Time Series Analysis using a Kernel based Multi-Modal Uncertainty Decomposition Framework

Rishabh Singh*
Computational NeuroEngineering Laboratory
University of Florida
Florida, FL 32611

Jose C. Principe
Computational NeuroEngineering Laboratory
University of Florida
Florida, FL 32611

Abstract

This paper proposes a kernel based information theoretic framework that provides a sensitive multi-modal quantification of time series uncertainty by leveraging a quantum physical description of the projected feature space of data in a Reproducing Kernel Hilbert Space (RKHS). We specifically modify the kernel mean embedding, which yields an intuitive physical interpretation of the signal structure, to produce a data based "dynamic potential field". This results in a new energy based formulation that exploits the mathematics of quantum theory and facilitates a multi-modal physics based uncertainty representation of the signal at each data sample. We demonstrate in this paper that such uncertainty features provide a better ability for online detection of statistical change points in time series data when compared to existing non-parametric and unsupervised methods. We also demonstrate a better ability of the framework in clustering time series sequences when compared to discrete wavelet transform features on a subset of VidTIMIT speaker recognition corpus.

1 INTRODUCTION AND MOTIVATION

A time series is defined as a sequence of measurements over different points of time to describe a system behavior (Hamilton, 1994). Such signals are abundantly found in a variety of applications such as economics, finance, engineering (speech signals), neuroscience and natural sciences (seismic signals, temperature measurements, etc.), thus making its study an active research arena.

A central problem associated with the analysis of real world time series datasets is their frequent defiance of statistical assumptions (Manuca & Savit, 1996). Real world time series signals are often non-stationary which means that they are characterized by time-varying statistical properties. Current machine learning approaches and information theoretic metrics that assume stable historical trends fail to characterize the intrinsic uncertainty and local statistical drifts in such signals. Therefore, there is a need for information processing tools for time series analysis that are sensitive towards local dynamical changes in datasets while also being able to take into account the global statistical properties of the signal.

We attempt to address this requirement by utilizing a recently introduced information theoretic framework (Singh & Principe, 2020) that leverages the kernel mean embedding (KME) for a universal characterization of data PDF in the RKHS and enables a physical interpretation of the data space as a potential field. This is followed by an imposition of the local dynamical structure of the time series data onto the characterized data PDF by using concepts of quantum physics (specifically the Schrödinger's equation). This modifies the KME into a *dynamic potential field* leading to an energy based reformulation of the time series dynamics, which in turn enables us to compute the various uncertainty modes associated with the data by using a moment decomposition procedure similar to that used in physics (to quantify particle-field interactions). Intrinsically, this approach decomposes local time realizations of the stochastic joint process PDF in terms of quantum uncertainty moments.

The rest of the paper is organized as follows. We begin with a survey of relevant literature associated with time series analysis in section 2 followed by a brief description of the proposed framework and contributions in section 3. We describe the framework details and associated background concepts in section 4. In section 5, we describe our experiments and discuss the results. We conclude the paper in section 6.

*Corresponding author. Email address: rish283@ufl.edu.

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

2 LITERATURE REVIEW

Time series clustering has been the principle approach towards extraction of relevant information from temporal data. The primary objective of clustering here is to discover interesting patterns in the associated data dynamics. A vast number of different subject areas utilize time series clustering for their respective applications. Examples include energy consumption pattern discovery (Košmelj & Batagelj, 1990) and discovery of seasonality patterns in finance (Kumar et al., 2002). Time series clustering can be broadly classified into three categories which are whole time series clustering, subsequence time series clustering and time point clustering (Aghabozorgi et al., 2015). From an approach based perspective, one can classify the methods based on whether they are dependent on raw data analysis, feature extraction or model parameters (Liao, 2005). Raw data based clustering refers to methods where clustering is directly performed on data. Examples include Agglomerative hierarchical clustering (Kakizawa et al., 1998) and κ-Means method. These methods often employ different distance measures such as Euclidean distance, Chernoff information divergence (Nielsen, 2013), Kullback Leibler divergence (Van Erven & Harremos, 2014) and dynamic time warping (Berndt & Clifford, 1994). Dynamic time warping has been well know in dealing with temporal drifts and datasets with unequal lengths. Feature based approaches include discrete Fourier transform (Bracewell & Bracewell, 1986), discrete wavelet transform (Heil & Walnut, 1989a) and LPC coefficients (Tierney, 1980). We specifically focus on whole time series clustering amongst these categories where clustering is performed on discrete objects, each of which represents an entire time series. An important example of representation based approach to whole time series clustering is discrete wavelet transform (Heil & Walnut, 1989a) which is known for capturing both frequency and location information and provides an advantage over discrete Fourier transform due to its multi-scale temporal resolution.

