Kernel Representation Learning for Time Sequence: Algorithm, Theory, and Applications

Kunpeng Xu
Department of Computer Science, University of Sherbrooke
Sherbrooke, Quebec, Canada
kunpeng.xu@usherbrooke.ca

Abstract

Time sequences are essential in fields such as finance, healthcare, and environmental science, where understanding temporal dependencies and making accurate predictions are crucial. These sequences often exhibit complexities like nonlinearity, noise, and concept drift. Traditional models struggle to capture the intricate dynamics of multivariate and co-evolving sequences, particularly in contexts where relationships between variables shift unpredictably. This thesis introduces a range of Kernel Representation Learning (KRL) methodologies to address these challenges. We develop kernel self-representation learning to capture the temporal dependencies and hidden structures, while identifying concept drift in co-evolving sequences. Additionally, we explore theoretical connections between KRL and advanced deep-learning models. The proposed methods are validated through real-world applications, showing improvements in predictive accuracy, interpretability, and robustness.

Introduction

Time series analysis is crucial in fields like finance, healthcare, and meteorology. The complexities of time series data—such as nonlinearity, noise, and concept drift—make accurate modeling and prediction challenging. Recent deep learning advances (Xu, Chen, and Wang 2024a,b), particularly Transformer-based architectures, have improved forecasting by capturing long-range dependencies. However, most models focus on prediction accuracy at the expense of interpretability.

This lack of interpretability is a concern in many fields. In finance, for instance, shifts in market regimes and nonlinear relationships are as important as prediction accuracy. Detecting and understanding these shifts is vital for adapting to changing market conditions. Similarly, in healthcare, understanding a model's reasoning is crucial, as clinical decisions can significantly impact patient outcomes.

An additional challenge arises with co-evolving time series, where multiple time series exhibit interdependent behavior over time. In such cases, understanding the interactions and dynamics among these series becomes crucial for both prediction and interpretation. A more holistic approach is required – one that not only forecasts future values but also explains the underlying behaviors and relationships that govern the evolution of these time series.

Therefore, the main research question addressed in my thesis is: **RQ** – *Can we model co-evolving or large-scale time series as an* ***ecosystem*** 🌍 *and develop a principled approach to forecast both future patterns and series values, while balancing predictive power and interpretability?*

An Incremental Research Plan

We address the challenges from a kernel perspective - a method to model complex, nonlinear relationships within data. Following an incremental strategy, the research will be guided by the following sub-research questions (SRQ):

SRQ.1 - *How can we adaptively learn kernel representations to capture temporal dependencies and hidden structures in high-dimensional, nonlinear time series?*
This question focuses on developing a flexible KRL framework that adaptively learns kernel representations to model both temporal patterns and hidden structures in evolving, nonlinear time series. Additionally, it seeks to design a domain-agnostic segmentation approach to address the challenge of determining the optimal segmentation or window size for capturing diverse temporal patterns.

SRQ.2 - *How can KRL model co-evolving time series as an ecosystem, and effectively detect and track concept drift?*
This question explores modeling time series as an ecosystem through KRL, leveraging self-representation properties to uncover hidden relationships and detect shifts in data distributions over time (concept drift), ensuring the model remains adaptable in dynamic environments.

SRQ.3 - *What are the theoretical connections between KRL and advanced deep learning methods, such as Transformer architectures and the latest Kolmogorov-Arnold Networks?*
This question investigates the theoretical links between KRL and these advanced models, aiming to enhance the modeling of long-range dependencies and capture complex functional relationships in time series data.

SRQ.4 - *How can KRL be applied to real-world scenarios?*
Validating the model's ability to capture nonlinear interactions and regime shifts is crucial in real-world contexts such as financial markets, bioinformatics, and healthcare.

SRQ.5 - *Can KRL generate high-quality series datasets?*
This question addresses the challenge of limited access to real-world time series data, particularly in finance and

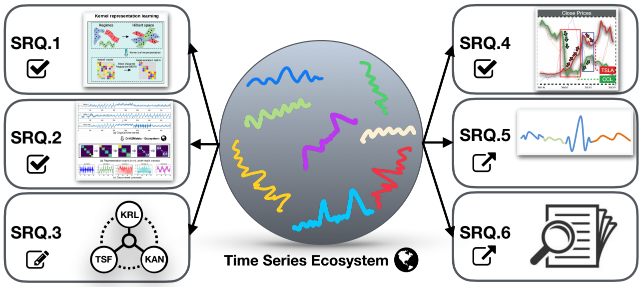

Figure 1: Key components of this thesis. ☑ = completed, ✎ = partially completed, ⧉ = not completed.

healthcare, while ensuring privacy by exploring KRL's potential for generating high-quality synthetic datasets.

