On analyzing time series: Clustering, causal discovery, and root cause identification

ABSTRACT

Time series analysis refers to the study of the dependence among observations at different points in time. What distinguishes time series analysis from general multivariate analysis is the temporal order imposed on the observations. Over the years, research on analyzing time series has presented challenges and methods for uncovering the underlying patterns, relationships, and dynamics that govern complex systems. Recently, the application of machine learning (ML) for use in the tasks of classification, forecasting, intervention, and making predictions on data, especially temporal data, has become popular. This Ph.D. research focuses on techniques from both machine learning and statistical methods to address some challenges and extract meaningful insights from time series. The research is structured around three interconnected areas of analysis: classification through robust clustering of noisy data, forecasting by discovering causality in non-linear systems, and intervention analysis by leveraging causal models.

In Chapter 2, we present a *robust clustering* to identify and group similar patterns within noisy datasets. Clustering enables the grouping of data into meaningful subsets, which can then be further analyzed more effectively. Traditional clustering methods like K-means and K-medoids iteratively group raw-data into similarity classes depending on their relative distances. These classical clustering algorithms work with raw-data and are not designed to be robust to uncertain/noisy data. However, data is naturally and inherently affected by the random nature of the physical generation process and measurement inaccuracies, sampling discrepancy, outdated data sources, or other errors, making it prone to noise/uncertainty. To this challenge, we propose a novel statistical metric known as the expectation distance (ED) of random variables that is demonstrated to be a statistical distance measure. Utilizing this newly proposed metric alongside the well-known 2-Wasserstein (W_2) distance, we develop noise-robust clustering algorithms that operate over data distributions rather than raw data points. By extending traditional K-means and K-medoids clustering algorithms with these proposed statistical metrics, our approach proves to be more effective in handling noisy data. Our research shows that while the W_2 distance relies only on marginal distributions and ignores the correlation information, the proposed ED metric captures this correlation, leading to superior noise-robust clustering results.

Chapter 3 of this research is centered on *causality*, specifically causal discovery, a process that seeks to uncover the causal relationships between variables in time series data. Understanding these causal relationships is essential not only for accurately modeling the data but also for making reliable predictions about future behavior, which is a step further from clustering. This part of the research emphasizes the Granger Causality to uncover these causal connections. To address the limitations of traditional linear Granger Causality methods, we introduce a novel framework called NeuroKoopman Dynamic Causal Discovery (NKDCD). The Koopman operator theory inspires this framework and harnesses the computational power of neural networks to learn the Koopman basis functions automatically. These basis functions aid in lifting the non-linear dynamics inherent in time series data

into a higher-dimensional space where they can be more effectively analyzed using linear techniques. NKDCD employs an autoencoder architecture that facilitates this transformation, enabling the application of Granger causality in non-linear settings. Our results show that NKDCD outperforms some existing non-linear Granger causality methods, offering a more robust process for discovering causal relationships in complex time series data. This research enhances our understanding of the underlying phenomena. It provides a foundation for developing more effective intervention analysis and decision-making processes.

In Chapter 4, the research focuses on the domain of *root cause analysis*, specifically intervention analysis, by focusing on the causal variables influencing hydrogen bonds in molecular dynamics simulations. Here, we leverage the statistical causal models to determine the "root causes" of changes observed in the joint and conditional probability distributions following a deviation in the time series. Specifically, we design a causal model that employs a dynamic Bayesian network (DBN) tailored for molecular dynamics applications. This model facilitates a detailed root cause analysis, enabling the identification of fundamental causes behind observed deviations, such as the formation or separation of hydrogen bonds. The ability to pinpoint these causes is crucial in fields like chemistry and materials science, where understanding the dynamics at the molecular level can lead to significant advancements in the development of new materials and drugs.

Throughout this research, the proposed methodologies are evaluated using both synthetic and real-world datasets to assess their implementation, performance, and robustness. By integrating these advanced techniques into a cohesive analytical framework, this research offers significant contributions to the field of time series analysis.