There has also been an increased research interest in change point detection (CPD) algorithms owing to their increased importance in various different fields such as speech recognition, image analysis, medical procedures and climate change monitoring (Aminikhanghahi & Cook, 2016). Change point detection refers to abrupt changes in data characteristics and it can be considered to be a subset of time series clustering. A detailed overview and survey of methods can be found in (Aminikhanghahi & Cook, 2016). CPD algorithms can be broadly classified as supervised or unsupervised methods. Supervised methods include naive Bayes (Rish et al., 2001), support vector machines (Scholkopf & Smola, 2001) and hidden Markov models (Fine et al., 1998). Unsupervised methods include Bayesian change point detector (Adams & MacKay, 2007) and divergence measures such as Kullback Leibler divergence (Van Erven & Harremos, 2014). Another import basis for classifying CPD algorithms is based on whether they are implemented online or offline. Offline methods consider the entire time series to determine points of change. In online methods, changes in the signals are tracked in real time. Hence, before the arrival of the new data point, existing data point must be evaluated to determine if a change has occurred at that point with respect to the older points. The main challenges faced in online CPD arises from the requirement of algorithms have have low reaction times towards changes and the requirement to use as few data samples as possible to effectively quantify data statistics and detect changes. Current ITL based divergence measures fail to simultaneously meet the requirements of quantifying global signal statistics and detecting changes in dynamics.

3 CONTRIBUTIONS

The goal of this paper is to introduce a new framework for representation of time series signals that is able to quantify local data dynamics relative to its overall general PDF in an unsupervised manner with high sensitivity and specificity. This is valuable for applications in time series analysis related to change point detection and whole series clustering. The proposed framework operates by first implementing an RKHS based representation of the data PDF in the form of kernel mean embedding (KME) (Muandet et al., 2017). Unfortunately, KME is only applicable to static multivariate data or stationary time series, which is too limiting for the current applications of time series analysis. The fundamental difficulty is that a stochastic process is a family of random variables over time, which requires for the non-stationary case, the analysis of the joint distribution over samples. Instead of using the traditional Markov assumption to simplify the problem, here we seek to employ the local dynamics of the time series data using the Schrödingers equation to improve the KME concept, which leads to an energy-based description of the signal structure as a dynamic potential field. This enables the extraction of the various modes of uncertainty related to the interaction of an upcoming data sample with the signals history by implementing a moment decomposition procedure based on orthogonal polynomial projections, similar to those used in quantum physics to extract the eigenmodes of a particle with respect to its neighboring force field. We also refer to extracted data uncertainty modes as information eigenmodes since they are essentially energy-based information data features. The advantages offered by such a framework for time series analysis are as follows:

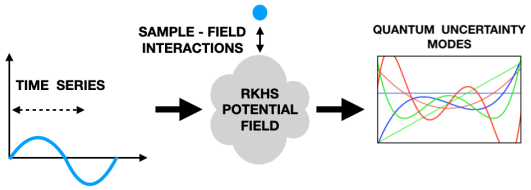


Figure 1: Abstract depiction of framework.

- Since the framework is based in the RKHS, it takes into account all even order data associations in an unsupervised manner.
- The extracted uncertainty modes provide an enhanced quantification of the tail regions of the data PDF, where uncertainty is maximum and statistical change points are most likely to occur.
- The framework can be implemented on a sample-by-sample basis so that each new sample can be characterized by uncertainty mode values online and determined the degree to which it belongs to the overall distribution of past samples.

The main idea in this paper is to introduce a new information theoretic framework to accomplish pattern recognition tasks associated with time series data that is usually not achievable by classical information theoretic methods due to their inability to quantify sample-by-sample dynamics of the signal (important in online time series analysis). It is for this reason that we look towards quantum physics, which provides a principled quantification of particle-particle dynamics in a physical system, to interpret data. Hence the Schrödinger's equation is introduced with similar Hamiltonian operators as is used in physics, to quantify the sample-to-sample dynamics in the data field described in a RKHS using the kernel mean embedding (also termed as information potential). The idea of extracting the spectrum of the Hamiltonian eigenfunctions thereafter (using orthogonal Hermite polynomial projections) is to try and quantify the interaction of a data sample with the past data along the various moments of the data PDF akin of the spectral modes in the power spectrum, which have practical value for analysis.

4 UNCERTAINTY DECOMPOSITION FRAMEWORK

The framework (fig. 1) implementation consists of three major steps - i) Representation of sample-by-sample data PDF characteristics using empirical estimate of the KME, that leads to a potential field based interpretation of the data space. ii) Physics based reformulation of the KME in terms of the Schrödinger's equation to create a dynamic potential field. iii) Extraction of uncertainty modes based on moment decomposition of the interaction of a new data sample with the dynamic potential field created by previous samples. We delve into the details of each step, first starting with a brief review of the kernel mean embedding metric.

4.1 KERNEL MEAN EMBEDDING

The reproducing kernel Hilbert space associated with positive definite kernels allows one to universally pose any non-linear relationship in the input (data) space as a linear relationship in a higher dimensional functional space, thanks to the acclaimed "kernel trick" (Aronszajn, 1950). Following similar intuition, another elegant property of the RKHS is their ability to embed statistical measures in the inner product structure of the reproducing kernel (the KME theory) which allows one to non-parametrically quantify a data distribution from the input space as an element of its associated RKHS (Muandet et al., 2017).

