



# VCU

Virginia Commonwealth University  
VCU Scholars Compass

---

Theses and Dissertations

Graduate School


---

2024

## Developing Machine Learning and Time-Series Analysis Methods with Applications in Diverse Fields

Muhammed Aljifri  
*Virginia Commonwealth University*

Follow this and additional works at: <https://scholarscompass.vcu.edu/etd>

 Part of the [Chemical Engineering Commons](#), [Data Science Commons](#), [Mathematics Commons](#), [Operations Research, Systems Engineering and Industrial Engineering Commons](#), [Risk Analysis Commons](#), and the [Statistics and Probability Commons](#)

© Muhammed Aljifri

---

Downloaded from

<https://scholarscompass.vcu.edu/etd/7691>

This Dissertation is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact [libcompass@vcu.edu](mailto:libcompass@vcu.edu).

©Muhammed Aljifri, May 2024

All Rights Reserved.



DISSERTATION ON DEVELOPING MACHINE LEARNING AND TIME-SERIES  
ANALYSIS METHODS WITH APPLICATIONS IN DIVERSE FIELDS

A Dissertation submitted in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy at Virginia Commonwealth University.

by

MUHAMMED ALJIFRI

B.S., King Abdulaziz University, KSA - 2011

M.S., Western Michigan University - 2018

Director: Yanjun Qian,

Assistant Professor, Department of Statistical Sciences and Operations Research

Virginia Commonwealth University

Richmond, Virginia

May, 2024



## Acknowledgements

First, I would like to thank God for the boundless love, mercy, and grace, which have been my guiding lights in every challenging moment.

I want to express my deepest appreciation to my Ph.D. advisor, Dr. Yanjun Qian, whose guidance has been a beacon of light throughout this journey. His expertise and dedication to excellence have profoundly influenced my research approach, teaching me the value of persistence and precision. I sincerely thank my committee members, Dr. QiQi Lu, Dr. Ye Chen, and Dr. Abdel-Salam Gomaa, for their invaluable contributions to my Ph.D. work. I am particularly thankful to Dr. QiQi Lu and Dr. Mo Li for the detailed feedback and enriching discussions that have notably improved the fourth chapter of my dissertation. Furthermore, the suggestions and feedback from Dr. Ye Chen have been instrumental in enhancing the quality of the third chapter.

I also want to express my heartfelt thanks to my parents, Zain and Safiyh, for their unwavering support and the incredible love they have shown me. I am truly blessed to be part of such a caring and supportive family.

I cannot express enough gratitude to my wife, Shahad, for her endless love and support, which have been my strength. My daughter, Aya, and my son, Ammar, have been a light in our lives, bringing unparalleled joy and laughter. Without the constant support of my family, I surely wouldn't be where I am today.

## TABLE OF CONTENTS

Chapter	Page
Acknowledgements . . . . .	iii
Table of Contents . . . . .	iv
List of Tables . . . . .	vi
List of Figures . . . . .	viii
Abstract . . . . .	xi
1 Introduction . . . . .	1
2 Machine Learning based Risk adjusted CUSUM control chart for Monitoring Readmission Rate following PTBD Catheter Placement . . . . .	6
2.1 Introduction . . . . .	6
2.2 Data and Methods . . . . .	10
2.2.1 Retrospective and Prospective Datasets . . . . .	10
2.2.2 RA CUSUM Chart . . . . .	11
2.2.3 Risk Prediction by Machine Learning . . . . .	13
2.2.3.1 Logistic Regression . . . . .	13
2.2.3.2 Tree-based Models . . . . .	14
2.2.4 Model Comparison for Risk Prediction . . . . .	15
2.2.5 RA CUSUM Chart Evaluation . . . . .	17
2.3 Results and Discussions . . . . .	18
2.3.1 Data Summary and Preliminary Analysis . . . . .	18
2.3.2 Predictive Model Comparison . . . . .	22
2.3.3 Control Charts Construction and Evaluation . . . . .	26
2.4 Conclusions . . . . .	32
3 Multilayer Modeling for Wide-Range Chemical Concentration Prediction in Spectroscopic . . . . .	33
3.1 Introduction . . . . .	33
3.1.1 High Dimensionality and Multicollinearity in Spectra Data . . . . .	34
3.1.2 Non-Linearity of Spectra Data . . . . .	36

3.2	Materials and Methods . . . . .	38
3.2.1	Materials and Instruments . . . . .	38
3.2.2	Spectral Data Analysis . . . . .	39
3.2.2.1	Principal Component Regression & Partial Least Squares algorithms . . . . .	39
3.2.2.2	Dynamical Layered Regression (DLR) Method . . . . .	43
3.2.2.3	Classified Layered Regression (CLR) Method . . . . .	44
3.2.2.4	Models Evaluation . . . . .	46
3.2.3	Parameter Tuning . . . . .	47
3.2.3.1	Tuning Number of Components . . . . .	48
3.2.3.2	Tuning Scaling Factor for DLR Method . . . . .	48
3.2.3.3	Tuning Number of Classes in the CLR Method . . . . .	49
3.3	Results and Discussion . . . . .	50
3.3.1	Data Summary and Preliminary Analysis . . . . .	50
3.3.2	Predictive Models Tuning . . . . .	52
3.3.3	Predictive Models Comparison . . . . .	59
3.4	Conclusions . . . . .	60
4	Multiple Changepoint Detection for Autocorrelated Ordinal Time Series . . . . .	63
4.1	Introduction . . . . .	63
4.1.1	Advancing from Single to Multiple Changepoint Detection . . . . .	63
4.1.2	Multiple Changepoint Detection Methods . . . . .	64
4.1.3	Multiple Changepoints in Categorical Time Series . . . . .	67
4.2	Methodology . . . . .	69
4.2.1	AOP Model . . . . .	69
4.2.2	Pairwise Likelihood Function for AOP Models . . . . .	72
4.2.3	Objective Function and Model Selection . . . . .	74
4.2.4	Genetic Algorithm . . . . .	75
4.2.5	Effective Number of Changepoints . . . . .	78
4.2.6	Evaluation Methods . . . . .	79
4.3	Simulation Studies . . . . .	80
4.3.1	Three Changepoints Setting (Up Down Up, $\kappa = 2$ ) . . . . .	81
4.3.2	Three Changepoints Setting (Up Down Up, $\kappa = 1$ ) . . . . .	82
4.3.3	Three Changepoints Setting (Up Down Up, $\kappa = 0.5$ ) . . . . .	84
4.3.4	Three Changepoints Setting (Up Up Up, $\kappa = 2$ ) . . . . .	89
4.3.5	Three Changepoints Setting (Up Up Up, $\kappa = 1$ ) . . . . .	91
4.3.6	Three Changepoints Setting (Up Up Up, $\kappa = 0.5$ ) . . . . .	92
4.3.7	Comparing Different Values of the Auto-correlation Parameter . . . . .	97



4.4 Los Angeles City AQI Data . . . . .	99
4.5 Conclusions . . . . .	102
5 Summary and Future work . . . . .	103
References . . . . .	105
Vita . . . . .	114

## LIST OF TABLES

Table		Page
1	Summary of the categorical characteristics in retrospective and prospective datasets. . . . .	19
2	Means of the numerical characteristics in retrospective and prospective datasets. . . . .	20
3	Preliminary analysis for the relationship between each characteristic and the readmission by the simple logistic regression. . . . .	21
4	The average and standard deviation (s.d.) of AUROC using 10-fold cross-validation for machine learning models with the subset in Section 2.3.2 and all variables. . . . .	24
5	Average run length ( $ARL$ ) for various $R_1$ and machine learning models estimated by bootstrapping from the retrospective and prospective dataset. $h^*$ is tuned for $ARL_0 \approx 500$ using the retrospective dataset. . . .	28
6	A example of the new range and number of observations in each layer for both DLR-PLS and DLR-PCR models. . . . .	57
7	Confusion matrix of PLS-DA model. . . . .	58
8	The concentrations range of High and Low classes. . . . .	59
9	RMSE values for the standard and proposed models . . . . .	60
10	Parameters of general setting. . . . .	81
11	Empirical proportions of the estimated number of changepoints, and the average distance (Up Down Up, $\kappa = 2$ ). The true value of $m$ is 3 in $Z_t$ . . . . .	82
12	Empirical proportions of the estimated number of changepoints and the average distance for the setting Up Down Up, $\kappa = 1$ . The true value of $m$ is 3 in $Z_t$ . . . . .	85

13	Differences from an effective number of changepoints for the setting Up Down Up, $\kappa = 1$ . . . . .	85
14	Empirical proportions of the estimated number of changepoints and average distance for the setting Up Down Up, $\kappa = 0.5$ . The true value of $m$ is 3 in $Z_t$ . . . . .	87
15	Differences from the effective number of changepoints for the setting Up Down Up, $\kappa = 0.5$ . . . . .	87
16	Empirical proportions of the estimated number of changepoints and the average distance in the setting Up Up Up, $\kappa = 2$ . The true value of $m$ is 3 in $Z_t$ . . . . .	89
17	Differences from the effective number of changepoints for the setting Up Up Up, $\kappa = 2$ . . . . .	90
18	Empirical proportions of the estimated number of changepoints and average distance for the setting Up Up Up, $\kappa = 1$ . The true value of $m$ is 3 in $Z_t$ . . . . .	91
19	Differences from the effective number of changepoints. For the setting Up Up Up, $\kappa = 1$ . . . . .	93
20	Empirical proportions of the estimated number of changepoints and average distance for the setting Up Up Up, $\kappa = 0.5$ . The true value of $m$ is 3 in $Z_t$ . . . . .	95
21	Differences from the effective number of changepoints for the setting Up Up Up, $\kappa = 0.5$ . . . . .	95
22	Average Distances for varies values of $\phi$ , for the setting Up Down Up, $\kappa = 2$ . . . . .	98

## LIST OF FIGURES

Figure	Page
1	The importance of all characteristics from the random forests model on the retrospective dataset. . . . . 22
2	The cross-validation prediction for readmitted vs. not readmitted groups for the retrospective model using different machine learning models. . . . . 25
3	The CUSUM chart without risk-adjustment for the combined dataset for $R_1 = 0.5$ . . . . . 29
4	The RA CUSUM chart by the logistical regression for the combined dataset for $R_1 = 0.5$ . . . . . 29
5	The RA CUSUM chart by the random forests for the combined dataset for $R_1 = 0.5$ . . . . . 30
6	The RA CUSUM chart by the GBM with 2000 trees for the combined dataset for $R_1 = 0.5$ . . . . . 30
7	The RA CUSUM chart by the GBM with 3000 trees for the combined dataset for $R_1 = 0.5$ . . . . . 31
8	UV-Vis spectral measurements. . . . . 51
9	Boxplots of the concentrations before (left) and after (right) the log transformation. . . . . 51
10	Sample of concentrations curves before the average (left with nine curves) and after the averaging (right with three curves). . . . . 53
11	Number of components vs RMSE values for PLS (left) and PCR (right). . . . . 55
12	$s$ values vs RMSE for DLR models . . . . . 56
13	Predictive accuracy comparison by all models. . . . . 59

14	Actual concentrations vs predicted values of PLS model. . . . .	61
15	Actual concentrations vs predicted values of PCR model. . . . .	61
16	Time series plots of latent variable ( $Z$ ) and categorical variable ( $X$ ) for the Up Down Up, $\kappa = 2$ setting. . . . .	83
17	Time series plots of latent variable ( $Z$ ) and categorical variable ( $X$ ) for the Up Down Up, $\kappa = 1$ setting. . . . .	84
18	Time series plots of latent variable ( $Z$ ) and categorical variable ( $X$ ) for the Up Down Up, $\kappa = 0.5$ setting. . . . .	86
19	Time series plots of latent variable ( $Z$ ) and categorical variable ( $X$ ) for the Up Up Up, $\kappa = 2$ setting. . . . .	90
20	Time series plots of latent variable ( $Z$ ) and categorical variable ( $X$ ) for Up Up Up, $\kappa = 1$ setting. . . . .	92
21	Time series plots of latent variable ( $Z$ ) and categorical variable ( $X$ ) for Up Up Up, $\kappa = 0.5$ setting. . . . .	94
22	Average distance with varies values of $\phi$ for the setting Up Down Up, $\kappa = 2$ . The true number of changepoints $m = 3$ in $Z_t$ . . . . .	98
23	Continuous ( $Z$ ) (Top) and the categorical ( $X$ ) daily AQI time series in Los Angeles from 2020 to 2022. . . . .	101

## **Abstract**

### DISSERTATION ON DEVELOPING MACHINE LEARNING AND TIME-SERIES ANALYSIS METHODS WITH APPLICATIONS IN DIVERSE FIELDS

By Muhammed Aljifri

A Dissertation submitted in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2024.

Director: Yanjun Qian,

Assistant Professor, Department of Statistical Sciences and Operations Research

This dissertation introduces methodologies that combine machine learning models with time-series analysis to tackle data analysis challenges in varied fields. The first study enhances the traditional cumulative sum control charts with machine learning models to leverage their predictive power for better detection of process shifts, applying this advanced control chart to monitor hospital readmission rates. The second project develops multi-layer models for predicting chemical concentrations from ultraviolet-visible spectroscopy data, specifically addressing the challenge of analyzing chemicals with a wide range of concentrations. The third study presents a new method for detecting multiple changepoints in autocorrelated ordinal time series, using the autoregressive ordered probit model in conjunction with a genetic algorithm. This technique is applied to the air quality index data for Los Angeles, aiming to detect significant changes in air quality over time.

# CHAPTER 1

## INTRODUCTION

Statistical techniques such as control charts and changepoint detection are widely utilized to analyze data patterns over time. These methods are frequently employed in quality control and process improvement settings to monitor and regulate system variability or process variability. Control charts help identify cases when a process is out of control and indicate a problem that needs to be addressed. Also, changepoint detection techniques are used to identify significant changes in the underlying patterns or trends of data. It helps to identify the points in time when a change occurs and estimate its magnitude and direction. Both control charts and changepoint detection use traditional models to model the relationships and calculate the probabilities. However, traditional models have shown drawbacks due to their limitations and strict assumptions. Recently, machine learning has emerged as a promising tool for improving the accuracy and efficiency of these methods. By leveraging large datasets and complex algorithms, machine learning models can identify patterns and changes in the data that traditional models might miss.

Through this dissertation, we aim to investigate and innovate in statistical analysis by leveraging the power of advanced techniques such as control charts, changepoint detection methods, and machine learning models. This dissertation develops three novel methods for improving statistical modeling and quality control by integrating machine learning methods and statistical process monitoring techniques. First, we developed risk-adjusted CUmulative SUM (CUSUM) control charts using machine learning models to monitor significant changes in hospital readmission rates. We

compared our proposed method with traditional linear models to evaluate its performance and effectiveness. Second, we introduced two innovative multi-layer models to address the non-linearities of chemical concentrations in spectroscopic data. These methods are applied in combination with two commonly used machine learning models to overcome the challenges of high dimensionality and multicollinearity found in spectral data. Finally, we developed a new technique for detecting changepoints that expands the use of multiple changepoint detection methods beyond continuous time series to autocorrelated categorical time series, using the autoregressive ordered probit model. A genetic algorithm addresses the searching challenge of detecting multiple changepoint configurations.

