection-diffusion equation (PDE) initialized with the initial condition $u(x,0)$ is given by:

$$\begin{aligned} &\dot{u}(x,t) = \nu u_x(x,t) + D u_{xx}(x,t), \quad x \in \Omega_x\,, t \in \mathbb{R}_+, \\ &\text{Initial condition: } u(x,0) = \frac{1}{\sqrt{4\pi}} e^{-\frac{x^2}{4}}, \end{aligned} \tag{C.1}$$

where ν and D denote the wave speed and the diffusion coefficient, respectively. A solution to the PDE defined in eq. (C.1) can be derived as:

$$u(x,t) = \frac{1}{\sqrt{4\pi D t}} e^{-\frac{(x-\nu t)^2}{4Dt}}. \tag{C.2}$$

The expression for the sensor data $s(t)$ at location $x = x_m$ is given by:

$$s(t) = \frac{1}{\sqrt{4\pi D t}} e^{-\frac{(x_m-\nu t)^2}{4Dt}}, \tag{C.3}$$

which can be represented using the transport generative model as:

$$s(t) = \dot{g}_{\nu,D}(t)\varphi_{\nu,D}(g_{\nu,D}(t)), \quad g_{\nu,D} \in \mathcal{G}_{\nu,D}, \tag{C.4}$$

where an inhomogeneous template $\phi_{\nu,D}(x\,,t) : \Omega_x \times \mathbb{R}_+ \to \mathbb{R}$ and a warping function in time are:

$$\begin{aligned} \varphi_{\nu,D}(t) &= f_{x_m}(t) e^{-t}, \\ g_{\nu,D}(t) &= \frac{(x_m - \nu t)^2}{4Dt}, \\ x_m, \nu, D > 0, &\quad t > \frac{x_m}{\nu}. \end{aligned} \tag{C.5}$$

Note that the condition $t > \frac{x_m}{\nu}$ ensures $g_{\nu,D}(t)$ to be monotonically increasing, i.e., $\dot{g}_{\nu,D}(t) > 0$. Then, we can choose a support of time t as:

$$\text{sup}(t) := \left[t_0 = \frac{x_m}{\nu} + a\,, t_1\right] := \Omega_t\,, \tag{C.6}$$

for $a > 0$ and a large enough $t_1 > t_0$. Here, $f_{x_m}(t)$ is an arbitrary function which will be derived later. First, we derive the inverse of the warping map, i.e., $g^{-1}_{\nu,D}(t)$. To distinguish between the warping map and its inverse, we replace the independent variable t in $g^{-1}_{\nu,D}(t)$ with another variable z. Meaning, $g^{-1}_{\nu,D}(z)$ is

defined by letting:

$$z = g_{\nu,D}(t) = \frac{(x_m - \nu t)^2}{4Dt} \tag{C.7}$$

$$\Leftrightarrow \ \nu^2 t^2 - 2\left(x_m \nu + 2Dz\right) t + x_m^2 = 0 \tag{C.8}$$

A condition of existence for solution is:

$$\Delta := \left(x_m \nu + 2Dz\right)^2 - \nu^2 x_m^2 = 4\left(D^2 z^2 + \nu D x_m z\right) > 0\,.$$

Since $\nu > 0$, we have $\Delta > 0$, i.e. equation (C.8) always have two solutions. Note that if $\nu < 0$, then for $\nu < -\frac{x_m}{t}$, the condition becomes (for $x_m\,, z > 0$):

$$\Delta > 0 \ \Leftrightarrow \ -\frac{Dz}{x_m} < \nu < -\frac{x_m}{t}\,,$$

In conclusion, this quadratic form has 2 solutions:

$$t_\pm := \frac{1}{\nu^2}\left(x_m \nu + 2Dz \pm \sqrt{\left(x_m \nu + 2Dz\right)^2 - \nu^2 x_m^2}\right)\,. \tag{C.9}$$

Under the condition that $g_{\nu,D}$ and $g_{\nu,D}^{-1}$ are one-one increasing functions, we choose the inverse of the warping map to be,

$$g_{\nu,D}^{-1}(z) = \frac{1}{\nu^2}\left(x_m \nu + 2Dz + \sqrt{\left(x_m \nu + 2Dz\right)^2 - \nu^2 x_m^2}\right)\,. \tag{C.10}$$

Eq. (C.10) can be written as a function t instead of z such that

$$g_{\nu,D}^{-1}(t) = \frac{1}{\nu^2}\left(x_m \nu + 2Dt + \sqrt{\left(x_m \nu + 2Dt\right)^2 - \nu^2 x_m^2}\right)\,. \tag{C.11}$$

Let, the family of the functions $g_{\nu,D}^{-1}$ forms a set $\mathcal{G}_{\nu,D}^{-1}$ which is non-convex. However, in section 4.4.3 of Chapter 4, we demonstrated that the set $\mathcal{G}_{\nu,D}^{-1}$ can be approximated as a convex set with respect to D. Here, we aim to derive a similar approximation of $\mathcal{G}_{\nu,D}^{-1}$ which is convex with respect to the parameter ν.