Definition 1 (Kernel Mean Embedding) *Consider the space $\mathcal{Z}(\mathcal{X})$ to consist of all probability measures $\mathbb{P}$ on a measurable space $(\mathcal{X}, \Sigma)$. The kernel mean embedding of probability measures in $\mathcal{Z}(\mathcal{X})$ into an RKHS denoted by $\mathcal{H}$ and characterized with a reproducing kernel $k : X \times X \to \mathbb{R}$ is defined by a mapping*

$$\mu : \mathcal{Z}(\mathcal{X}) \to \mathcal{H}, \;\; \mathbb{P} \mapsto \int k(x, .) d\mathbb{P}(x).$$

The kernel mean embedding (KME), therefore, represents the probability distribution in terms of a mean function associated with the kernel feature map in the space of the distribution. In other words,

$$\phi(\mathbb{P}) = \mu_{\mathbb{P}} = \int k(x, .) d\mathbb{P}(x). \tag{1}$$

There are several useful properties associated with the KME (Muandet et al., 2017). For characteristic kernels, the KME is injective, meaning that $\mu_{\mathbb{P}} = \mu_{\mathbb{Q}}$ only when $\mathbb{P} = \mathbb{Q}$, thus allowing for unique characterizations of data distributions. In most applications, the nature of $\mathbb{P}$ is not known or pre-defined. One therefore depends on the empirical estimation of the KME. This can be approximated using its unbiased estimate given by

$$\hat{\mu} = \frac{1}{n} \sum_{t=1}^{n} k(x_t, .). \tag{2}$$

Here $\hat{\mu}$ converges to μ for $n \to \infty$, according to the law of large numbers.

4.2 INFORMATION POTENTIAL FIELD

One can also derive the empirical estimate of the KME from Rényi entropy, where it takes the form of information potential field. To elaborate further, let us consider the Rényi quadratic entropy (Rényi et al., 1961) given by

$$H_2(X) = -log \int p(x)^2 dx = -logV(X). \quad (3)$$

One notices here that the argument of the logarithm in Rényi entropy, $V(X)$, is an important quantity called the information potential (IP) of the data set (Principe et al., 2000), which is simply the mean value of the PDF. One can estimate this quantity by using the Parzen density estimator (Parzen, 1962) for estimating $p(x)$. Hence, assuming a Gaussian kernel window of kernel width σ, one can readily estimate directly from experimental data $x_i, i = 1, ..., N$ the information potential as

$$\begin{aligned} V(X) &= \int p(x)^2 dx = \int \left(\frac{1}{N} \sum_{i=1}^{N} G_\sigma(x - x_i) \right)^2 dx \\ &= \frac{1}{N^2} \int \left(\sum_{i=1}^{N} \sum_{j=1}^{N} G_\sigma(x - x_j).G_\sigma(x - x_i) \right) dx \\ &= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \int G_\sigma(x - x_j).G_\sigma(x - x_i) dx \\ &= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G_{\sigma/\sqrt{2}}(x_j - x_i) \end{aligned} \quad (4)$$

Hence the IP is a number obtained by the double sum of the Gaussian functions centered at differences of samples with a larger kernel size. The same result is obtained when using the empirical estimate of the KME in a RKHS defined by the Gaussian function. There is a physical interpretation of $V(X)$ if we think of the samples as particles in a potential field, hence the name information potential. It can also be interpreted as the total potential created by the data set in an RKHS, i.e.

$$V(X) = \frac{1}{N} \sum_{j=1}^{N} V(x_j), \quad (5)$$

where,

$$V(x) = \frac{1}{N} \sum_{i=1}^{N} G(x - x_i) \quad (6)$$

represents the *field* due to each sample, which can be interpreted as an information particle. One refers to $V(x)$ as the *information potential field* (IPF). One can notice that it is basically a continuous function over the RKHS obtained by the sum of Gaussian bumps centered on the samples. We now delve into the quantum physical interpretation of the empirical KME which we now refer to as the IPF henceforth.

4.3 QUANTUM FORMULATION OF THE IPF

It is well known in the field of physics that, unlike a classical system which is characterized by deterministic parameters, a quantum-physical system is characterized by discrete transitions induced by Hamiltonian operators that lead to increased stochasticity and uncertainty in the measurement of system state (Griffiths & Schroeter, 2018). In such a case, the probabilistic wave-function determines the system behavior. As an example, one can consider the case of single particle of mass $m = 1$ in a general quantum system. Its associated time independent Schrödinger's equation is then given by

$$\hat{H}\psi = \left(-\frac{\hbar^2}{2m} \nabla^2 + V_r(x) \right) \psi(x) = E\psi \quad (7)$$

Here, $\hat{H}$ denotes the Hamiltonian operator and is given by $\hat{H} = -\frac{\hbar^2}{2m}\nabla^2 + V(x)$, where $-\frac{\hbar}{2m}$ is the kinetic energy operator with $\hbar$ and m being the Planck's constant and particle mass respectively. $V_r(x)$ represents the potential energy of the particle at position x. ∇^2 denotes the Laplacian operator and $\psi(x)$ denotes the wave-function value at position x that also implies that the probability of finding the particle at that position given by $p(x) = |\psi(x)|^2$.