The first part of the dissertation introduced the utilization of a statistical approach for monitoring the readmission rate of hospitals. Readmissions are a costly issue for hospitals, as they can result in increased healthcare costs, longer hospital stays, and poorer patient outcomes [1]. Hence, control charts can monitor the readmission rate and identify when the rate is outside of its expected variability. CUSUM control charts are a type of control chart commonly used for monitoring readmission rates in healthcare. They are specifically designed to identify small or gradual changes in a process mean, which may be difficult to detect using other control charts [2]. To create a control chart for the readmission rate, it is essential to consider each patient's underlying risk level. This is necessary due to the potential impact of variations in patient characteristics on the effectiveness of CUSUM charts. Therefore, an adjustment can be made to the CUSUM method by factoring in the patient's risk level along with the risk associated with the surgical procedure. This modification is called risk-adjusted CUSUM charts [3]. Previous studies have primarily relied on linear logistic regression to establish a connection between a patient's individual risk and surgical outcomes. However, such studies have identified certain limitations



associated with using this particular model [4]. This dissertation proposes a novel risk-adjusted CUSUM chart that employs a machine-learning model. This new machine learning-based risk-adjusted CUSUM chart will be able to detect minor changes more effectively by leveraging the improved accuracy of probabilities derived from a powerful predictive model. In order to address the issue of overfitting, which arises from the varying degrees of complexity in machine learning models, we have devised an approach for comparing predictive models with different levels of complexity. By implementing this approach, we can compare the models' performance based on a standardized metric and identify the optimal model that balances complexity and accuracy.

In the second part of the dissertation, we proposed two novel multi-layer models to tackle various challenges in predicting chemical concentrations from spectroscopic data. These challenges include high-dimensional data with a large number of spectral features, nonlinear relationships between these features, and multicollinearity [5] [6] [7]. The two methods were employed to predict the concentrations of Nickel and Sulfate ions, i.e., NiSO<sub>4</sub> chemical. However, the concentration range of this chemical compound was extremely wide, leading to a nonlinear relationship between the features. To overcome this challenge, we developed dynamically layered regression and classified layered regression models, which help to predict concentrations based on local linearity instead of the global one. By using these multi-layer models, the accuracy of the concentration prediction was significantly improved compared to standard methods. Furthermore, these multi-layer models have been designed to complement two commonly used models in chemometrics: partial least squares and principal component regression. Both models effectively handle high-dimensional and multicollinear spectral data. Therefore, integrating the proposed methods further enhances the accuracy and reliability of the concentration prediction from the

spectroscopic data.

The third part of this dissertation presents a newly developed technique for detecting multiple changepoints in autocorrelated ordinal time series, a significant advancement in this field of study. This project addresses the shortcomings of traditional single changepoint detection methods, which often need to be revised for real-world phenomena where multiple changepoints are common [8]. The core of our method involves using the autoregressive ordered probit model, now enhanced with a pairwise likelihood function for more accurate parameter estimation. We also incorporate penalized likelihood methods for detecting changepoints and employ a genetic algorithm to improve computational efficiency in handling complex datasets. To validate our methods, we have conducted thorough simulation studies. Additionally, the practical utility of our approach is demonstrated through its application to real-world data, offering a concrete example of its effectiveness in analyzing complex time series data.

The remainder of this dissertation is organized as follows. Chapter 2 proposes a novel monitoring method to monitor hospital readmission rates. CUSUM control charts are specifically designed for this purpose, and a modification called risk-adjusted CUSUM charts factor in patient risk levels. Chapter 3 presents two multi-layer models designed for predicting chemical concentrations from spectroscopic data, addressing challenges such as high dimensionality and nonlinearities. The proposed methods were tested on the  $\text{NiSO}_4$  chemical and showed enhanced prediction accuracy over standard approaches. They are compatible with established machine learning techniques like partial least squares and principal component regression, offering a valuable extension to current analytical methods. Chapter 4 proposes a new method to detect multiple changepoints in autocorrelated categorical time series. The method uses a model selection technique with a penalty term on the number and locations

of changepoints and a penalized likelihood-based approach optimized by the genetic algorithm. The method is applied to real data to identify significant changes over time. In Chapter 5, we present a summary of our main contributions and explore potential future extensions of our work.

## CHAPTER 2

# MACHINE LEARNING BASED RISK ADJUSTED CUSUM CONTROL CHART FOR MONITORING READMISSION RATE FOLLOWING PTBD CATHETER PLACEMENT

### 2.1 Introduction

The quality of hospital care is considered one of the most critical aspects of the healthcare system, as poor quality of hospital care can lead to serious adverse consequences. Nicolay et al. [1] stated that only 55% of patients receive proper care in the USA, while between 44,000 – 98,000 patients died in hospitals due to preventable medical errors. One critical indicator of hospital care is the readmission rate, defined as the hospital admission in a short time (e.g., 30 days) after the original admission [9]. Besides the high readmission rate's negative impact on healthcare quality, the financial burden of readmissions causes healthcare systems to encounter substantial financial risk. For example, the Medicare Hospital Readmissions Reduction Program (HRRP) penalizes around 2,000 hospitals over \$280 million in one year [10].

In response to this challenge, the literature on quality engineering proposed methods to analyze, optimize, predict, and monitor the readmission rate by leveraging healthcare data sources. Statistical analysis is a popular tool to draw meaningful interpretations and significant associations with the readmission rate. For instance, Sarwar et al. [11] used a logistic regression model to investigate the causes of 30-day readmission for patients who underwent percutaneous transhepatic biliary drainage (PTBD) procedures. Using a multivariate logistic regression model, Pathak et al. [12] calculated the odds of being readmitted after revision Total Hip Arthroplasty

(rTHA) with different patient characteristics. Sarwar et al. [13] showed the association between 30-day readmission rates and 90-day mortality in a dataset with 3653 procedures in 12 different categories using chi-square tests and simple logistic regression models. In a retrospective cohort study, Panagiotou et al. [14] found that Medicare Advantage beneficiaries have higher 30-day readmission rates than traditional Medicare beneficiaries using a hierarchical logistic regression model. Those statistical tools are suitable for finding causes and factors associated with the increase in the readmission rate and performing the prediction task. However, when we want to monitor and improve the readmission rate, they have limitations in detecting and monitoring changes in a process [15]. Thus, statistical process control charts draw much attention to handling this task.

Statistical Process Control (SPC) techniques are widely used in the industrial field and healthcare for process monitoring and improving the quality of the procedure. One of the most effective SPC methods for monitoring the processes in healthcare applications is control charts, which can visualize stability and variability in automated processes over time [16]. Control charts were developed by Walter Shewart at the AT&T Bell laboratories in the early 1920s [17]. Those chronological graphs, which display the data process to better understand the process's variability, have been applied in a wide range of healthcare systems [2] [18]. For different types of data processes, several control charts, such as  $\bar{x}$  charts,  $p$  charts, and  $c$  charts, have been developed for data from normal, binomial, and Poisson distributions. Then, more advanced charts have been developed, such as the Exponential Weighted Moving Average (EWMA) [19] and CUSUM control chart [20]. These charts are known as control charts with memory, which indicates that they utilize both the previous and the current information to calculate the plotted statistics, making them less sensitive to outliers but good at detecting small but constant shifts [21]. Among charts with

memory, the CUSUM chart has received more attention in the medical field because of its intuitive formulation and the faster detection ability [22]. The CUSUM chart was first introduced by Page [23]. De Leval et al. [24] proposed one of the first studies that used the CUSUM chart in healthcare to monitor 104 consecutive neonatal switch operations between 1987 and 1993. Neuburger et al. [25] showed that CUSUM charts detected the changes in clinical performance rates faster than the Shewhart p-chart and EWMA chart.

The standard CUSUM charts perform well as long as the process is naturally homogeneous and there is no great variation among subjects. However, in the health care application, we must consider the baseline risk for subjects, i.e., patients, due to the heterogeneity that will affect the performance of CUSUM charts [26]. As stated by Sego et al. [27], without adjusting the CUSUM chart using postoperative risk for each patient, the outcomes will be confounded with the preexisting risk factors. For this reason, the CUSUM method needs to be adjusted by adding the patient's risk to monitoring the procedure's risk [3], known as risk-adjusted (RA) CUSUM charts. Li et al. [28] indicated that using RA CUSUM charts helps reduce the bias of the CUSUM outcomes due to heterogeneity among patients. Using a nonadjusted CUSUM chart for high-risk operations might produce misleading and less accurate outcomes than the risk-adjusted one. Grigg et al. [26] pointed out that the standard CUSUM chart indicated the unrealistic performance of surgeons in cardiac surgeries and suggested adjusting the risk in the chart to address this issue. Moreover, Novick et al. [15] showed the incremental advantages of RA CUSUM charts over the standard CUSUM charts using data that includes patients who underwent a coronary PROBIT grafting (CABG) procedure. Likewise, Steiner et al. [3] claimed faster detection of RA CUSUM charts than the standard CUSUM charts to monitor the surgical performance by adding the Parsonnet score [29] of each patient as a prior risk using data from

a UK center for cardiac surgery. At last, Rasmussen et al. [22] showed how RA CUSUM charts help detect 30-day mortality rates in 24 hospitals in Denmark.

In the RA CUSUM procedure, most studies use a logistic regression model to access the patient’s own risk based on his/her characteristics [3], [15], [22]. However, recent work highlighted the limitations when using the generalized linear model regarding the model’s restrictions and its prediction performance. For instance, Rossi et al. [4] proposed a risk-adjusted Bernoulli CUSUM (RA-B-CUSUM) chart to alleviate this impact of changes in the parameters of the risk distribution on the outcomes of RA CUSUM. Li et al. [28] introduced the varying-coefficient logistic regression (VCLR) as a nonparametric model to reduce the modeling bias of the logistic regression in the case of nonlinear relationships among patient characteristics. Moreover, other studies showed poor predictive performance of regression models compared to other methods, especially advanced machine learning models. Thanks to the increasing availability of healthcare data, machine learning models draw great attention in healthcare applications, such as predicting readmission and mortality rates after surgical procedures. In a systematic review study, Huang et al. [30] surveyed 25 studies that used machine learning models to predict hospital readmission rates and found Decision Trees and Random Forests were the most popular algorithms. Shin et al. [31] discovered machine learning models such as Random Forests and Support Vector Machines outperformed the traditional regression models for predicting heart failure readmission and mortality rates in a comparison study. Also, Futoma et al. [32] showed that Random Forests enjoy better predictive performance than the standard logistic regression in predicting early readmission rates using a dataset that includes around 3.3 million hospital admissions in New Zealand.

To further improve the performance of the RA CUSUM charts, we adopted machine learning models to estimate the probability of the risk when constructing the

chart. With the help of these powerful models, the proposed RA CUSUM charts will provide enhanced accuracy in detecting small changes in the readmission rate. To select the most suitable machine learning method, we developed a model selection criterion based on the area under the receiver operating characteristic curve (AUROC) and the cross-validation, which achieves a trade-off between the model’s bias and variance. At last, we applied the proposed charts to monitor the Percutaneous Transhepatic Biliary Drainage (PTBD) catheter placement procedures, detecting a possible decrease in the readmission rate after the improved post-procedure care paradigm.

## **2.2 Data and Methods**

### **2.2.1 Retrospective and Prospective Datasets**

We collected two readmission datasets from the electronic health record at the medical college of Virginia Commonwealth University (VCU). A total of 243 PTBD catheter placement procedures were recorded between January 2013 and May 2019 as the retrospective dataset. This dataset has 30 variables, including the primary diagnosis type, the indicator for benign vs. malignant disease, American Society of Anesthesiologists (ASA) classification score, lab test values prior to the procedure such as creatinine (Cr), liver transaminases alanine transaminase (ALT)/aspartate transaminase (AST), alkaline phosphatase (Alk Phos), total/direct bilirubin (T/D bili), white cell count (WBC), hemoglobin (Hgb), platelets (Plt), and international normalized ratio (INR). Moreover, the data records the patient’s characteristics, such as age, gender, whether the patient is insured, race, and the procedure’s characteristics, such as access position (Access), size in the French scale (Fr), and access type (Type). After that, the last variable indicates whether the patient was readmitted within 30 days after the procedure.



Relatively high rates of readmissions have been reported in the literature following the PTBD catheter placement [11]. The medical team at VCU hypothesized that post-procedure care might be inadequate and inconsistent for these patients. Hence, a new interdisciplinary standard of care paradigm was created for other patients based on best practice guidelines after 2019. The new procedural paradigm includes standardized documentation, dedicated education to interventional radiology (IR) nursing on drain management, standardized order set in electronic health records (EHR), ensuring care coordination consult and adequate supply for drain management post-discharge, as well as follow-up phone calls two and seven days post-discharge. Then, we collected a prospective dataset, including 67 PTBD procedures carried out after implementing the new procedural care paradigm. The prospective dataset includes the same set of variables in the retrospective dataset. In this study, we will design a new technique of RA CUSUM charts to identify where there is a significant reduction in the readmission rate after adopting the new procedural care paradigm for PTBD catheter placement.

### 2.2.2 RA CUSUM Chart

A standard CUSUM chart can be adjusted by the patient's prior risk  $p_t$ , estimated from his/her characteristics preoperatively. The characteristics are denoted as a vector  $X_t$  where  $t = 1, 2, \dots$  are the indices for the patients undergoing PTBD procedures. The response is a binary variable  $y_t$  to indicate whether the patient was readmitted within 30 days. The details of predicting  $p_t$  from  $X_t$  will be discussed in Section 2.2.3. When interested in detecting a change in the readmission rate, the hypotheses are based on the odds ratio of the readmission due to the procedure's risk [3]. In the RA CUSUM chart scenario, the odds ratio plays a crucial role in quantifying the relationship between the odds of readmission for patients undergoing

a specific medical procedure (e.g., PTBD procedures) and the odds of readmission for other patients. It is computed as the ratio of the odds of readmission for patients undergoing the procedure to the odds of readmission for those not undergoing the procedure. Then, the null and alternative hypotheses are written as follows:

$$H_0: \text{odds ratio} = R_0 \quad \text{vs.} \quad H_1: \text{odds ratio} = R_1, \quad (2.1)$$

where  $R_0$  indicates the current procedure performance and  $R_1$  corresponds to the new readmission odds ratio we want to detect. The null hypothesis ( $H_0$ ) assumes that the odds ratio is equal to a reference value ( $R_0$ ), typically representing the current performance level of the medical procedure in terms of readmission rates. Conversely, the alternative hypothesis ( $H_1$ ) posits a specific change in the odds ratio ( $R_1$ ) that the analysis aims to detect. This change could signify either an improvement or deterioration in the readmission rate following the procedure.  $R_0$  is usually set as 1 in default and  $R_1$  can be chosen in  $(0, 1)$  to detect a decrease in the readmission rate. For a patient with the prior risk  $p_t$  under  $H_0$ , the combined odds of being readmitted is  $R_0 p_t / (1 - p_t)$  and the corresponding probability is  $R_0 p_t / (1 - p_t + R_0 p_t)$ . Under  $H_1$ , the combined odds of being readmitted is  $R_1 p_t / (1 - p_t)$  and the corresponding probability is  $R_1 p_t / (1 - p_t + R_1 p_t)$ . Derived from those assumptions, the hypotheses in Equation (2.1) are tested using the following two log-likelihood ratio scores [3]:

$$W_t = \begin{cases} \log\left[\frac{(1-p_t+R_0 p_t)R_1}{(1-p_t+R_1 p_t)R_0}\right], & \text{if } y_t = 1, \log\left[\frac{(1-p_t+R_0 p_t)}{(1-p_t+R_1 p_t)}\right], \\ \text{if } y_t = 0. \end{cases} \quad (2.2)$$

$$Z_t = \max(0, Z_{t-1} + W_t), t = 1, 2, 3, \dots \quad (2.3)$$

where  $W_t$  is an assigned weight for each subject and  $Z_0 = 0$ . Equation (2.3) only assigns positive  $Z_t$ , producing a one-sided CUSUM chart. A one-sided chart only

focuses on the increasing trend of  $W_t$ , which means the alternative  $H_1$  becomes more likely than  $H_0$ . RA CUSUM charts will signal an alarm if the value of  $Z_t > h$ , indicating the process is in the out-of-control state  $H_1$ ; otherwise, it is in the in-control state  $H_0$ . The threshold  $h$  will be determined based on the targeted average running length (ALR) of the charts, which will be discussed in Section 2.2.5. In our study, we set  $R_1 = 0.5$  to detect halving of the odds of the readmission rate [3], and also show the chart performances when  $R_1 = 0.2$  or  $0.8$ .