Convexify set of inverses of warping maps: Similar to 4.4.3, we aim to convexify the set $\mathcal{G}_{\nu,D}^{-1}$ by taking the 1st order Taylor expansion of the function $g_{\nu,D}^{-1}$ with respect to ν at ν_0 (since, we assume $\nu \sim U\left(\nu_0 - \epsilon\,, \nu_0 + \epsilon\right)$):

$$\begin{aligned}
g_{\nu,D}^{-1}(t) \approx & g_{\nu_0,D}^{-1}(t) + \frac{\partial g_{\nu,D}^{-1}(t)}{\partial \nu}\Big|_{\nu=\nu_0}(\nu - \nu_0) \\
= & \frac{1}{\nu_0^2}\left(x_m \nu_0 + 2Dt + \sqrt{\left(x_m \nu_0 + 2Dt\right)^2 - \nu_0^2 x_m^2}\right) \\
& - \frac{1}{\nu_0^2}\left(x_m + \frac{4Dt}{\nu_0} + \frac{2}{\nu_0}\sqrt{\left(x_m \nu_0 + 2Dt\right)^2 - \nu_0^2 x^2} - \frac{2Dt x_m}{\sqrt{\left(x \nu_0 + 2Dt\right)^2 - \nu_0^2 x_m^2}}\right)(\nu - \nu_0) \\
:= & \tilde{g}_{\nu,D}^{-1}(t)\,.
\end{aligned}$$

We then evaluate the $\tilde{g}^{-1}_{\nu,D}(t)$ at the boundary values of ν, i.e., $\nu = \nu_0 \pm \epsilon$ to obtein:

$$
\begin{aligned}
g^{-1}_{\nu_0\pm\epsilon,D}(t) \approx & \tilde{g}^{-1}_{\nu_0\pm\epsilon,D}(t) \\
= & \frac{1}{\nu_0^2}\left(x_m\nu_0 + 2Dt + \sqrt{\left(x_m\nu_0+2Dt\right)^2 - \nu_0^2 x_m^2}\right) \\
& \pm \frac{1}{\nu_0^2}\left(x_m + \frac{4Dt}{\nu_0} + \frac{2}{\nu_0}\sqrt{\left(x_m\nu_0+2Dt\right)^2 - \nu_0^2x^2} - \frac{2Dtx_m}{\sqrt{\left(x\nu_0+2Dt\right)^2 - \nu_0^2x_m^2}}\right)\epsilon \\
= & g^{-1}_{\nu_0,D}(t) \pm \frac{1}{\nu_0^2}\left(x_m + \frac{4Dt}{\nu_0} + \frac{2}{\nu_0}\sqrt{\left(x_m\nu_0+2Dt\right)^2 - \nu_0^2x^2} - \frac{2Dtx_m}{\sqrt{\left(x\nu_0+2Dt\right)^2 - \nu_0^2x_m^2}}\right)\epsilon,
\end{aligned}
$$

which is convex with respect to ϵ. It indicates that the set $\mathcal{G}^{-1}_{\nu,D}$ is convex with respect to $\nu \sim U(\nu_0-\epsilon, \nu_0+\epsilon)$ for a given D.

Derive template $\varphi_{\nu,D}(t)$: To derive the template $\varphi_{\nu,D}(t)$ defined in eq. (C.5), we need to find an expression for the function $f_{x_m}(t)$. We choose $f_{x_m}(t)$ such that:

$$
\frac{1}{4D}\left(\nu^2 - \frac{x_m^2}{t^2}\right) f_{x_m}\left(\frac{(x_m-\nu t)^2}{4Dt}\right) e^{-\frac{(x_m-\nu t)^2}{4Dt}} = \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{(x_m-\nu t)^2}{4Dt}}.
$$

The above equality holds if we choose $f_{x_m}(t)$ in such a way that:

$$
\begin{aligned}
& \frac{1}{\sqrt{4\pi Dt}} = \frac{1}{4D}\left(\nu^2 - \frac{x_m^2}{t^2}\right) f_{x_m}\left(\frac{(x_m-\nu t)^2}{4Dt}\right) \\
\Leftrightarrow \; & f_{x_m}\left(\frac{(x_m-\nu t)^2}{4Dt}\right) = \sqrt{\frac{4D}{\pi}}\frac{t^{\frac{3}{2}}}{\nu^2t^2 - x_m^2} \\
\Leftrightarrow \; & f_{x_m}(t) = \frac{1}{\nu}\sqrt{\frac{4D}{\pi}}\frac{\left(\left(x_m\nu+2Dt\right) + \sqrt{\left(x_m\nu+2Dt\right)^2 - \nu^2x_m^2}\right)^{\frac{3}{2}}}{\left(\left(x_m\nu+2Dt\right) + \sqrt{\left(x_m\nu+2Dt\right)^2 - \nu^2x_m^2}\right)^2 - \nu^2x_m^2}.
\end{aligned}
$$

Thus, the template is given by:

$$
\varphi_{\nu,D}(t) = \frac{1}{\nu}\sqrt{\frac{4D}{\pi}}\frac{\left(\left(x_m\nu+2Dt\right) + \sqrt{\left(x_m\nu+2Dt\right)^2 - \nu^2x_m^2}\right)^{\frac{3}{2}}}{\left(\left(x_m\nu+2Dt\right) + \sqrt{\left(x_m\nu+2Dt\right)^2 - \nu^2x_m^2}\right)^2 - \nu^2x_m^2} e^{-t}. \tag{C.12}
$$