One can extend a similar interpretation towards data systems (Principe, 2010). It can be deduced that the IPF is always positive and regions of space with more samples will have a larger IP, while regions of the space with few samples will have a lower IP. Here, one notices that the shape of the kernel function will determine the "gravity", instead of the inverse square law of physics. Following this intuition, one can readily extend the idea of a potential field over the space of the samples with quantum theoretical concepts (Principe, 2010) to enhance the paradigm for conditions where the time series statistics change over time, and our goal is to quantify it using the local spatial structure. One can formulate a Schrödinger's time-independent equation to define a new potential energy function $V_s(x)$ that characterizes the data space. Here, $V_s(x)$ is based on a wave-function defined by using the IPF as the probability measure $p(x)$. Since $p(x) = |\psi(x)|^2$, it follows that for a set of information particles with a Gaussian kernel, the wave-function for one dimensional information particle becomes,

$$\psi(x) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} G_\sigma(x - x_i)} \quad (8)$$

One can assume that all information particles have the same mass(i.e. $m = 1$) and that $V_s(x)$ can be rescaled such that σ (bandwidth of the kernel window) is the only free parameter that replaces all physical constants. This reformulates (7) to yield the Schrödinger's time-independent equation for information particles as

$$H\psi(x) = \bigg(-\frac{\sigma^2}{2}\nabla^2 + V_s(x)\bigg)\psi(x) = E\psi(x) \quad (9)$$

where H denotes the Hamiltonian. Solving for $V_s(x)$, we obtain:

$$V_s(x) = E + \frac{\sigma^2/2\nabla^2\psi(x)}{\psi(x)} \quad (10)$$

which was called the *quantum information potential field* (QIPF) denoted by $V_s(x)$. To determine the value of $V_s(x)$ uniquely, it is required that $min(V_s(x)) = 0$, which makes

$$E = -min\frac{\sigma^2/2\nabla^2\psi(x)}{\psi(x)} \quad (11)$$

where $0 \leq E \leq 1/2$. Here, $\psi(x)$ is the eigenfunction of H and E is the lowest eigenvalue of the operator, which corresponds to the ground state. Given the data set, one can expect $V_s(x)$ to increase quadratically outside the data region and to exhibit local minima associated with the locations of highest sample density (clusters). One can interpret this as clustering since the potential function attracts the data distribution function $\psi(x)$ to its minima, while the Laplacian drives it away, producing a complicated potential function in the space. It should be noted that, in this framework, E sets the scale at which the minima are observed. One can also extend this derivation to multidimensional data. It can be seen that $V_s(x)$ in (10) is also a potential function that differs from $V(x)$ in (6) because it is now an energy based formulation associated with the quantum description of the IPF.

4.4 EXTRACTION OF QUANTUM UNCERTAINTY MODES

Unlike the classical interpretation, the quantum interpretation provides a much more detailed decomposition of the system dynamics since it consists of a large (potentially infinite) number of stochastic *features*, given by the energy modes. Likewise, the same interpretation holds when applying this quantum field potential to data. Owing to the finite number of samples, the local structure of the PDF in the space of samples is very difficult to quantify. In the input space, one normally uses clustering or other techniques to achieve this goal, but thee is still an enormous difficulty in characterizing the tails of distributions. Here it is relevant to remember the characteristic function of the PDF and the cumulants, which has been a work horse of statistics. The issue with the cumulants is the complexity of estimating the higher order moments in high dimensional data. In this approach, one instead follows the teachings of quantum theory and employs a model decomposition of the wave function to subsequently extract uncertainty modes characterizing the PDF tails. To understand the decomposition procedure of the data wave function, it is helpful to first analyze the quantum harmonic oscillator which is popular example of a quantum model that is pervasively used in many fields to describe system behavior (econometrics, for instance (Meng et al., 2016; Ahn et al., 2018)). In this case, one can describe the decomposition of the system's wave function in the following manner (Griffiths & Schroeter, 2018).

Definition 2 (Quantum Harmonic Oscillator) *The potential energy of a particle can be generalized using Hooke's law as* $V(X) = \frac{1}{2}m\omega^2x^2$. *The Hamiltonian of the particle characterizes its dynamic parameters (position and momentum) and is formulated as*

$$\hat{H} = \frac{\hat{p}^2}{2m} + \frac{1}{2}m\omega^2x^2 \quad (12)$$

where $\omega = \sqrt{\frac{k}{m}}$ *is the angular frequency of the oscillator,* x *is the position and* $\hat{p} = -i\hbar\frac{d}{dx}$ *represents the momentum operator. Given this Hamiltonian, the time-independent Schrödinger's equation can be formulated as*

$$\hat{H}\psi(x) = \bigg[-\frac{\hbar^2}{2m}\frac{d^2}{dx^2} + \frac{1}{2}m\omega^2x^2\bigg]\psi(x) = E\psi(x) \quad (13)$$

This differential equation can be treated as an eigenvalue problem and solved using the spectral method to yield a family of wave-function modes, $\psi_n(x)$*, that amount to successive Hermite polynomial moments. The solutions are given as:*

$$\begin{aligned} E_0 &= \frac{\hbar w}{2}, \quad \psi_0 = \alpha_0 e^{\frac{-y^2}{2}} \\ E_1 &= \frac{3\hbar w}{2}, \quad \psi_1 = \alpha_0(2y)e^{\frac{-y^2}{2}} \\ E_2 &= \frac{5\hbar w}{2}, \quad \psi_2 = \alpha_0(4y^2-2)e^{\frac{-y^2}{2}} \\ &\cdot \end{aligned} \quad (14)$$

Here, $y = \sqrt{\frac{mw}{\hbar}}x$, $\psi_0, \psi_1, \psi_2...$ *are the obtained wave-function modes and* $E_0, E_1, E_2...$ *are their corresponding eigenvalues. Therefore the solution to the Schrödinger equation for the harmonic oscillator yields infinite eigenfunctions successively associated with each other through Hermite polynomials.*