### 2.2.3 Risk Prediction by Machine Learning

The readmission probability  $p_t$  related to the patient’s own risk plays a critical role in RA CUSUM charts. For cardiac surgery monitoring, a logistic regression based on the Parsonnet score [29] is a popular choice [3]. However, there is a lack of such existing work for the readmission of the PTBD procedures. Different types of machine learning models can be applied to obtain  $p_t$  from the preoperative characteristics  $X_t$ , and the main differences between these models lie in their underlying assumptions, the complexity of their decision boundaries, and the performance on different types of data [30], [33]. In this study, we consider three popular models: logistic regression [34], random forests [35], and gradient boosting machine [36], where the latter two belong to tree-based models.

#### 2.2.3.1 Logistic Regression

Logistic regression (LR) is one of the most widely used models for the relationships between the patient’s characteristics  $X_t$  and the surgical outcome. The probability  $p_t$  of a patient being readmitted ( $y_t = 1$ ) versus being not readmitted ( $y_t = 0$ ) due to his/her own risk is predicted as a linear combination of elements in

$X_t$  after the logit transformation [34]:

$$\text{logit}(p_t) = \log\left(\frac{p_t}{1-p_t}\right) = \beta_0 + \sum_{j=1} \beta_j X_{tj}, \quad (2.4)$$

where  $X_{tj}$  is the  $j$ th element in  $X_t$ . Then  $p_t$  is calculated using the following equation:

$$p_t = \frac{e^{\beta_0 + \sum_{j=1} \beta_j X_{tj}}}{1 + e^{\beta_0 + \sum_{j=1} \beta_j X_{tj}}}. \quad (2.5)$$

The parameters  $\beta_0$  and  $\beta_j$  can be estimated using the maximum likelihood estimation (MLE) from the training data.

### 2.2.3.2 Tree-based Models

Despite the popularity of logistic regression in healthcare applications, the assumption of the linear combination of characteristics in Equation (2.4) limits its prediction performance for readmission rate. Tree-based models have been adopted for this application to overcome the limitation. A decision tree has a hierarchical structure that starts from the top to the bottom, consisting of nodes  $R_m$ ,  $m = 1, 2, \dots, J$  connected by branches. The classification task in a decision tree uses a sequence of decision rules based on the value of  $X_{tj}$ , splitting the training data into terminal nodes  $R_m$ . The proportion of the readmission for a single node  $R_m$  can be calculated using the following equation [37]:

$$\hat{p}_m = \frac{1}{N_m} \sum_{X_t \in R_m} I(y_t = 1) \quad (2.6)$$

where  $N_m$  is the number of training observations in the terminal node  $R_m$ . The proportion  $\hat{p}_m$  can be used to approximate the readmission probability of a new observation falling in the node  $R_m$  [30].

The decision tree method suffers from its sensitivity to the training data [37]. Random forests (RF), introduced by Breiman [35], use the bagging to overcome this

drawback. RF consists of  $B$  uncorrelated decision trees, where each tree carries information from a bootstrap sample drawn from the training data. By averaging those shallow decision trees, RF can alleviate the overfitting problem. Hence, it can achieve accurate predictions from the averaging, even if some shallow trees are weak estimators. The variance of the RF model will be reduced by  $B$  compared to a single decision tree.

Gradient boosting machine (GBM) is another popular tree-based model for regression and classification problems by adopting the boosting and gradient descent methods [36]. GBM works by sequentially adding  $B$  decision trees to an ensemble, and the final model is the sum of all the individual trees. GBM achieves state-of-the-art performance for readmission prediction [33]. However, as the algorithm ensembles decision trees sequentially, GBM is sensitive to overfitting and requires careful tuning of hyperparameters, such as the number of trees, the shrinkage rate, and the interaction depth of the input variables.

In the work, we skip the single decision tree due to its poorer performance compared to RF and GBM [33]. We fix the number of trees  $B = 1000$  for RF as its prediction performance is robust to a large enough  $B$ . For GBM, we choose the shrinkage rate as 0.01, the interaction depth as 4, and the number of trees  $B$  from 100 to 10,000 to evaluate the effect of model complexity on the risk prediction. At last, we also consider incorporating the variable selection [38] by choosing a subset of the characteristics for model training.

#### **2.2.4 Model Comparison for Risk Prediction**

To ensure the performance of RA CUSUM charts, we need to find a suitable model to predict the patient's risk. To examine the prediction performance of different models, we need to address challenges in model fitting using the retrospective dataset,

such as small sample sizes and unbalanced response variables. We develop a pipeline incorporating cross-validation and the AUROC for model comparison and selection for risk prediction. The proposed pipeline can balance the prediction variance and bias of the machine learning models with various complexities, which will be shown in Section 2.3.2.

As there are only 243 patients in the retrospective dataset, models with high complexity, such as GBMs with a large  $B$ , can overfit the training data. It will provide the perfect predictions for observations used in training but poor results for unseen ones. Thus, we need the cross-validation, first proposed by Larson et al. [39], to evaluate the true prediction performance. The cross-validation will train and evaluate the model by separating the data into two parts: one for training the model and the second for validation. In the  $K$ -fold CV, the cross-validation procedure will be iterated  $K$  times. For each iteration,  $K - 1$  folds are employed for training the model, while the remaining fold is for evaluating the model. As cross-validation is able to evaluate the model performance for future data, we adopt the 10-fold cross-validation to compare different models' risk prediction abilities.

Due to the small sample size of the retrospective dataset, a model with high complexity may show an excessive prediction variance, even with a good training fit. Such a model will provide a perfect prediction in the retrospective dataset and show a CUSUM chart close to 0. If we apply such a chart to a new dataset, the false alarm rate will be high. Thus, in the model selection, we must consider both the bias and variance of the prediction, avoiding an overfitting model for the small training dataset. In Section 2.3.2, we will demonstrate the prediction bias and variance for each machine learning model, showing that the proposed selection pipeline can achieve a trade-off for risk prediction.

### 2.2.5 RA CUSUM Chart Evaluation

As explained in Section 2.2.2, an RA CUSUM chart signals when  $Z_t$  exceeds a limit value  $h$ . An optimal value of  $h$  is determined by the targeted performance of the RA CUSUM procedure. We can measure the performance of the RA CUSUM chart using the Average Run Length (ARL), which is the expected observation number where a control chart from the start of the process triggers the first signal. The in-control ARL is denoted as  $ARL_0$  while the out-of-control ARL is denoted as  $ARL_1$ . Ideally, as the process is in control,  $ARL_0$  should be large since the signals here represent false alarms. On the other hand,  $ARL_1$  should be short while the process is out of control to ensure an early detection [28]. We can also consider the role of  $ARL_0$  as the Type I error and  $ARL_1$  as the power in the traditional hypothesis test [3]. So,  $h$  should be selected to ensure a large  $ARL_0$ , such as 500, using in-control data, and  $ARL_1$  will be evaluated using out-of-control data with the selected  $h$ .

As there is no explicit formula of ARL for CUSUM charts, several approximation methods are recommended to determine the control limit  $h$ . Reynolds et al. [40] suggested using Markov chain simulation, and Jones et al. [41] proposed using a bootstrap resampling method to adjust the control limits. As there is no parametric model for the readmission risk of the PTBD procedure, we adopt the bootstrap resampling method to compute the ARL. Given a control limit value  $h$ , we generate  $M_B$  bootstrap samples with length  $N_B$  from the retrospective dataset. For the  $j$ th sample, we predict each patient's risk  $p_{jt}$  and plot  $Z_{jt}$  according to Equations (2.2) and (2.3). The run length  $RL_j$  is using the following equation:

$$RL_j = \inf(t : Z_{jt} \geq h, t = 1, 2, \dots, N_B). \quad (2.7)$$

Finally, the in-control  $ARL_0$  is estimated by  $\sum_{j=1}^{M_B} RL_j / M_B$ . The optimal  $h^*$  will

be found according to a targeted  $ARL_0$ . An  $ARL$  can be evaluated using the same procedure, resampling from the prospective data. To ensure the performance of the  $ARL$  estimation, we choose  $M_B$  as 1000 and  $N_B$  as 5000.

We summarize all the steps in the proposed RA CUSUM charts in Algorithm 2.2.5.

1. Calculate the means and standard deviations of cross-validation AUROC using the retrospective dataset for 8 machine learning models, including the logistic regression, random forests ( $B = 1000$ ), and GBMs with  $B \in \{100, 200, 500, 1000, 2000, 5000\}$ .
2. Select an appropriate machine learning model based on the cross-validation AUROC and retrain the model using all retrospective data.
3. Estimate the patient's risk  $p_t$  for the retrospective and prospective datasets.
4. Find the optimal  $h^*$  to let the retrospective  $ARL_0 = 500$  by bootstrap resampling from the retrospective data.
5. Evaluate the prospective  $ARL$  with  $h^*$  by bootstrap resampling from the prospective data.
6. Plot the RA CUSUM charts for the retrospective and prospective datasets.

## 2.3 Results and Discussions

### 2.3.1 Data Summary and Preliminary Analysis

In the retrospective data, among 243 patients who underwent PTBD procedure, there were 56.0% males. The procedure was performed due to benign obstructions in 47.7% of patients. Most patients were insured (89.3%), and the average age was 59.6



years. Most catheters were positioned in the right biliary system (62.6%), and the remaining were positioned in the left (33.7%) and bilateral (3.7%). 33.7% of patients were readmitted within 30 days. On the other hand, the prospective data includes 67 patients who underwent PTBD with the new procedural paradigm. The procedure was performed due to benign obstructions in 37.3% of patients. Most patients were insured (95.5%), and the average age was 61.3 years old. The catheters were positioned in the right biliary system (47.8%), left (44.8%), and bilateral (7.4%). The prospective readmission rate was slightly lower than the retrospective data (29.9%). For more details, Tables 1 & 2 show the summary of all the categorical and numerical variables, where the meaning of each characteristic can be found in Section 2.2.1. We can identify different patterns in some characteristics between the two datasets, justifying using an RA CUSUM chart.

Categorical characteristics		% in retrospective	% in prospective
Readmitted (vs. Not Readmitted)		33.7%	29.9%
Benign (vs. Malignant)		47.7%	37.3%
Male (vs. Female)		56.0%	58.2%
Insured (vs. Uninsured)		89.3%	95.5%
Type I/E (vs. Type E)		86.8%	86.6%
Race	Caucasian	63.8%	49.3%
	African American	30.0%	37.3%
	Asian	0.8%	3.0%
	American Indian	5.4%	10.4%
Access	Left	33.7%	44.8%
	Right	62.6%	47.8%
	Bilateral	3.7%	7.4%

Table 1.: Summary of the categorical characteristics in retrospective and prospective datasets.

Then, we apply a preliminary analysis to evaluate the relationship between those characteristics and the readmission. Table 3 shows the estimated coefficients with their confidence intervals (C.I.) and p-values of patients being readmitted vs. not

Numerical characteristic	Mean in retrospective	Mean in prospective
Age	59.6	61.3
ASA	3.06	3.03
Cr	1.15	1.01
AST	141.6	254.3
ALT	137.9	183.7
Alk.phos	557.3	803.6
T.bili	8.04	9.98
D.bili	6.34	7.17
WBC	10.0	12.5
Hgb	10.6	9.78
PLT	260.1	260.5
INR	1.31	1.27
Size in Fr	8.87	9.24

Table 2.: Means of the numerical characteristics in retrospective and prospective datasets.

readmitted for each characteristic in the retrospective and the prospective datasets using a simple logistic regression model. We found that none of the listed related characteristics has a significant linear relationship ( $p$ -value  $< 0.10$ ) with the readmission variable in the retrospective dataset. Several significant characteristics are identified from the prospective datasets, such as gender, WBC, and Plt. However, considering the small sample size (67 patients), those results may not be reliable. Thus, using those characteristics, a nonlinear model is preferred for the readmission risk prediction.

As the simple logistic regression fails to find significant predictors for the readmission risk, we follow [38] to find important characteristics using the RF method. The variable importance plots are shown in Figure 1, where the left and right figures illustrate the average decrease in prediction accuracy and Gini index, respectively, if the corresponding characteristic is removed. Here, we find a subset of the characteristics that give us a positive decrease in accuracy and a relatively large decrease ( $> 6.0$ )

Characteristics	Retrospective		Prospective	
	Coefficients (95% C.I.)	P-value	Coefficients (95% C.I.)	P-value
Benign (0) vs. Malignant (1)	-0.14 (-0.67, 0.40)	0.61	-0.46 (-1.53, 0.62)	0.40
Male (0) vs. Female (1)	0.07 (-0.47, 0.60)	0.81	1.07 (0.00, 2.18)	0.05
Type I/E (0) vs. Type E (1)	0.03 (-0.78, 0.80)	0.94	-0.45 (-2.42, 1.08)	0.59
Insured (0) vs. Uninsured (1)	-0.59 (-1.63, 1.31)	0.23	1.63 (-0.77, 4.74)	0.19
Caucasian (0) vs. African American (1)	-0.03 (-0.63, 0.56)	0.92	0.73 (-0.39, 1.89)	0.20
Caucasian (0) vs. Asian (1)	0.68 (-2.56, 3.93)	0.63	-15.4 (-∞, 250)	0.99
Caucasian (0) vs. American Indian (1)	0.21 (-1.02, 1.36)	0.72	0.22 (-1.85, 1.97)	0.81
Left (0) vs. Right (1)	0.11 (-0.45, 0.69)	0.69	1.01 (-0.10, 2.20)	0.08
Left (0) vs. Bilateral (1)	-1.37 (-4.31, 0.40)	0.21	0.00 (-3.07, 2.14)	1.00
Age	-0.009 (-0.028, 0.010)	0.35	-0.023 (-0.063, 0.016)	0.26
ASA	-0.173 (-0.677, 0.320)	0.50	-0.661 (-2.070, 0.598)	0.32
Cr	0.080 (-0.161, 0.316)	0.50	-0.912 (-2.359, 0.061)	0.14
AST	-0.001 (-0.003, 0.001)	0.43	-0.001 (-0.005, 0.001)	0.28
ALT	0.000 (-0.003, 0.002)	0.73	0.000 (-0.003, 0.002)	0.95
Alk.phos	0.000 (-0.001, 0.001)	0.99	0.000 (-0.001, 0.000)	0.41
T.bili	-0.034 (-0.075, 0.005)	0.10	0.008 (-0.051, 0.063)	0.76
D.bili	-0.037 (-0.090, 0.012)	0.15	0.014 (-0.068, 0.091)	0.72
WBC	0.016 (-0.031, 0.063)	0.49	-0.127 (-0.279, -0.020)	0.06
Hgb	0.023 (-0.104, 0.151)	0.72	0.139 (-0.144, 0.429)	0.34
Plt	0.001 (-0.001, 0.003)	0.24	0.003 (-0.000, 0.007)	0.08
INR	0.120 (-0.596, 0.802)	0.73	-0.760 (-2.821, 0.832)	0.40
Size in Fr	0.097 (-0.137, 0.333)	0.42	0.211 (-0.283, 0.775)	0.43

Table 3.: Preliminary analysis for the relationship between each characteristic and the readmission by the simple logistic regression.

in the Gini index. The subset includes 8 variables: Age, AST, ALT, T.bili, D.bili, Hgb, Plt, and INR. With a small sample size, using a subset can reduce overfitting and improve risk prediction [37]. In Section 2.3.2, we will compare models with this subset and with all available characteristics.