SRQ.6 – *How can a survey on time series representation learning capture the field's state and guide future research?* To conclude the doctoral research, a comprehensive survey will summarize key advancements, methodologies, and challenges in time series representation learning, offering insights that can guide future research and development.

Current Works and Results

SRQ.1 We have developed Data-driven Kernel Learning (Xu, Chen, and Wang 2022a, 2025b) and Multiple Kernel Learning method (Xu, Chen, and Wang 2022b) that adaptively learn kernel representations and integrate multiple kernels from the data itself. Additionally, we proposed a Twin Learning Framework for Domain-Agnostic Time Series, which adaptively identifies behaviors and segmentations while learning the transitions.

SRQ.2 We introduced a kernel induced self-representation model (Xu, Chen, and Wang 2025a) that establishes a new theoretical paradigm in co-evolving time series analysis, enhancing our ability to capture hidden structures and track concept drift.

SRQ.3 We have made partial progress in deriving theoretical connections between KRL and self-attention in Transformer models. Specifically, we found that the Transformer kernel can be approximated as a dot product between spectral feature maps.

SRQ.4 We extended our work to real-world applications, including ecosystem-based modeling for financial time series (Xu et al. 2024a,b; Chen et al. 2022b,a) and DNA/RNA categorical sequences (Xu, Chen, and Wang 2022b). In these applications, we validated the proposed models by uncovering nonlinear interactions, identifying regime shifts, and demonstrating significant improvements in both predictive accuracy and model interpretability.

Future Works

Leading up to the DC Workshop, we will continue refining the theoretical connections between Kernel Representation Learning (KRL), Transformers, and Kolmogorov-Arnold Networks (KAN) as part of **SRQ.3**. For **SRQ.4**, I plan to extend the application domain to include healthcare time series, further validating the proposed models in this critical real-world context. Regarding **SRQ.5**, we will develop a new kernel-based diffusion model to generate a public synthetic time series dataset. This dataset will address the challenges of limited access to real-world data, ensuring privacy while adapting to distribution shifts in the time series. For **SRQ.6**, the goal is to complete a comprehensive survey paper before graduation, consolidating key advancements, methodologies, and challenges in the research process. This survey will offer valuable insights, highlight emerging trends, and guide future developments in the field.

To further promote collaboration and innovation, we will develop up-to-date websites for both the dataset generated in **SRQ.6** and the survey work in **SRQ.7**. These platforms will serve as resources for researchers interested in time series representation learning and public datasets, expanding the impact of our findings and fostering ongoing collaboration.

References

Chen, R.; Sun, M.; Xu, K.; Patenaude, J.-M.; and Wang, S. 2022a. Clustering-based cross-sectional regime identification for financial market forecasting. In *International Conference on Database and Expert Systems Applications*. Springer.

Chen, R.; Xun, K.; Patenaude, J.-M.; and Wang, S. 2022b. Dynamic Cross-sectional Regime Identification for Financial Market Prediction. In *2022 IEEE 46th Annual Computers, Software, and Applications Conference*. IEEE.

Xu, K.; Chen, L.; Patenaude, J.-M.; and Wang, S. 2024a. Kernel Representation Learning with Dynamic Regime Discovery for Time Series Forecasting. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer.

Xu, K.; Chen, L.; Patenaude, J.-M.; and Wang, S. 2024b. RHINE: A Regime-Switching Model with Nonlinear Representation for Discovering and Forecasting Regimes in Financial Markets. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, 526–534. SIAM.

Xu, K.; Chen, L.; and Wang, S. 2022a. Data-driven Kernel Subspace Clustering with Local Manifold Preservation. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, 876–884. IEEE.

Xu, K.; Chen, L.; and Wang, S. 2022b. A Multi-view Kernel Clustering framework for Categorical sequences. *Expert Systems with Applications*, 197: 116637.

Xu, K.; Chen, L.; and Wang, S. 2024a. KAN4Drift: Are KAN Effective for Identifying and Tracking Concept Drift in Time Series? In *NeurIPS Workshop on Time Series in the Age of Large Models*.

Xu, K.; Chen, L.; and Wang, S. 2024b. Kolmogorov-Arnold Networks for Time Series: Bridging Predictive Power and Interpretability. *arXiv preprint arXiv:2406.02496*.

Xu, K.; Chen, L.; and Wang, S. 2025a. Drift2Matrix: Kernel-Induced Self Representation for Concept Drift Adaptation in Co-evolving Time Series. *arXiv preprint arXiv:2501.01480*.

Xu, K.; Chen, L.; and Wang, S. 2025b. Towards Robust Nonlinear Subspace Clustering: A Kernel Learning Approach. *arXiv preprint arXiv:2501.06368*.