Hence it is noticeable that the quantum interpretation enables one to extract the various intrinsic energy modes associated with the system, along with the corresponding eigenvalue of each mode. In the previous section, we described the Schrödinger's equation associated with the quantum IPF (QIPF) given by (9) which essentially provides a quantum interpretation of data dynamics similar to how (13) does for the harmonic oscillator. The objective is to now extract successive energy modes (analogical to those obtained in (14)) associated with the QIPF given by $V_s(x) = E + \frac{\sigma^2/2\nabla^2\psi(x)}{\psi(x)}$. The ground state of the wave-function, in the case of the QIPF formulation, is already probabilistically defined as an expression of the empirical KME given by $\psi(t) = \sqrt{p(t)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} k(x_i, t)}$. This information leads to the following summary of the QIPF state extraction procedure in the following postulate.

Postulate 1 (Extraction of QIPF Energy Modes) *Consider the QIPF of the data samples x as $V_s(x) = E + \frac{\sigma^2/2\nabla^2\psi(x)}{\psi(x)}$ with the associated ground state wave-function given by $\psi(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} k(x_i, t)}$. The approximate higher order energy modes of $\psi(x)$ can be extracted by projecting the ground state wave-function into the corresponding order Hermite polynomial given by $\psi_k(x) = H_k^*(\psi(x))$, where H_k^* denotes the normalized k^{th} order Hermite polynomial, normalized so that $H_k^* = \int_{x=-\infty}^{\infty} e^{-x^2}[H_k(x)]^2 dx = 1$. This leads to the evaluation of the higher order QIPF states as*

$$\begin{aligned} V_s^k(x) &= E_k + \frac{\sigma^2/2\nabla^2 H_k^*(\psi(x))}{H_k^*(\psi(x))} \\ &= E_k + \frac{\sigma^2/2\nabla^2\psi_k(x)}{\psi_k(x)} \end{aligned} \tag{15}$$

where k denotes the order number and E_k denotes the corresponding eigenvalues of the various modes and is given by

$$E_k = -min\frac{\sigma^2/2\nabla^2\psi_k(x)}{\psi_k(x)} \tag{16}$$

Algorithm 1 Quantum decomposition of IPF

Input:
x: Signal
σ: Kernel width
m: Number of quantum modes
Initialization:
ψ: Wave-function
$\psi^1, \psi^2, ..., \psi^m$: Wave-function Hermitian embeddings
$V_s^1, V_s^2, ..., V_s^m$: QIPF modes
$E^1, E^2, ..., E^m$: Eigenvalue of each mode
Computations:
for $i = 1$ to length(x) **do**
  $\psi = 0$
  for $j = 1$ to i **do**
    $\psi \leftarrow \psi + e^{-\frac{(x_i - x_j)^2}{2\sigma^2}}$
  end for
  $\psi_i \leftarrow \sqrt{mean(\psi)}$

  $[\psi_i^1, \psi_i^2, ..., \psi_i^m] \leftarrow HermiteProjections(\psi_i)$
  $[\nabla^2\psi_i^1, ..., \nabla^2\psi_i^m] \leftarrow Laplacians$

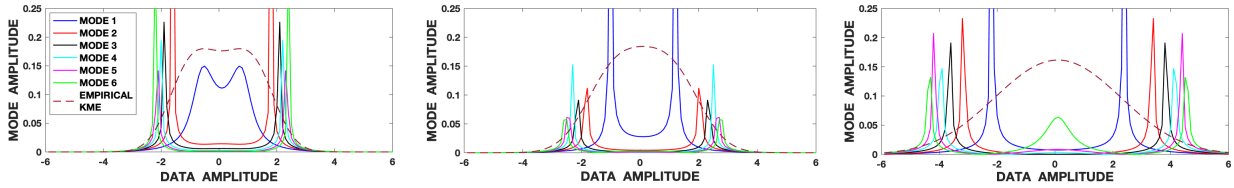
  for each mode k **do**
    $E_i^k = -\min_{q=1...i} \frac{\sigma^2/2\nabla^2\psi_q^k}{\psi_q^k}$

    $V_{s(i)}^k = E_i^k + \frac{\sigma^2/2\nabla^2\psi^k}{\psi^k}$
  end for
end for

The extracted modes of the data QIPF given by $V_s^k(x)$ are thus stochastic functionals depicting the different moments of *potential energy* of the data at any point x. This is different from the IPF formulation of (6) because $V_s^k(x)$ is an energy based metric resembling the potential energy operator in a quantum harmonic oscillator at various energy levels (Eigenstates).

We provide a summary of the framework in terms of a pseudocode in algorithm 1. As a pedagogical example, we show how the different QIPF modes get localized in the space of a sine wave signal, which represents one of the most fundamental dynamical systems.