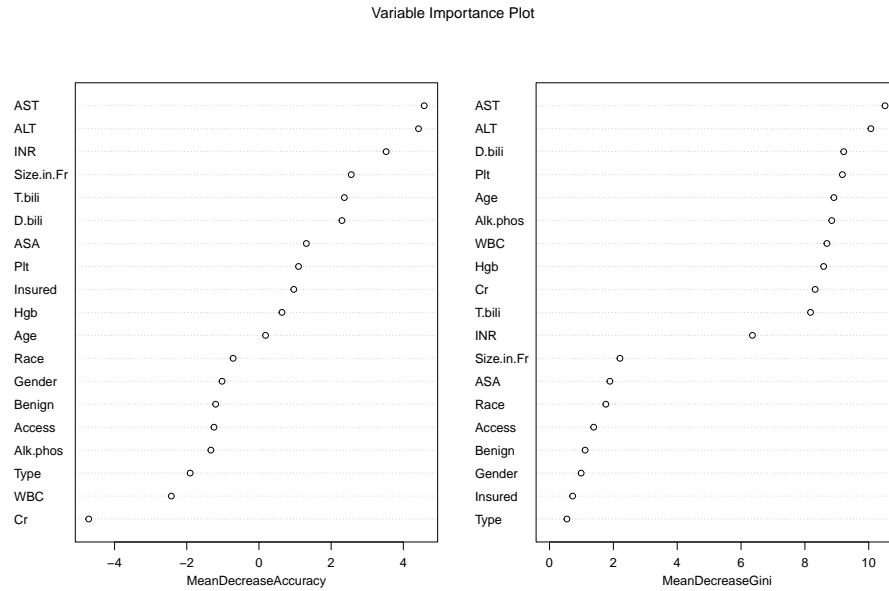


Fig. 1.: The importance of all characteristics from the random forests model on the retrospective dataset.

### 2.3.2 Predictive Model Comparison

Our preliminary analysis shows that a logistic regression model fails to identify significant readmission prediction characteristics, indicating that tree-based models could be more appropriate. We evaluate machine learning models with different complexity using 10-fold cross-validation on the retrospective dataset. For each model, we compare the performance between models with the subset identified in Section 2.3.1 and with all available patients' characteristics. As the RF and GBM will gen-

erate different models from the same dataset due to the randomness in their training process, we keep the fold separation and repeat the cross-validation by 100 times, obtaining the mean and standard deviation (s.d.) of the AUROC. The results of the 26 models are shown in Table 4, and the highest AUROC is marked in red.

From Table 4, we have the following discoveries for the risk prediction for the PTBD procedure. First, the tree-based models, i.e., RF and GBM, work better than the logistic regression, whose average AUROCs are less than 0.5 due to the weak linear relationship and imbalanced classes. Second, the GBMs with relatively large complexity achieve the highest AUROC, indicating the complicated association between the characteristics and a patient’s readmission risk. However, the performance of GBMs using the subset begins to decrease when the complexity becomes too high. Third, the models using all variables have smaller AUROCs than their counterparts using the subset, which confirms that the important variables found by the RFs can improve all those models. In future sections, we will only consider models trained by the subset. At last, the best AUROC, about 0.57, is still not satisfying. However, readmission risk prediction is a challenging problem. In their survey, Huang et al. [33] achieved a 0.66 AUROC using 372,293 patients’ information for pneumonia readmission. Considering the small training sample size, 243, we believe the current model selection is comprehensive for the PTBD readmission prediction. Based on Table 4, the best available model is the GBM with 3,000 trees. If we use the “one-standard error” rule[37], i.e., finding the most parsimonious model within one standard error of the best one, the GBM with 2,000 trees should also be considered. We will implement the RA CUSUM charts with both models in the next sections to verify the robustness of the detection.

At last, we take a close look at the cross-validation prediction using different models with subset selection, shown in Figure 2. In each subfigure, we show the

Model	Mean $\pm$ s.d. (Subset)	Mean $\pm$ s.d. (All)
Logistic Regression	0.4874 $\pm$ 0.0000	0.3936 $\pm$ 0.0000
Random Forests	0.5112 $\pm$ 0.0044	0.4992 $\pm$ 0.0052
GBM100	0.5204 $\pm$ 0.0095	0.4857 $\pm$ 0.0103
GBM200	0.5317 $\pm$ 0.0085	0.4922 $\pm$ 0.0095
GBM300	0.5403 $\pm$ 0.0075	0.4945 $\pm$ 0.0083
GBM500	0.5476 $\pm$ 0.0057	0.4952 $\pm$ 0.0075
GBM800	0.5578 $\pm$ 0.0061	0.4991 $\pm$ 0.0067
GBM1,000	0.5603 $\pm$ 0.0052	0.5014 $\pm$ 0.0064
GBM2,000	0.5664 $\pm$ 0.0051	0.5116 $\pm$ 0.0054
GBM3,000	0.5668 $\pm$ 0.0043	0.5165 $\pm$ 0.0045
GBM5,000	0.5665 $\pm$ 0.0033	0.5207 $\pm$ 0.0040
GBM8,000	0.5651 $\pm$ 0.0030	0.5234 $\pm$ 0.0037
GBM10,000	0.5640 $\pm$ 0.0031	0.5240 $\pm$ 0.0032

Table 4.: The average and standard deviation (s.d.) of AUROC using 10-fold cross-validation for machine learning models with the subset in Section 2.3.2 and all variables.

boxplots for the predicted readmission risk for the not-readmitted and readmitted patients. For a simple model, such as the logistic regression (LR) and the GBM with only 100 trees (GBM100), the within-group variations are small, but the two groups are almost not differentiable. That explains why such models only achieve AUROCs close to 0.5. When we increase the complexity of the machine learning model, the predicted risks of the two groups begin to show differences at the expense of a larger within-group variation. However, if the model complexity is too high such as the GBM with 5,000 trees (GBM5000), the within-group variation decreases the model performance. Thus, we must carefully choose the model by balancing its bias and variance for the risk prediction of the RA CUSUM charts.

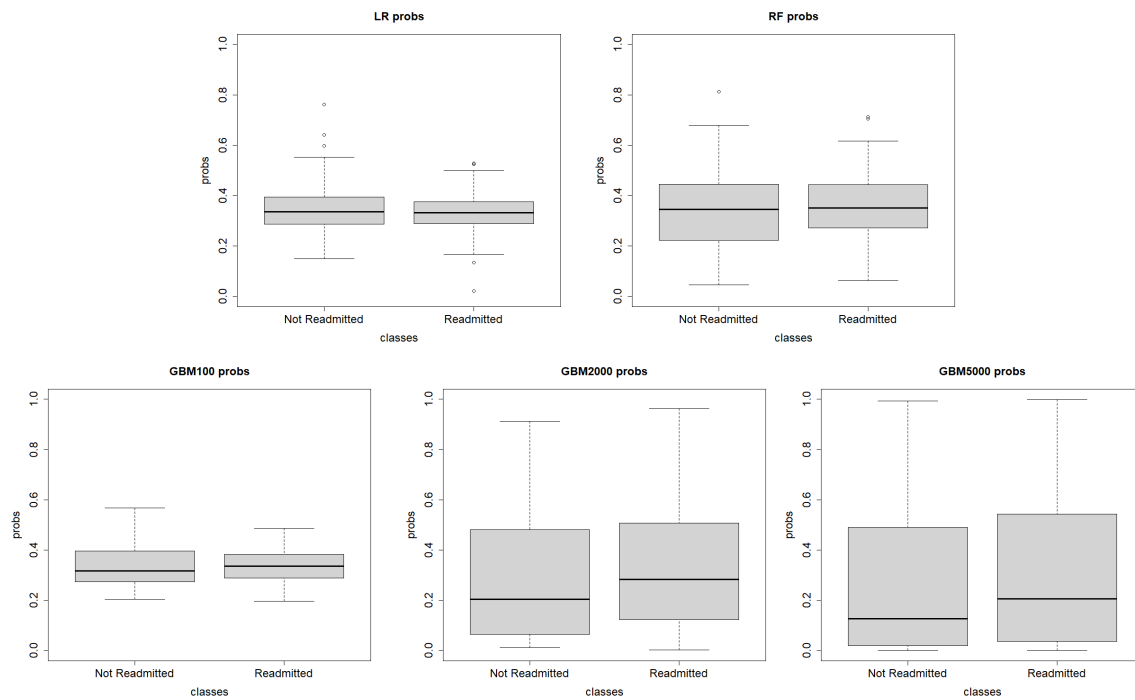


Fig. 2.: The cross-validation prediction for readmitted vs. not readmitted groups for the retrospective model using different machine learning models.

### 2.3.3 Control Charts Construction and Evaluation

In this section, we construct the RA CUSUM charts using the PTBD retrospective (sample size 243) and prospective datasets (sample size 67). The RA CUSUM procedure is divided into two phases: Phase 1 trains the model and estimates the parameters using existing observations, and Phase 2 employs the obtained model to monitor an ongoing process. The retrospective data is used in Phase 1, and the prospective data is used in Phase 2 analysis. Based on the results in Section 2.3.2, we use the GBM model with 3000 trees (GBM3000) trained from the retrospective dataset to estimate the probability  $p_t$  of the readmission rate due to the patient’s own risk. We also consider the model selected by the “one-standard-error” rule, GBM2000. The random forests and logistic regression are compared to the two GBM models. The RA CUSUM procedure is constructed to detect a halving of the odds of the readmission  $R_0 = 1$  vs.  $R_1 = 0.5$ , and the performance for  $R_1 = 0.2$  and  $0.8$  are also evaluated. Since we have  $R_1 < R_0$ , an increase in the chart means an increase in the likelihood of the alternative  $R_1$ , which indicates a decrease in the readmission rate. To highlight the effects of the risk adjustment, we implement a standard CUSUM chart where the patient’s risk is set as a constant 0.337 estimated from the average readmission rate of the retrospective dataset.

As it was explained in Section 2.2.5, we use  $ARL$  to measure the RA CUSUM chart performance, where a good RA CUSUM chart will have a large  $ARL_0$  indicating a low false alarm rate, and a short  $ARL_1$  indicating a fast detection. Using the bootstrap method, we evaluate  $ARL$  using different control limits  $h$  and find the optimal  $h^*$  targeting the  $ARL \approx 500$  for the retrospective dataset. In our problem, the retrospective dataset sets the baseline for the readmission rate. Thus, this  $ARL$  can be considered as an in-control one. Then, we use the same procedure to evaluate



the  $ARL$ s for the prospective dataset using  $h^*$ . Whether a change to the readmission rate in the prospective dataset is unknown before the analysis, so we can not say that its  $ARL$  is in control or out of control.

Table 5 displays the optimal  $h^*$ ,  $ARL_0$  for the retrospective dataset, and  $ARL$  for the prospective dataset for different alternative  $R_1$ 's for RA CUSUM charts using logistic regression, random forests, GBM2000 and GBM3000. We also show the results using the CUSUM chart without risk adjustment as a baseline. The logistic regression, which is the most popular in existing RA CUSUMs, gives us a longer  $ARL$  for the prospective dataset than the  $ARL_0$  for the retrospective one. However, we do not think it indicates no readmission rate change in the prospective dataset. In Figure 2, the logistic regression failed to differentiate the not-readmitted and readmitted groups in the cross-validation prediction for the retrospective data, and in Table 4, it gave an average AUROC less than 0.5. We also find that the logistic regression's  $h^*$  is close to the no-risk-adjusted one, and the prospective  $ARL$  is longer. Considering all those results, we can conclude that the logistic regression risk adjustment is inappropriate for monitoring the PTBD procedure readmission rate.

On the other hand, the three tree-based models, random forests, GBM2000, and GBM3000, show much shorter  $ARL$  for the prospective dataset. For  $R_1 = 0.5$  and  $ARL_0 \approx 500$  for the retrospective dataset, the prospective  $ARL$  is 27.9 for the random forests, 44.3 and 20.7 for the GBMs with 2000 and 3000 trees. A larger  $R_1$  gives a longer  $ARL$  as closer null and alternative hypotheses will decrease the power of control charts and the detection speed. For the GBMs, GBM3000 yields a smaller  $h^*$  for the retrospective  $ARL_0 \approx 500$  and a shorter prospective  $ARL$  compared to GBM2000 as the former's discriminant ability for the readmitted vs. non-readmitted groups is better based on Table 4. Also, the random forests obtain similar prospective  $ARL$  despite their different AUROCs. One possible reason is that the

random forests model has a smaller in-group prediction variance than GBM2000 and GBM3000, as shown in Figure 2, compensating for the former’s weaker discriminant ability. In conclusion, the risk adjustment by the tree-based models, incorporating the subset selection and complexity tuning, is able to identify possible readmission changes between the retrospective and prospective datasets.

Model	$R_1$	$h^*$	$ARL_0$ (Retrospective)	$ARL$ (Prospective)
No Risk-Adjustment	0.2	4.10	512.38	259.03
	0.5	3.05	503.44	212.69
	0.8	1.58	495.55	191.26
Logistic Regression	0.2	4.19	500.82	650.97
	0.5	3.09	500.78	621.63
	0.8	1.58	501.69	605.89
Random Forests	0.2	1.42	497.38	19.29
	0.5	1.07	504.91	27.91
	0.8	0.61	495.42	49.65
GBM2000	0.2	1.35	501.08	26.53
	0.5	0.99	497.82	44.26
	0.8	0.53	510.82	88.59
GBM3000	0.2	0.91	506.19	11.77
	0.5	0.66	491.69	20.69
	0.8	0.35	509.49	38.30

Table 5.: Average run length ( $ARL$ ) for various  $R_1$  and machine learning models estimated by bootstrapping from the retrospective and prospective dataset.  $h^*$  is tuned for  $ARL_0 \approx 500$  using the retrospective dataset.

Figures 3 -7 show the no risk-adjusted CUSUM chart and the RA ones using the four models in Table 5 for the combined retrospective and prospective datasets when  $R_1 = 0.5$ . The patient indices of the retrospective one range from 1 to 243, and the prospective ones range from 244 to 310. The blue vertical lines separate the two datasets, and we restart the CUSUM charts by setting  $Z_t = 0$  in Equation (2.3) when the prospective dataset begins. The red dashed lines show the optimal  $h^*$  found in Table 5 for the retrospective  $ARL_0 \approx 500$ . If detected, the first out-of-control points

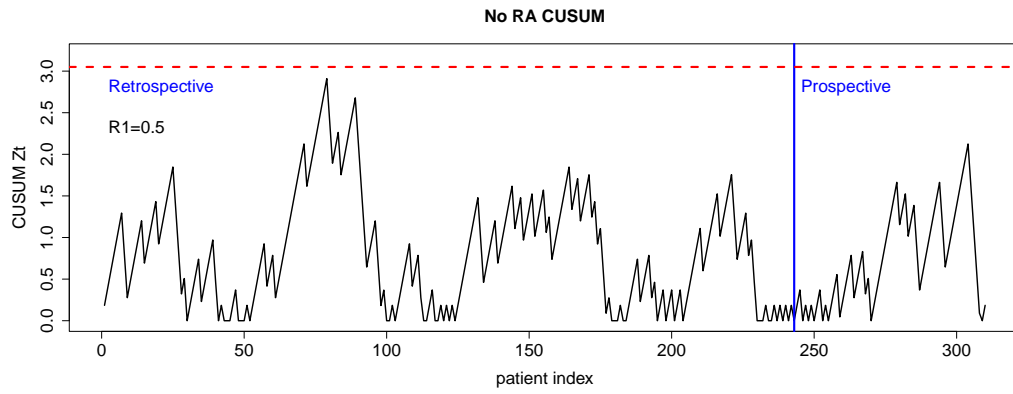


Fig. 3.: The CUSUM chart without risk-adjustment for the combined dataset for  $R_1 = 0.5$ .

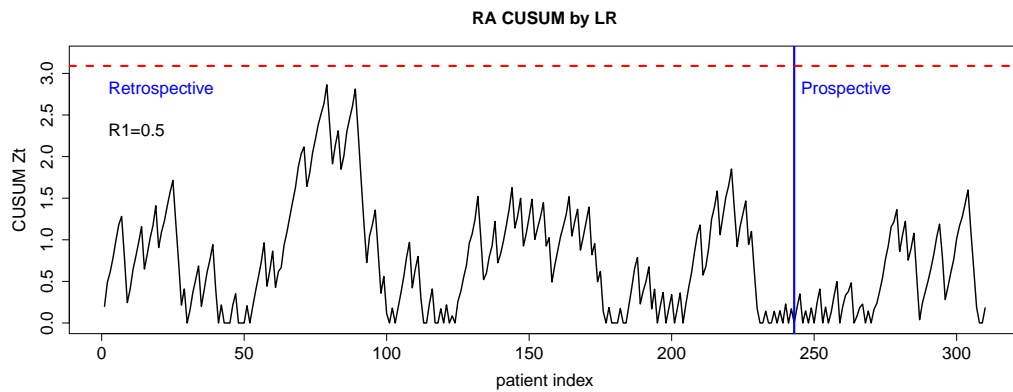


Fig. 4.: The RA CUSUM chart by the logistical regression for the combined dataset for  $R_1 = 0.5$ .

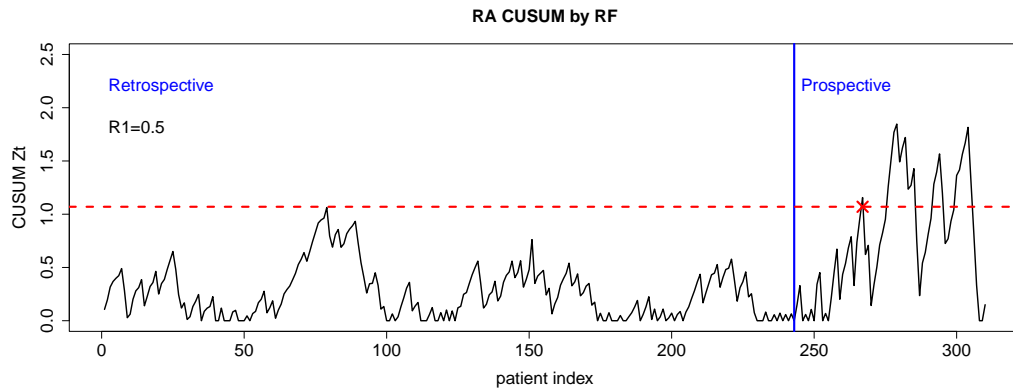


Fig. 5.: The RA CUSUM chart by the random forests for the combined dataset for  $R_1 = 0.5$ .

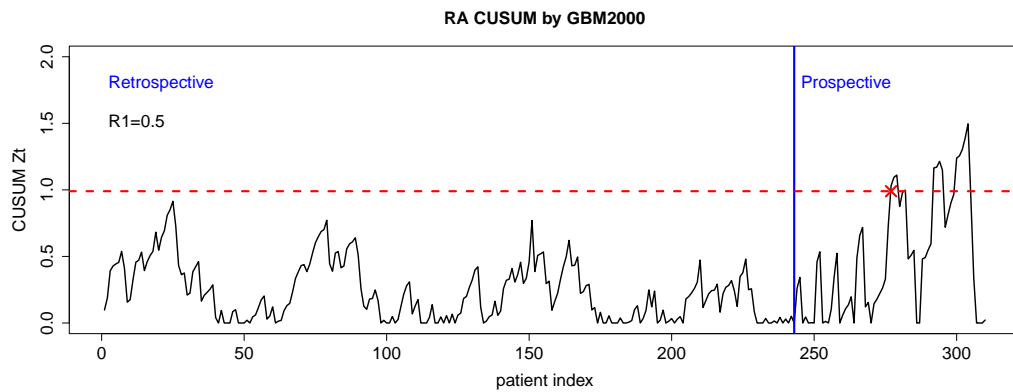


Fig. 6.: The RA CUSUM chart by the GBM with 2000 trees for the combined dataset for  $R_1 = 0.5$ .

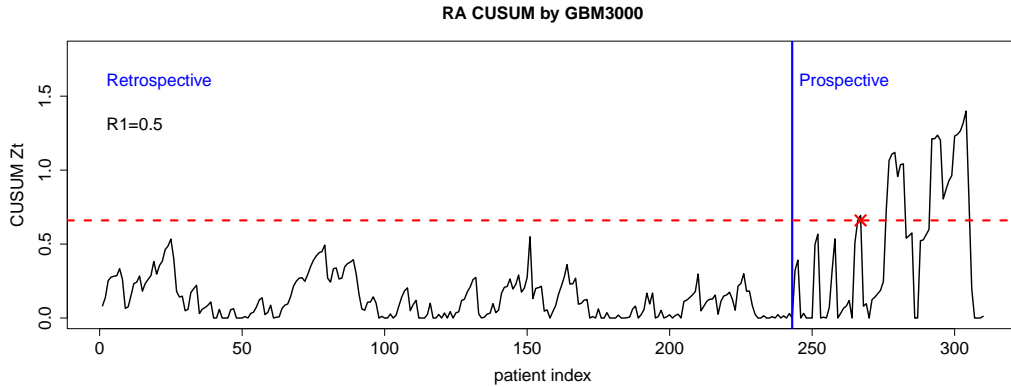


Fig. 7.: The RA CUSUM chart by the GBM with 3000 trees for the combined dataset for  $R_1 = 0.5$ .

in the prospective dataset are marked as red crosses in the CUSUM charts.

The CUSUM chart using the logistic regression (Figure 4) gives similar plots for the retrospective and prospective parts and provides no out-of-control point. The chart is similar to the no-risk-adjusted one (Figure 3) in the retrospective part, and its values in the prospective part are even lower. The linear regression provided an AUROC of less than 0.5 in the cross-validation using the retrospective data, indicating that the model has no or even negative effect in evaluating the risks for the readmitted and not readmitted groups. Those two plots highlight the importance of the model selection for the RA CUSUMs. Without a model with good discriminant power, the risk adjustment cannot utilize the patient’s characteristics to improve the CUSUM chart performance.

The remaining three charts show the results of RA CUSUM charts with tree-based models. The random forests chart (Figure 5) finds the out-of-control signal at the 24th patient in the prospective dataset, and the GBM2000 (Figure 6) and GBM3000 (Figure 7) charts find the signals at the 34th and 24th patients, respec-

tively. The retrospective curves no longer have oblivious peaks and fluctuate below the control limits, thanks to the tree-based models' flexibility for risk prediction. At the same time, an increasing trend appears in the second half of the prospective dataset. If we increase  $h^*$  by a small margin in Figures 5 and 7, the three charts will give similar locations for the out-of-control signals, around 30th patients in the prospective procedures. Considering the discriminant ability of the random forests and GBMs shown in Figure 2 and Table 4, it is reasonable to believe that those detection results show real signal changes, indicating that the readmission rate of the PTBD procedure begins to decrease after the new post-procedural care paradigm was implemented by a certain period.

## 2.4 Conclusions

In this chapter, we propose RA CUSUM charts combined with machine learning models for estimating the patient's risk from their preoperative characteristics and monitoring the change of the readmission rates in healthcare applications. The tree-based models have demonstrated better discriminant ability and faster detection of changes compared to traditional charts based on logistic regression. By evaluating those models with cross-validation, AUROC, and bias-variance analysis, we show that checking the models' performance is critical before applying them to control charts. Furthermore, the findings of the proposed charts show that implementing a new post-procedural care paradigm has reduced readmission rates for the PTBD procedure. Overall, this study provides an effective method in the field of healthcare monitoring to improve patient outcomes while reducing healthcare costs. Also, we demonstrate the use of advanced machine learning models in control chart design and implementation.

## CHAPTER 3

### MULTILAYER MODELING FOR WIDE-RANGE CHEMICAL CONCENTRATION PREDICTION IN SPECTROSCOPIC

#### 3.1 Introduction

In chemical experiments, various types of spectroscopic techniques are widely used to analyze the characterization of the electronic and structural properties of a wide range of chemical compounds. They are beneficial in analyzing solid and liquid mixtures due to their simplicity, low cost, and fast analysis time [42]. However, the complexity of spectra datasets and the presence of overlapping peaks make it challenging to accurately quantify the composition of the mixture. In recent years, there has been a growing interest in using statistical methods to analyze the spectra data, as this can provide a more rapid and cost-effective alternative to traditional wet chemistry methods, which typically require a significant amount of time, expertise, and resources, and may be limited by the sensitivity and selectivity of the analytical techniques used. However, analyzing spectra data presents challenges rooted in high dimensionality, multicollinearity, and the absence of global linearity. The large number of variables complicates analysis and interpretation, while multicollinearity can destabilize models. Moreover, the non-linear nature of spectral data poses difficulties in capturing accurate relationships. These challenges underscore the need for tailored statistical techniques to ensure robust and more efficient results.

In the next section, we will examine some of the most commonly used statistical methods in different types of spectra datasets and their advantages and applications. Each statistical method will be introduced to address a common challenge associated

with spectra data.

### 3.1.1 High Dimensionality and Multicollinearity in Spectra Data

Analyzing spectra data presents a dual challenge involving both high dimensionality and multicollinearity. The abundance of variables in spectra data introduces the curse of dimensionality, complicating the development of accurate analyses. Simultaneously, the presence of multiple highly correlated absorbance values gives rise to multicollinearity, further complicating the identification of relevant variables for predictive tasks. This interplay of challenges underscores the need for sophisticated statistical techniques, emphasizing both dimensionality reduction strategies and methods to address multicollinearity, extract meaningful insights, and enhance the robustness of predictive models in spectroscopic analyses.

Principal component regression and partial least squares are valuable methods in addressing the complexities of high dimensionality and multicollinearity in spectra data. Principal component regression tackles dimensionality by transforming the original variables into a reduced set of principal components, capturing the essential variability in the data. This not only mitigates the curse of dimensionality but also enhances computational efficiency. On the other hand, partial least squares combines the strengths of dimensionality reduction and regression, allowing for the extraction of latent variables that capture both the spectral information and the response variable. partial least squares effectively mitigate multicollinearity challenges by emphasizing the relationships between absorbance values and prediction targets.

Principal component regression and partial least squares thus serve as powerful tools to address these challenges and build robust predictive models. In a study of ultraviolet-visible spectra data of fingerprinting for quality control analysis of food and functional food, Farag et al. [5] used principal component regression to reduce the



dimensionality of large ultraviolet-visible datasets in the food sample [5]. The study showed that after reducing the number of components, they were able to identify any patterns or trends in the spectral data that are relevant to quality control and visualize the spectra data. Also, Shi et al. [43] used principal component regression and other methods to analyze spectral data from an online ultraviolet-visible spectrophotometer to detect changes in the concentration of dissolved organic carbon in drinking water. They stated that principal component regression helped improve the efficiency of the data analysis process and make it easier to identify any significant changes in water quality parameters [43].

For other spectra data such as infrared spectra data, Togkalidou et al. [44] compared different forms of principal component regression models to predict the solute concentrations in aqueous solutions obtained from Infrared spectroscopic data. The study used the mean width of the prediction interval as a criterion in the model selection [44]. Suhandy et al. [7] used Principal component regression to develop a calibration model in the spectral data in the ultraviolet-visible region for the quantification of adulteration in Indonesian Palm Civet coffee. The study found that the principal component regression models effectively predicted luwak content in luwak-arabica coffee blends with a high level of accuracy. Jiao et al. [45] extended the principal component regression model and used interval partial least squares and Moving Window Partial Least Squares (MWPLS) models to predict the enantiomeric composition of tryptophan using spectral data. The study results showed that MWPLS was the best model to build a calibration model for the spectral region using leave-one-out cross-validation. Moreover, In a study that obtained the spectroscopic data from micro-Raman spectra, Chawla et al. [46] used Partial Least Squares Discriminant Analysis (PLS-DA) with other statistical models to classify bacterial pathogens from mixed data. The study showed that the accuracy of the prediction reached 80%.

### 3.1.2 Non-Linearity of Spectra Data

Another critical issue with the spectroscopic data is that they do not always exhibit a linear relationship. This non-linearity can occur for various reasons, such as overlapping spectral features, the non-linear form of the response variable, and the non-linear interaction between molecules [6] [47] [48]. However, when the relationship between the variables is nonlinear, using a linear regression model can lead to biased and inaccurate predictions of the target variable. Some researchers handle this problem by simply discretizing the numeric variable into discrete values (classes). For example, Diaz et al. [49] defined five classes of organic acid concentrations based on their clinical significance: low, low-normal, normal, normal-high, and high. They used clinical reference values to determine the threshold concentrations for each class. While other researchers keep the target variable in its numerical form by applying techniques like logarithmic and exponential transformations. Kvalheim [50] used logarithmic transformation to study dissolved organic matter in natural waters. The study showed that using the logarithm of the concentration or absorbance values, the data can be compressed to a more manageable range, which makes it easier to visualize and analyze. Also, the author used techniques such as Near-Infrared Spectroscopy (NIRS) to predict the concentration of multiple compounds in raw coffee beans. NIRS is a non-destructive analytical technique used to measure the chemical composition of a sample. It is based on the absorption of near-infrared light by molecules in the sample, which produces a unique spectrum that can be used to identify and quantify the different chemical components. The authors stated that using NIRS spectra data helped reduce the range of the samples' composition when performing PLS models.