We generated 3000 samples of a 50 Hz sine wave signal using a sampling frequency of 6000 samples per second to mimic a continuous signal. The signal was also normalized to zero mean and unit standard deviation. We used all 3000 samples as centers to construct the wave-function given by (8) and then evaluated it at each point in the data space range $x = (-6, 6)$ using a step size of 0.1. We then evaluated the Hermite projections of the wave-function value at each point to subsequently extract 6 QIPF modes using the formulation given by (15). This was done for three different kernel widths whose corresponding QIPF plots (represented by solid color lines) are shown in fig. 2. The dashed line represents the empirical KME estimate (or simply the IP) given by $p(x) = \psi^2(x) = \frac{1}{N}\sum_{i=1}^{N} \kappa(x, x_i)$, which essentially gives an estimate of the data PDF. All plots were



(a) Kernel width = 0.6 (b) Kernel width = 1 (c) Kernel width = 1.8

Figure 2: Analysis of mode locations of the sine wave in the space of data using different kernel widths. Solid colored lines represent the different QIPF modes. Dashed line represents the empirical KME (IP).

normalized for easier visualization. An important property of the extracted QIPF modes that can be observed from the plots is that, regardless of the kernel width, they consistently emphasize the more uncertain regions of the data space closer to the tails of the data PDF. One can observe the significant increase in the density (or clustering) of the extracted QIPF modes as one moves farther away from the mean ($x = 0$) and towards the PDF tails for the different kernel widths, thus demonstrating a greater emphasis of the QIPF modes on more uncertain regions of the signal. Furthermore, we observe here that the modes appear sequentially based on their orders, with the lower order modes emphasizing regions closer to the mean and the higher order modes clustering together at the PDF tails.

5 EXPERIMENTS

We ran simulations to evaluate the performance of the QIPF framework in the applications of change point detection and whole time series clustering. For change point detection, we compared the performance of the QIPF framework with the online version of Bayesian change point detector (Adams & MacKay, 2007) and Kullback Leibler divergence measure. The reason for choosing these algorithms is that they fall under the same taxonomy of methods as the QIPF framework (i.e. non-parametric, online and unsupervised). For whole time series clustering application, we chose to compare the QIPF uncertainty features with those of discrete wavelet transform (DWT) on a subset of the VidTIMIT speaker recognition corpus. All simulations were performed using python 3.7. We first provide a brief overview of the datasets used for each application before delving into the results and related analysis.

5.1 DATASETS

For change point detection, artificial datasets were generated and change points were inserted manually at particular intervals to cause the statistical drifts. Two datasets were generated using an auto-regressive model and change points were inserted to simulate mean jumps and variance scaling. The same datasets are also used in (Takeuchi & Yamanishi, 2006). For time series clustering, we use the VidTIMIT dataset (Sanderson & Lovell, 2009).

5.1.1 Mean jumps

For simulating a time series with mean jumps at regular intervals of time, we synthesized 5000 samples from the following auto-regressive model.

$$y(t) = 0.6y(t-1) - 0.5y(t-2) + \epsilon_t. \qquad (17)$$

Here ϵ_t represents the Gaussian noise with mean μ and standard deviation set as 1.5. We set the initial values as $y(1) = y(2) = 0$. We insert a change point at every 100 time steps by setting the noise mean μ at time t as

$$\mu_N = \begin{cases} 0 & \text{for } N = 1 \\ \mu_{N-1} + \frac{N}{16} & \text{for } N = 2, ..., 49 \end{cases} \qquad (18)$$

Here N is a natural number set such that $100(N-1) + 1 \leq t \leq 100N$. The idea of synthesizing such a dataset is to create drifts in the data without the exact change points being visible to the human eye. This creates a challenging detection task for algorithms.

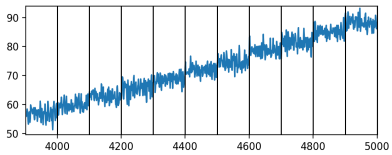
5.1.2 Variance jumps

We use the same auto-regressive model as Dataset 1, but here a change point is inserted at every 200 time steps by setting the noise standard deviation μ at time t as

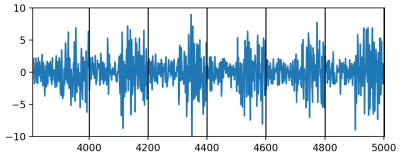
$$\sigma = \begin{cases} 1 & \text{for } N = 1 \\ ln\left(e + \frac{N}{4}\right) & \text{for } N = 2, ..., 49 \end{cases} \qquad (19)$$

5.1.3 VidTIMIT

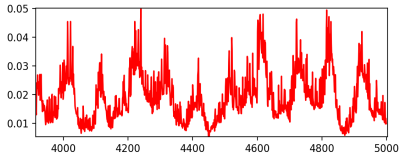
For the purpose purpose of time series clustering, we tested the framework on the subset of the VidTIMIT



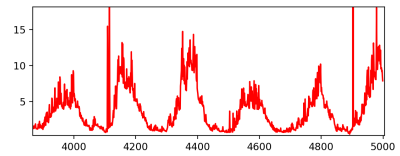
(a) Mean Jump Data



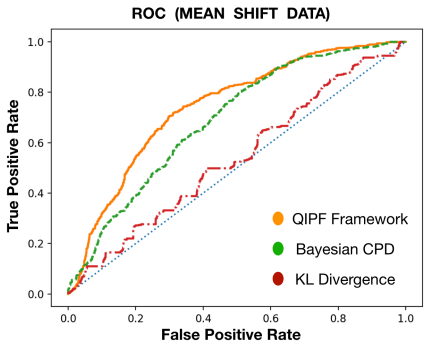
(b) Variance Jump Data



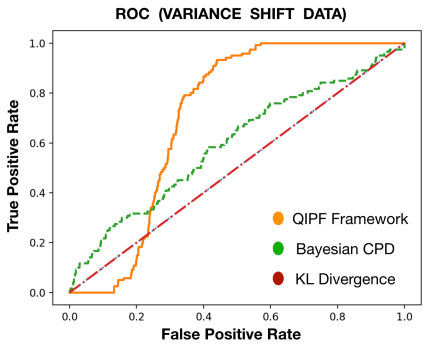
(c) Std. Dev. of 10 QIPF modes (mean jump)



(d) Std. Dev. of 10 QIPF modes (variance jump)



(e) Std. Dev. of 10 QIPF modes (mean jump)



(f) Std. Dev. of 10 QIPF modes (variance jump)

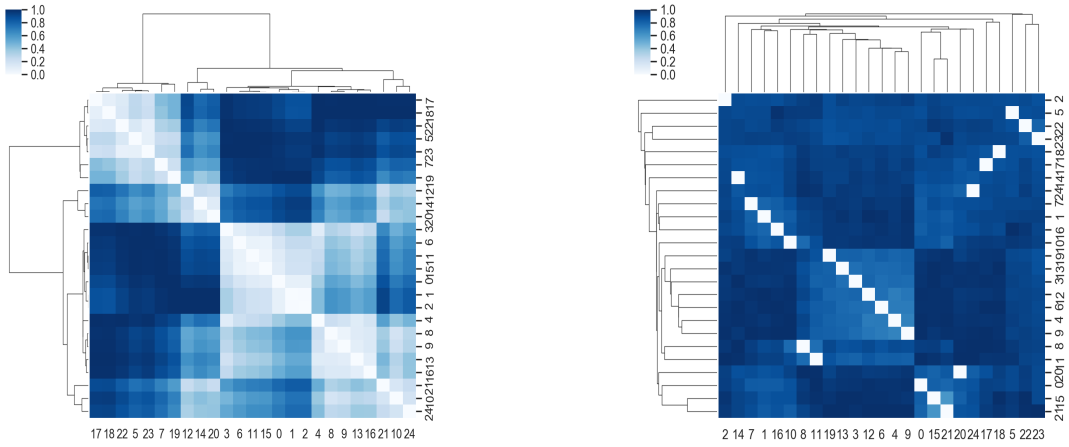
Figure 3: Last 1000 samples of drift datasets (top row), their corresponding QIPF mode standard deviations measured at each point (middle row) and corresponding the ROC curves (bottom row) for different methods measured in the range of 2000-3000 samples for both datasets. Black vertical lines (in the top row) mark the actual change points.

dataset which is a speaker recognition dataset that consists of voice recordings of 43 different speakers (Sanderson & Lovell, 2009). The dataset is made more challenging by the fact that there are also head movements involved while recording the data. For our experiments, we randomly chose 5 different speakers, each having 5 different voice recordings. Thus we worked with 25 time series datasets in total with the goal of clustering (unsupervised) them into their corresponding speaker classes. We downsampled each signal using a rate of 20 and chose the only middle 2000 samples for our experiments.

5.2 RESULTS

5.2.1 Change Point Detection

We implemented the QIPF framework on the synthesized datasets (after z-normalizing them) in an online manner using a fixed window length of 50 samples (kept short to make the framework more sensitive towards detecting changes). This means that the QIPF mode values at each point were evaluated only with respect to the previous 50 data points. We extracted the first 10 QIPF modes at every sample using a data-based bandwidth of 20 times Silvermans rule of thumb. To detect changes, we measured the standard deviation of the extracted QIPF modes at each sample thereby quantifying uncertainty at those points with respect to previous samples. Our conjecture is that the points where the signal characteristics change would exhibit increased uncertainty with respect to information field created by the previous samples leading to increased variations in extracted energy modes. For demonstration, the last 1000 samples of the both data sets are shown in fig. 3 with the change points marked with black vertical lines. The corresponding standard deviation values of the 10 extracted QIPF modes at each sample are also shown in fig. 3. We see here that, in both datasets, the peaks of the standard deviations of the QIPF modes match the actual change points. Furthermore, the ROC curves by measuring the true positive rates and false positive rates associated with the three methods over a range of thresholds from 2000-3000 samples of the datasets show the QIPF framework to have a significant better performance.



(a) QIPF framework

(b) Discrete Wavelet Transform

Figure 4: Heat-maps and corresponding dendograms representing pair-wise feature vector distances (with each feature vector representing one of the 25 voice files) of the QIPF framework (left) and the discrete wavelet transform (right).