Moreover, Hossain et al. [6] used a non-linear kernel function for a machine learning model, support vector machines, to detect the presence of disinfectants in

drinking water samples obtained from spectroscopic data. On the other hand, other studies utilize the capability of Artificial Neural Network (ANN) to address the non-linear challenge of the spectra data. For instance, Takahashi et al. [47] compared linear models, PLS-DA, and non-linear models, ANN, to identify the presence of vinegar in blends using Ultraviolet-visible spectra spectroscopic data. The study's findings indicate that non-linear models using ANNs are better suited to identifying the components of binary mixtures of vinegar compared to PLS-DA models, as they demonstrate higher accuracy.

While the mentioned methods have demonstrated efficacy in handling non-linearity in spectroscopic data, it is essential to note that they often assume global linearity in the relationship between variables. Despite their successful applications, these techniques encounter challenges when the concentration range is exceptionally wide, and there is no clear global linearity present. The inherent assumption of a uniform relationship across the entire concentration spectrum may limit the accuracy of predictions, especially in cases where the data exhibits significant variability or non-linear patterns within different concentration ranges. Recognizing this limitation, novel approaches are presented to address the unique challenges posed by spectroscopic data that encompasses a wide concentration range. The proposed methods seek to overcome these limitations by introducing innovative multi-layer models specifically designed to adapt to the non-linearity inherent in wide-ranging concentration profiles.

This chapter proposes two novel multi-layer models to address the non-linearity of chemical concentrations in spectral data. These proposed methods can be integrated with commonly used regression models such as partial least squares and principal component regression to address the dimensional and multicollinearity challenges associated with spectra data. By integrating these regression models with the proposed methods, the chapter suggests a comprehensive solution for handling the challenges

associated with chemical concentrations in spectral data.

## 3.2 Materials and Methods

### 3.2.1 Materials and Instruments

Ultraviolet-Visible (UV-Vis) spectra employing wavelengths from 200-950 nm were obtained using a model 440 UV-Vis spectrophotometer manufactured by SI Photonics, which uses tungsten and deuterium light sources for the visible light spectrum and ultraviolet spectrums, respectively. The light source switches from deuterium to tungsten at 460 nm. The cuvette was a semi-micro quartz cuvette with a path length of 10 mm produced by Fisherbrand. To develop a baseline, a blank of deionized water was collected before measuring the UV-Vis spectra for the sample, which was inserted into the cell holder. The SI 400 program was utilized to collect each spectrum, and the raw data was stored in CSV files.

To create the reflectance element, nickel sulfate hexahydrate ( $\text{NiSO}_4 \cdot 6\text{H}_2\text{O}$ , > 98%) obtained from Sigma-Aldrich was dissolved in deionized water at molar concentrations ranging from  $1 \times 10^{-6}M$  to  $0.9M$ , where  $M$  notates mol per liter. The samples were prepared by taking  $1M$   $\text{NiSO}_4$  stock solution, diluting small samples of stock to  $0.1 - 0.9M$  samples, and then serial diluting by a factor of 10 down to the order of magnitude of  $10^{-6}M$ . Each sample was vortexed for about 15 seconds before the next step of dilution.

The spectra were collected in triplicate, meaning that each concentration was sampled 3 times in the spectrometer, and each sample had 3 spectra taken each time. This gives 9 total spectra for each concentration coming from the same batch. This was done to maximize data collected from the same batch of solutions.

### 3.2.2 Spectral Data Analysis

In the introduction section, it was highlighted that the analysis of spectral data poses several challenges, including high dimensionality, multicollinearity, and non-linear relationships among the predictor variables. To address these challenges, two novel multi-layer methods will be introduced that aim to effectively reduce the wide range of chemical concentrations, which is the main cause of non-linearity in spectral data. These novel techniques have been designed to work in conjunction with two widely used regression models in chemometrics: principal component regression and partial least squares. By combining these regression models with the new multi-layer methods, a comprehensive solution is proposed for handling the challenges associated with chemical concentration variability in spectral data. The proposed approach can help to improve the accuracy and robustness of chemometric models and facilitate the analysis of complex chemical systems.

#### 3.2.2.1 Principal Component Regression & Partial Least Squares algorithms

In a multiple linear regression model with  $j$  predictor variables, the Ordinary Least Squares (OLS) objective is to minimize the sum of squared differences between the predicted values ( $X\beta$ ) and the actual response ( $Y$ ):

$$\min_{\beta} \|Y - X\beta\|_2^2, \quad (3.1)$$

where  $X$  is a matrix of dimensions  $i \times j$ , and  $i$  denotes the number of samples and  $j$  represents the number of spectral variables. Also, an  $i \times 1$  vector  $Y$  represents the true response variable, and  $\beta$  is the vector of coefficients to be estimated. OLS

solution is given by:

$$\beta_{ols} = (X^T X)^{-1} X^T Y, \quad (3.2)$$

and

$$\hat{Y} = X\beta_{ols}. \quad (3.3)$$

Now, let's consider the case where  $j$  (the number of columns in  $X$ ) exceeds  $i$  (the number of observations). The matrix  $X^T X$  becomes singular (non-invertible) because its rank is at most  $i$ , and it cannot be inverted. Therefore, the solution for  $\beta$  does not exist, and the OLS estimation fails [51].

Principal Component Regression (PCR) and Partial Least Squares (PLS) are powerful statistical techniques employed in regression analysis to address the issue of high-dimensional data. PCR entails projecting predictor variables onto a set of uncorrelated variables called principal components derived from a Principal Component Analysis (PCA). These components, representing linear combinations of the original predictors, are utilized in a linear regression model to predict the response variable. By reducing data dimensionality and selecting a subset of principal components that capture key information, PCR enhances model accuracy and addresses multicollinearity concerns, especially in the presence of numerous correlated predictors [52]. On the other hand, PLS constructs latent variables through linear combinations of original predictors that explain maximum variation in the response variable. PLS proves advantageous when the number of predictors surpasses the number of observations, making it a valuable tool in scenarios characterized by substantial predictor dimensions. Moreover, The PLS algorithm goes beyond merely maximizing variance; it processes information from both predictors and predicted variables, seeking factors that not only explain maximum variance but also provide maximum correlation between them. A notable advantage of PLS is its ability to assign equal importance to

predictors and predicted variables [53].

The PCR approach involves a two-step process. First, it conducts a PCA on the matrix  $X$ . In PCA,  $X$  is decomposed using its singular value decomposition:

$$X = R\Lambda V^T \quad (3.4)$$

with

$$R^T R = V^T V = I, \quad (3.5)$$

where  $R$  and  $V$  are matrices of left and right singular vectors, and  $\Lambda$  is a diagonal matrix with singular values. These singular vectors are ordered based on their corresponding singular values, representing the square root of the variance (or eigenvalue) of  $X$  explained by each vector. The columns of  $V$ , known as loadings, and the columns of  $G = R\Lambda$  are the factor scores or principal components of  $X$ . In the second step,  $V$  is used to predict  $Y$  through a standard linear regression as the following:

$$X_{pc} = XV_k, \quad (3.6)$$

$$\hat{\alpha} = (X_{pc}^T X_{pc})^{-1} X_{pc}^T Y, \quad (3.7)$$

where

$$\hat{\alpha} = V_k \hat{\beta}. \quad (3.8)$$

The prediction will be:

$$\hat{Y} = X_{pc} \hat{\alpha} \quad (3.9)$$

where the original data matrix  $X$  is first transformed into a lower-dimensional space using principal components  $X_{pc}$ , as represented in Equation 3.6, and  $V_k$  contains the first  $k$  principal components from the SVD of  $X$ . The new regression coefficients  $\hat{\alpha}$  are then estimated in this reduced space as shown in Equation 4.6, which leverages

the least squares method. Equation 4.7 shows the final predictions by combining the dimensional-reduced data with the new estimated coefficients  $\hat{\alpha}$ .

This process uses the captured variance in  $X$  to predict the response variable  $Y$ . This method addresses the multicollinearity of the spectra data due to the orthogonal nature of the singular vectors. However, it is crucial to note that these components were initially chosen to explain  $X$  rather than  $Y$ . Therefore, there is no assurance that the principal components, which optimally explain  $X$ , will necessarily be pertinent for predicting  $Y$ . Hence, PLS emerges as a valuable alternative. PLS derives a set of latent variables from the predictor matrix  $X$  with the aim of maximizing the predictive power for the response matrix  $Y$ . Unlike PCR, which focuses on explaining variance in  $X$ , PLS seeks to capture the covariance between  $X$  and  $Y$  in its components. PLS includes a set of orthogonal factors or loadings that decompose the dependent variables as:

$$X = TP^T, \quad (3.10)$$

where  $T$  is the scores matrix from the latent variables, and  $P$  is the loading matrix for  $X$ . The relationship between the latent variables and the response matrix  $Y$  is formulated as:

$$Y = TBQ^T, \quad (3.11)$$

where  $B$  is a diagonal matrix of regression coefficients, and  $Q$  is the loading matrix for  $Y$ . PLS iteratively extracts latent variables by maximizing covariance between the residuals of  $X$  and  $Y$ . Here, the columns of  $T$  matrix are the *latent variables*. Optimal estimation of  $Y$  is achieved with a subset of latent variables; using too many can result in complexity akin to PCR without enhancing predictive performance. The prediction for new data  $X_{new}$  utilizes the latent variable space derived from the original dataset. The scores  $T_{pc}$  are obtained by transforming  $X_{new}$  using the loading



matrix  $P$ :

$$T_{pc} = X_{new}P. \quad (3.12)$$

The vector of fitted values from PLS can be represented by the first PLS linear combinations in  $T_{pc}$ . The predicted response  $\hat{Y}$  is then calculated from these scores, employing the regression coefficients matrix  $B$  and the loading matrix  $Q$  for  $Y$ :

$$\hat{Y} = T_{pc}BQ^T. \quad (3.13)$$

This approach is particularly advantageous in scenarios where predictors exhibit multicollinearity and the relationship between predictors and response is complex. By focusing on the covariance between  $X$  and  $Y$ , PLS ensures that the components selected are both representative of the predictors and relevant for predicting the response [54].

### 3.2.2.2 Dynamical Layered Regression (DLR) Method

To obtain an initial estimate of each concentration from the training set, we first fit a regression model. However, the wide range of concentrations from the spectra data often leads to poor initial estimates. To address this, we dynamically create narrower layers for each rough estimate by identifying neighboring data and truncating any values outside the interval. By creating these narrower layers, which only include data close to each rough estimate, we can improve the accuracy of our predictions. After that, we fit the regression model again using the narrower layers from the training data and obtain final estimates using the testing data. If the interval contains very few data points, making the prediction inefficient, we replace the new estimate with the original rough estimate. The following are the steps of the DLR method:

- Obtain an initial estimate  $y_1$  using the first layer regression model from all the training data  $\{X, y\}$ .
- Define a scaling factor  $s$ , such that  $s > 1$ , to adjust the width of the range.
- Define the lower and upper bounds of the range as  $l = \frac{y_1}{s}$  and  $u = y_1 * s$ .
- Find a subset of the training data  $\{X_R, y_R\}$ , whose concentrations fall within the range  $[l, u]$ :

$$\text{For all } y \in y_R : \frac{y_1}{s} \leq y \leq y_1 * s. \quad (3.14)$$

- If the size of the subset is too small, set  $y_2 = y_1$ .
- Otherwise, train the second layer regression model from the subset  $\{X_R, y_R\}$  and then obtain the final prediction  $y_2$  using the new model.

Equation 3.14 defines the range as the set of values in  $X_R$  that fall within the range of  $y_1/s$  to  $y_1 * s$ . Where  $y_1$  is the initial prediction from the model,  $X_R$  is the set of neighboring data used to fit the regression model, and  $s$  is the scaling factor that adjusts the width of the range.

### 3.2.2.3 Classified Layered Regression (CLR) Method

Initially, the CLR method starts by dividing the training data into distinct classes based on concentration values using a data-driven approach. To ensure objectivity and avoid prior knowledge, we employ appropriate methods, such as the median for two classes, to evenly divide the data into distinct class intervals. The number of classes, denoted by  $C$ , determines the partitioning of the data into  $C$  non-overlapping intervals. Each interval represents a separate class, indexed from 1 to  $C$ . As the concentrations in the testing data are unknown, we utilize a classification model,

PLS-DA, to predict their classes. In the case of two classes ( $C = 2$ ), The results of the PLS-DA model are then used to divide the testing data into high and low layers similar to the training data. After that, the regression model is fitted for each class. By dividing the data into classes, we can effectively reduce the wide range of concentrations without refitting the model for each prediction. The following are the steps of the CLR classes in case of two classes (High - Low):

- Divide the training data into high and low classes based on a predetermined threshold (e.g., median):

$$H_{tr} = \{X_{tr}, y_{tr} | y_{tr} > \text{median}(y_{tr})\}, \quad (3.15)$$

$$L_{tr} = \{X_{tr}, y_{tr} | y_{tr} \leq \text{median}(y_{tr})\}, \quad (3.16)$$

where  $H_{tr}$  is the set of training data in the high class  $L_{tr}$  is the set of training data in the low class,  $y_{tr}$  is the concentration value of a training data, and  $\text{median}(y_{tr})$  is the threshold value.

- Train a classification model (such as PLS-DA) on all the training data to predict the classes for the testing data.
- Divide the testing data into high and low classes based on the predicted class labels (High - Low):

$$H_{ts} = \{X_{ts}, y_{ts} | \hat{y}_{ts} = \text{High}\}, \quad (3.17)$$

$$L_{ts} = \{X_{ts}, y_{ts} | \hat{y}_{ts} = \text{Low}\}, \quad (3.18)$$

where  $\hat{y}_{ts}$  is the predicted class of testing data using the PLS-DA model.  $H_{ts}$  is the set of testing data in the high class, and  $L_{ts}$  is the data set in the low class.

- Finally, obtain the prediction of each class ( $H_{ts}$  and  $L_{ts}$ ) by fitting the regression model from their corresponding training sets  $H_{tr}$  and  $L_{tr}$ .

### 3.2.2.4 Models Evaluation

#### Evaluation Metrics:

In this study, we evaluated the accuracy of the regression models trained to predict the concentration of chemicals from spectral data. To assess the predictive performance of these models, we used two commonly used measures: mean squared error (MSE) and root mean squared error (RMSE). MSE is calculated as the average of the squared differences between the predicted values and the actual values. On the other hand, RMSE is the square root of the MSE and measures the typical magnitude of the error in the predictions. A lower value of the MSE indicates that the model has a smaller average error in its predictions, while a higher value of the MSE indicates a larger average error. Similarly, a lower value of the RMSE indicates that the model has a smaller typical magnitude of error, while a higher value of the RMSE indicates a larger typical magnitude of error[55] [7].

MSE:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2. \quad (3.19)$$

RMSE:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (3.20)$$

where  $y_i$  are the actual values,  $\hat{y}_i$  are the predicted values,  $n$  is the number of observations.