5.2.2 Time Series Clustering

For time series clustering, we implemented the QIPF framework and discrete wavelet transform on a subset of VidTIMIT data as described earlier. Our goal here is to represent each time series signal using features that would maximize the distance between the different speaker classes and minimize the intra class distances. We chose DWT as the baseline for comparison since it has been established as a powerful unsupervised feature extraction method that uses both spatial and frequency components of the signals to extract their useful properties (Heil & Walnut, 1989b). We implemented the QIPF framework by extracting 50 modes corresponding to the samples of each time series signal. For each sample, we recorded which mode dominated (had the highest value) and thereafter characterized a histogram vector associated with each signal that characterized the frequency of domination of the different modes. All experiments were performed using z-normalized data and a kernel width equal to 40 times Silvermans ideal value. For DWT, we implemented maximum level decomposition using Debauchy-2 wavelet to extract the coefficients. Hence, we have the domination frequency of QIPF modes that represent the feature vector associated with each signal on one hand and the DWT coefficients on the other. To compare the quality of features extracted, we computed the pair-wise euclidean distances between the features of different signals for each method in order to ascertain how close the features of each signal are to its true class. The related heat maps representing relative euclidean distances between the different signals are shown in fig. 4. Further, hierarchical clustering was also performed on the set of pair-wise distances represented by the heatmaps to determine the quality of the clustering induced by QIPF and DWT representations. One can notice here from the euclidean distance heat-maps that the QIPF framework produces features having significantly better class discrimination properties. This is also evident in the hierarchical clustering characteristics.

Conclusion

In this paper, we introduced a recently proposed information theoretic framework for the purpose of characterizing time series data. We specifically applied the framework as a sensitive uncertainty feature extraction method for the applications of change point detection and time series clustering. We demonstrated through real world and synthetic datasets that out framework performs significantly better than established related methods for both applications. We intend to explore the capabilities of the QIPF framework in a more detailed manner in the future.

References

Adams, R. P., & MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering–a decade review. *Information Systems*, *53*, 16–38.

Ahn, K., Choi, M., Dai, B., Sohn, S., & Yang, B. (2018). Modeling stock return distributions with a quantum harmonic oscillator. *EPL (Europhysics Letters)*, *120*(3), 38003.

Aminikhanghahi, S., & Cook, D. J. (2016). A survey of methods for time series change point detection. *Knowledge and Information Systems*, *51*, 339-367.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, *68*(3), 337–404.

Berndt, D. J., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *Kdd workshop* (Vol. 10, pp. 359–370).

Bracewell, R. N., & Bracewell, R. N. (1986). *The fourier transform and its applications* (Vol. 31999). McGraw-Hill New York.

Fine, S., Singer, Y., & Tishby, N. (1998). The hierarchical hidden markov model: Analysis and applications. *Machine learning*, *32*(1), 41–62.

Griffiths, D. J., & Schroeter, D. F. (2018). *Introduction to quantum mechanics*. Cambridge University Press.

Hamilton, J. D. (1994). *Time series analysis* (Vol. 2). Princeton New Jersey.

Heil, C. E., & Walnut, D. F. (1989a). Continuous and discrete wavelet transforms. *SIAM review*, *31*(4), 628–666.

Heil, C. E., & Walnut, D. F. (1989b). Continuous and discrete wavelet transforms. *SIAM review*, *31*(4), 628–666.

Kakizawa, Y., Shumway, R. H., & Taniguchi, M. (1998). Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association*, *93*(441), 328–340.

Košmelj, K., & Batagelj, V. (1990). Cross-sectional approach for clustering time varying data. *Journal of Classification*, *7*(1), 99–109.

Kumar, M., Patel, N. R., & Woo, J. (2002). Clustering seasonality patterns in the presence of errors. In *Proceedings of the eighth acm sigkdd international conference on knowledge discovery and data mining* (pp. 557–563).

Liao, T. W. (2005). Clustering of time series dataa survey. *Pattern Recognition*, *38*(11), 1857 - 1874. doi: https://doi.org/10.1016/j.patcog.2005.01.025

Manuca, R., & Savit, R. (1996). Stationarity and nonstationarity in time series analysis. *Physica D: Nonlinear Phenomena*, *99*(2), 134 - 161. doi: https://doi.org/10.1016/S0167-2789(96)00139-X

Meng, X., Zhang, J.-W., & Guo, H. (2016). Quantum brownian motion model for the stock market. *Physica A: Statistical Mechanics and its Applications*, *452*, 281–288.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, *10*(1-2), 1–141.

Nielsen, F. (2013). An information-geometric characterization of chernoff information. *IEEE Signal Processing Letters*, *20*(3), 269–272.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, *33*(3), 1065–1076.

Principe, J. C. (2010). *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media.

Principe, J. C., Xu, D., Fisher, J., & Haykin, S. (2000). Information theoretic learning. *Unsupervised adaptive filtering*, *1*, 265–319.

Rényi, A., et al. (1961). On measures of entropy and information. In *Proceedings of the fourth berkeley symposium on mathematical statistics and probability, volume 1: Contributions to the theory of statistics.*

Rish, I., et al. (2001). An empirical study of the naive bayes classifier. In *Ijcai 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, pp. 41–46).

Sanderson, C., & Lovell, B. C. (2009). Multi-region probabilistic histograms for robust and scalable identity inference. In *International conference on biometrics* (pp. 199–208).

Scholkopf, B., & Smola, A. J. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Singh, R., & Principe, J. C. (2020). Towards a kernel based physical interpretation of model uncertainty. *arXiv preprint arXiv:2001.11495*.

Takeuchi, J.-i., & Yamanishi, K. (2006). A unifying framework for detecting outliers and change points from time series. *IEEE transactions on Knowledge and Data Engineering*, *18*(4), 482–492.

Tierney, J. (1980). A study of lpc analysis of speech in additive noise. *IEEE Transactions on Acoustics, speech, and signal processing*, *28*(4), 389–397.

Van Erven, T., & Harremos, P. (2014). Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, *60*(7), 3797–3820.