#### Cross Validation:

In the  $K$ -fold cross-validation test, the original dataset undergoes a randomized partition into  $K$  sets. These sets are iteratively employed as validation datasets, with the remaining  $K - 1$  sets serving as training datasets [49]. Mathematically, if  $N$  is the total number of data points, each of the  $K$  folds will contain approximately  $\frac{N}{K}$

data points. During each iteration, the model is trained on  $K - 1$  folds and validated on the remaining fold. This process is repeated  $K$  times, ensuring that each data point experiences inclusion in both training and validation phases exactly once. The performance metric for each iteration is denoted as  $\text{RMSE}_j$ . The overall average RMSE, denoted as  $\overline{\text{RMSE}}$ , is calculated as follows:

$$\overline{\text{RMSE}} = \frac{1}{K} \sum_{j=1}^K \text{RMSE}_j, \quad (3.21)$$

where  $\text{RMSE}_j$  represents the RMSE for the  $j$ -th iteration of the cross-validation process.  $K$ -fold cross-validation, using RMSE as the evaluation metric, provides a robust assessment of the model’s generalization performance across different subsets of the data, aiding in the prevention of overfitting or underfitting.

### 3.2.3 Parameter Tuning

Parameter tuning is a crucial aspect of employing the proposed multi-layer methods. These methods often rely on various parameters significantly influencing their performance and predictive accuracy. The importance of parameter tuning lies in its ability to enhance the generalization capabilities of the proposed methods. By carefully adjusting these parameters, we ensure the models can effectively capture underlying patterns and relationships within the data.

The parameter tuning process is not arbitrary or subjective; rather, it is a systematic approach to optimize the model’s performance. Through various techniques like random search and cross-validation, we iteratively explore different combinations of parameter values to identify the configuration that yields the best results. This rigorous approach empowers us to fine-tune the models, making them better suited to handle diverse datasets and real-world scenarios.

### 3.2.3.1 Tuning Number of Components

In regression models like PLS and PCR, the number of components is crucial in determining model performance. These techniques are especially useful for handling high-dimensional data in spectroscopic datasets, where noise and multicollinearity can pose significant challenges. High-dimensional spectroscopic datasets often contain redundant or irrelevant information, leading to multicollinearity, where predictor variables are highly correlated. This can adversely affect the stability of the regression models. By tuning the number of components, we aim to capture the most relevant and informative latent factors while reducing the impact of noise and multicollinearity. PLS and PCR are dimensionality reduction techniques that project the original high-dimensional data into a lower-dimensional space represented by the chosen components. These components are combinations of the original variables that explain the maximum variability in the data. By selecting an appropriate number of components, we can retain enough information to make accurate predictions while mitigating the effects of noise and multicollinearity [50] [56]. Selecting the best number of components is extremely important to avoid adding noise or losing important information in the data.

### 3.2.3.2 Tuning Scaling Factor for DLR Method

Tuning the scaling factor, denoted as  $s$ , during the DLR method in Section 3.2.2.2 is important as it significantly impacts the accuracy and effectiveness of the predictions. Here,  $s$  serves as a critical parameter that determines the width of the response variable range in each layer of the method. Finding the right balance for  $s$  is crucial because it involves a trade-off between two essential factors.

On one hand, if  $s$  is set too high, the resulting range becomes excessively wide. In

this scenario, the range could encompass a large amount of data, potentially reverting back to the initial estimate obtained from the first-layer regression model.

On the other hand, setting  $s$  too low results in very narrow ranges. While this may capture localized information around each rough estimate more precisely, it comes at the cost of excluding potentially relevant data points that lie just outside the narrow range. As a result, critical information might be lost, leading to suboptimal performance and reduced accuracy in the final predictions.

Hence, finding an appropriate value of  $s$  is crucial to strike the right balance between these trade-offs. It allows the DLR method to dynamically adjust the width of the range, ensuring that the subsequent regression models are trained on relevant and informative subsets of the data.

### **3.2.3.3 Tuning Number of Classes in the CLR Method**

The number of classes is another critical parameter in the effectiveness of the proposed CLR method. The choice of the number of classes directly affects how the data is categorized, subsequently impacting the training of the regression models. To ensure a robust and objective classification model, it is important to make informed decisions about the number of classes. This is especially important when dealing with a wide range of response variables.

We employ a systematic approach to exploring different class configurations. By selecting different setups and comparing two and three classes, we gain valuable insights into how the classification results differ. This comparative analysis allows us to understand how the choice of the number of classes affects the method's performance, especially when dealing with data that spans a wide range. The method's sensitivity to the number of classes becomes particularly significant in scenarios where the target variable varies significantly across the data. By tuning this parameter and

investigating the classification outcomes, we can tailor the model to achieve optimal results, ensuring it captures the underlying patterns effectively and adapts well to diverse datasets.

### 3.3 Results and Discussion

#### 3.3.1 Data Summary and Preliminary Analysis

In Figure 8, we observe a collection of 911 UV-Vis spectra, each coded from blue to red, representing concentrations in the dataset. The x-axis corresponds to wavelength, and the y-axis shows absorbance levels. This dataset stands out for its high dimensionality due to the abundance of curves. Notably, red curves indicate higher concentrations, while blue curves signify lower concentrations. The connection between concentration and absorbance levels is evident, with red shades signaling increased absorption. Moreover, the wavelength shows two distinct absorbance peaks, contributing to non-linearities in the dataset. Also, correlations among variables are evident throughout the entire dataset, offering a clear depiction of the interconnected nature of the spectral variables. This observation proves instrumental in unraveling patterns of association and potential dependencies within the intricate UV-Vis spectroscopic dataset.

Figure 9 shows two boxplots of the concentrations. As shown in the left boxplot, the concentrations in this dataset exhibit a wide range, spanning from  $1e-06$  to  $9e-02$ . The second boxplot, which follows the log transformation, demonstrates that the concentrations become more uniformly distributed around the median value. This transformation results in a more balanced distribution, helping to mitigate the impact of the wide concentration range.

Given the repetitive nature of the dataset (explained in Section 4.2), wherein



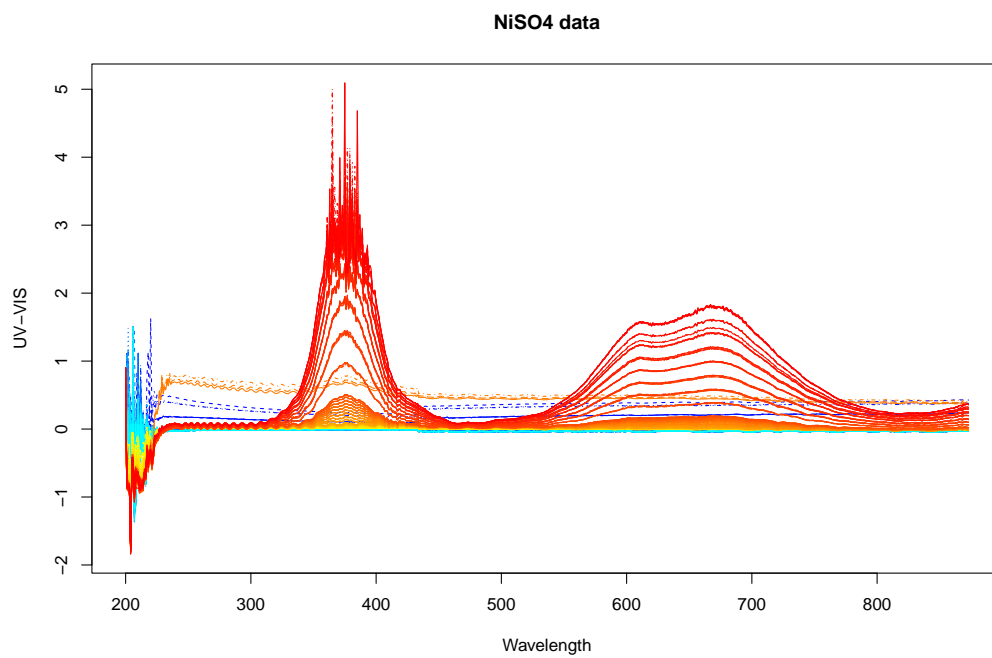


Fig. 8.: UV-Vis spectral measurements.

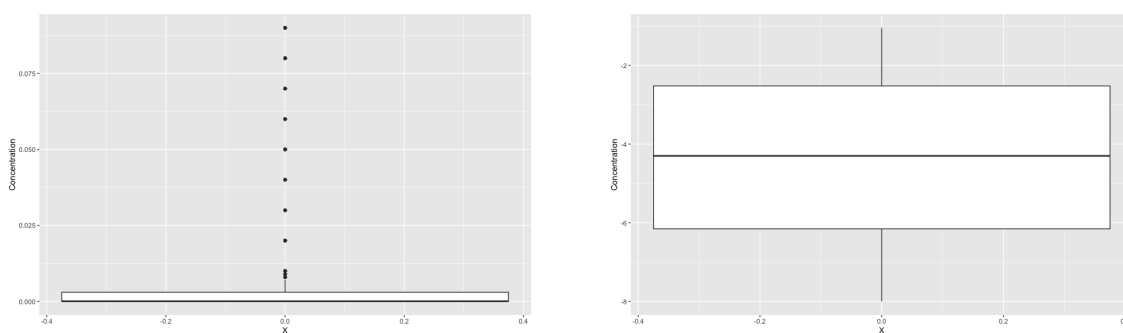


Fig. 9.: Boxplots of the concentrations before (left) and after (right) the log transformation.

each set of nine curves corresponds to a single concentration, we calculated the average of every three curves. This decision was guided by the inherent structure of the provided data, where measurements are repeated three times for each concentration. Averaging every three curves allows us to capture the underlying trend more robustly by mitigating the impact of potential outliers or fluctuations within each set of three measurements. This approach aligns with the intrinsic repetition in the dataset, providing a more reliable representation of the concentration-specific spectral characteristics and enhancing the overall stability of the analysis. This can help to alleviate redundancy within the dataset by condensing the information related to each concentration. By doing so, we streamline the data while preserving its essential characteristics, making it more manageable and easier to work with and preparing it for further analysis. Figure 10 shows a pair of plots showing the nine and the three curves of three different concentrations. In the first set of plots on the left, where we have nine curves representing a specific concentration, we see all the fine details and variations in the data. On the right set, the plots show a simpler representation with only three curves, achieved through averaging. Importantly, both sets of plots demonstrate that our data reduction process hasn't altered the essential characteristics of the data or these individual curves. We've managed to keep the main patterns, variations, and trends intact.

### **3.3.2 Predictive Models Tuning**

As explained in Section 3.2, our approach involves using two regression models, PLS and PCR. We further improved these models by integrating two innovative multi-layer models, DLR and CLR. These modifications aim to address challenges related to chemical concentrations in spectral data. We employed a cross-validation method to fine-tune various parameters, such as the number of components for PLS

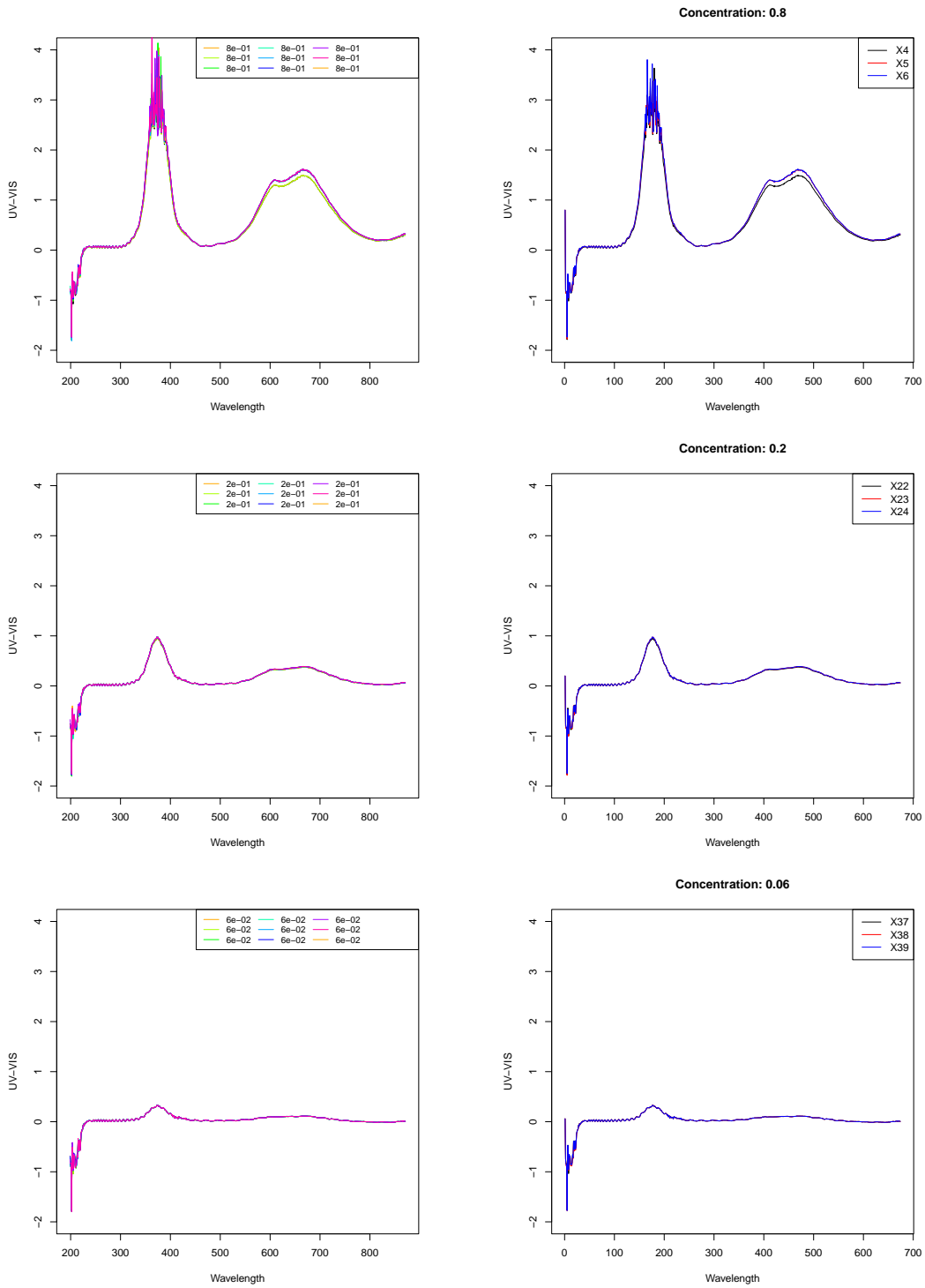


Fig. 10.: Sample of concentrations curves before the average (left with nine curves) and after the averaging (right with three curves).

and PCR models, the scaling factor in the DLR method, and the number of classes in the CLR method. Finally, we compared the original PLS and PCR models with the proposed methods using the evaluation metrics discussed in Section 3.2.2.4. We first split the entire dataset into training and testing datasets. With a split proportion of 70% for the training set with 131 observations and 30% for the testing set with 56 observations. Figure 11 depicts the results of tuning the number of components for both PLS and PCR models. The  $x$ -axis spans the range of 1 to 100 components, while the  $y$ -axis illustrates the RMSE obtained through cross-validation. Notably, the plots reveal that PLS achieved its minimum RMSE with 17 components, whereas PCR attained the lowest RMSE with 39 components. The resulting plots exhibit a U-shape, wherein both low and high numbers of components lead to high RMSE, while the minimum RMSE is observed at a moderate number of components. This characteristic U-shape pattern indicates the trade-off between model complexity and predictive performance. When the number of components is too low, the model may oversimplify, resulting in high RMSE due to an inability to capture essential patterns in the data. Conversely, an excessively high number of components may lead to poor prediction. Therefore, identifying the point where the U-shape occurs is pivotal for selecting an optimal number of components, striking a balance that ensures both model simplicity and accuracy in prediction. This nuanced approach to tuning contributes to the models' robustness in handling unseen data. This process of tuning the number of components is crucial as it helps optimize the model's performance. Achieving the minimum RMSE signifies the ideal configuration for each model, contributing to their effectiveness in the prediction task.

For the DLR method, we also tuned the value of  $s$ , which determines the width of the concentration range in each layer of the method. Figure 12 shows the results of this tuning process, illustrating the relationship between  $s$  values range 1 : 300 and RMSE

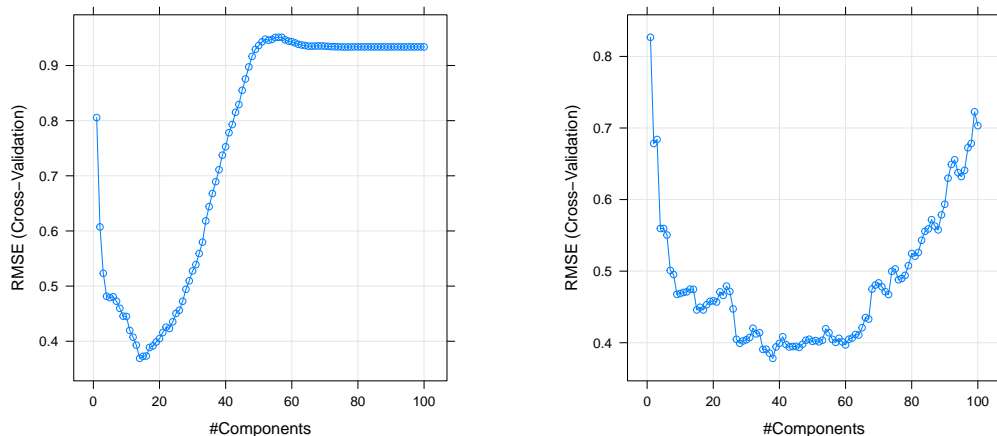


Fig. 11.: Number of components vs RMSE values for PLS (left) and PCR (right).

for both DLR-PLS and DLR-PCR models. Interestingly, both plots exhibit what is known as elbow shape, where the plot often resembles an arm, and the “elbow” point is where the performance improvement begins to slow down, suggesting that adding larger  $s$  values does not significantly improve the model’s performance. The figure indicates that extreme values of  $s$  lead to higher RMSE, while an optimal balance is achieved with balanced values. Specifically, for DLR-PLS, the minimum RMSE is attained at  $s = 70$ , while for DLR-PCR, the minimum occurs at  $s = 69$ .

These findings align with the theoretical considerations discussed earlier. Setting  $s$  too high widens the concentration range excessively, potentially incorporating irrelevant data and returning to the initial estimate. Conversely, setting  $s$  too low creates overly narrow ranges, sacrificing relevant data and compromising the model’s accuracy. The observed minima for  $s$  in the plots reflect the optimal trade-off, allowing the DLR Method to dynamically adjust the range width, ensuring that subsequent DLR models capture informative subsets of data for precise concentration predictions. Additionally, during the DLR process, we further refined our approach by tuning the

number of components when re-fitting the regression models for each layer. The inherent variability in the number of observations across layers, determined by the width of the concentration range, necessitated a nuanced adjustment of the number of components. Recognizing that each subset of the DLR models demands a tailored tuning of the number of components, we aimed to ensure the selection of an optimal number for each layer. This adaptive tuning strategy acknowledges the diverse characteristics of the data subsets within the DLR method, contributing to the precision and effectiveness of the regression models.

Table 6 shows a sample of new concentration ranges and the corresponding count of observations within each layer using the optimal  $s$  value for both DLR-PLS and DLR-PCR models. Notably, the table shows the impact of using only neighboring observations that are related to a specific concentration. This approach results in a narrower concentration range than all training observations with a wide range (1e-06 to 9e-02). By considering these relevant observations within a more constrained range, there is a potential enhancement in prediction performance. This approach contrasts with standard regression models that incorporate all observations, as the targeted focus on relevant data points contributes to refined and potentially more accurate modeling results.

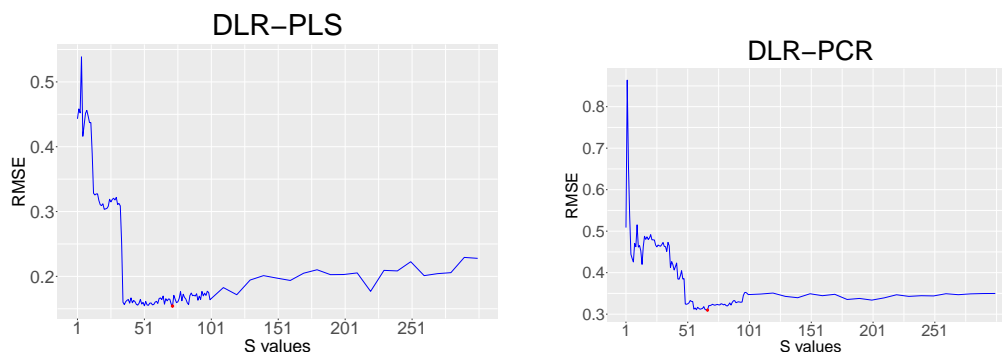


Fig. 12.:  $s$  values vs RMSE for DLR models

Concentration	Range (Count) DLR-PLS	Range (Count) DLR-PCR
1e-06	1e-06, 1e-05 (14)	1e-06, 1e-05 (14)
2e-06	1e-06, 3e-05 (17)	1e-06, 3e-05 (17)
3e-06	1e-06, 1e-04 (29)	1e-06, 1e-04 (29)
4e-06	1e-06, 2e-04 (32)	1e-06, 1e-04 (29)

Table 6.: A example of the new range and number of observations in each layer for both DLR-PLS and DLR-PCR models.

For the CLR method, we will tune the number of classes through various configurations to identify the most effective classification approach. Hence, we partitioned the training data into two configurations (Two classes and Three classes). The classes in the two configurations, named “High” and “Low”, leverage the median concentration value as the threshold. Similarly, in the three-class configuration, the classes are named “High”, “Middle”, and “Low”. This approach utilizes quantile-based thresholds, dividing the concentration values into three distinct groups, each representing a third of the data range as calculated from the training dataset. Specifically, we use the one-third and two-thirds markers of the concentration distribution as thresholds to define these categories. This ensures a balanced distribution of data between the two classes. The effectiveness of both configurations was assessed through cross-validation, using RMSE as the metric. The evaluation revealed that the two-class configuration had a superior performance, registering an average RMSE of 0.22, compared to the three-class configuration, which had an average RMSE of 2. This assessment was conducted for the PLS model, and similar outcomes were observed when applying the PCR model. Consequently, we selected the two-class configuration for implementing the CLR method as described below.

The two-class classification resulted in 72 concentrations being assigned to the High class and 59 concentrations to the Low class. Utilizing the median for data classification offers objectivity in-class assignment. Subsequently, the PLS-DA model was trained on the classified training data to predict the classes for the testing dataset. The purpose of building the PLS-DA model is to leverage its ability to classify observations with unknown concentrations in the testing data. The model identified 25 concentrations in the High class and 31 concentrations in the Low class. The classification performance is shown by the confusion matrix presented in Table 7. The matrix reveals that out of 51 correctly classified observations, 20 belong to the Actual High class and 31 to the Actual Low class. However, the model only misclassified 5 observations. The accuracy of the PLS-DA model is calculated as the ratio of correctly classified observations to the total number of observations, resulting in an accuracy rate of 91%. This robust classification performance demonstrates the effectiveness of the PLS-DA model in accurately assigning concentrations to their respective classes. This approach significantly narrowed down the concentration range within each class. Originally spanning from (1e-06, 9e-02), the concentration range for the High class now lies between (1e-06, 5e-05), while the Low class is confined to the range (6e-05, 9e-02) as demonstrated in Table 8. This reduction in concentration range provides a more focused and specific context for the subsequent prediction task using CLR-PLS and CLR-PCR models. By classifying the concentrations into distinct classes, the model is better equipped to make precise predictions within these refined ranges.

	Predicted High	Predicted Low
Actual High	20	0
Actual Low	5	31

Table 7.: Confusion matrix of PLS-DA model.



Data	Range
Full data	1e-06, 9e-02
Classified data (Low class)	1e-06, 5e-05
Classified data (High class)	6e-05, 9e-02

Table 8.: The concentrations range of High and Low classes.

### 3.3.3 Predictive Models Comparison

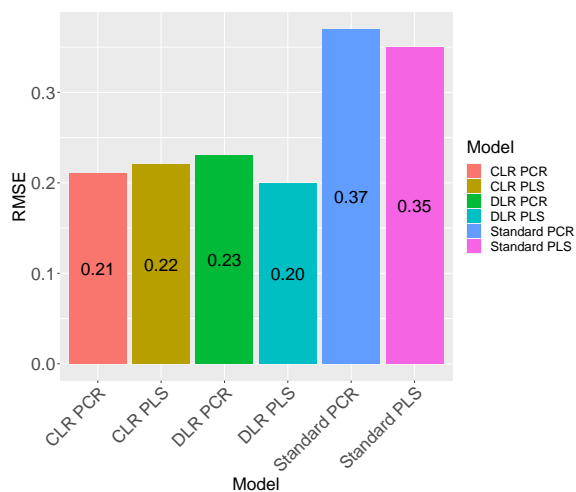


Fig. 13.: Predictive accuracy comparison by all models.

Table 9 and Figure 13 display the RMSE values for six models, highlighting the consistent outperformance of our proposed multi-layer models in comparison to the standard models, which establish the baseline performance, with RMSE values of 0.35 and 0.37, respectively. The CLR models exhibit notable improvement, reducing RMSE to 0.22 for CLR-PLS and 0.21 for CLR-PCR. Similarly, the DLR models demonstrate significant performance, particularly with DLR-PLS reducing RMSE by 43% to 0.20. The success of our proposed multi-layer models lies in their effectiveness in addressing challenges associated with wide concentration ranges. The proposed models excel in capturing local linearity and narrowing concentration ranges, result-

ing in a significant enhancement in predictive accuracy compared to the standard models. Figures 14 and 15 show actual concentrations and predicted values for the four proposed models. Due to the wide range of actual values, traditional plotting made details less visible. To improve visualization, a log transformation was applied to both actual and predicted values. This compresses the scale, making patterns more apparent by reducing the impact of outliers. The plots of the proposed models clearly demonstrate the improved fit of the regression line compared to the standard models, especially at high concentrations, addressing the challenges associated with a wide concentration range.

Model	RMSE (PLS)	RMSE (PCR)
Standard models	0.35	0.37
CLR models	0.22	0.21
DLR models	0.20	0.23

Table 9.: RMSE values for the standard and proposed models

### 3.4 Conclusions

This study tackles the challenge of predicting concentrations in spectroscopic data, a task rendered complex by high dimensionality, multicollinearity, non-linearity, and especially the wide range of concentrations. The introduction of DLR and CLR methods marks a significant stride in addressing these issues. Both methods significantly reduce the wide concentration range, with DLR employing dynamic layering and CLR using strategic data classification to enhance predictive accuracy. Among them, the DLR-PLS method particularly excels, demonstrating its efficacy in managing the wide range concentration, as reflected by its lower RMSE values. Crucially, the success of these methods is also attributed to careful parameter tuning, underscoring its importance in optimizing their performance. This research, therefore,

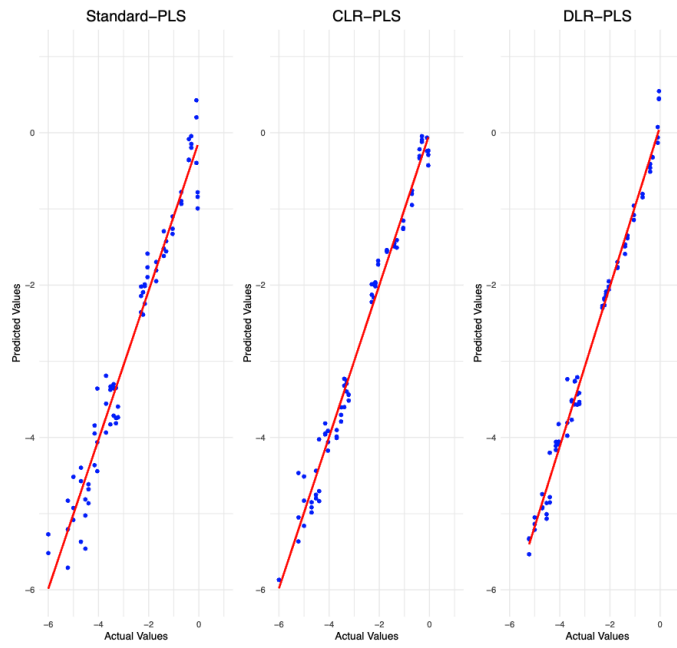


Fig. 14.: Actual concentrations vs predicted values of PLS model.

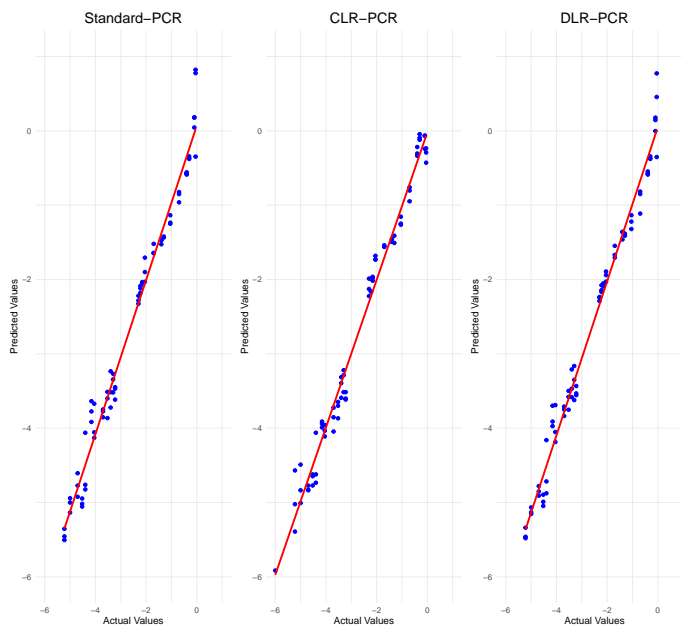


Fig. 15.: Actual concentrations vs predicted values of PCR model.

not only provides effective solutions for complex spectroscopic analysis but also sets a new benchmark in chemometric methodologies, enabling more accurate and insightful concentration predictions.

## CHAPTER 4

### MULTIPLE CHANGEPOINT DETECTION FOR AUTOCORRELATED ORDINAL TIME SERIES

#### 4.1 Introduction

Changepoint detection is a powerful method used to identify significant changes in the statistical properties of time series data. A changepoint can signal a change in the time series data's mean, variance, or correlation structure. This technique is applicable in diverse fields, including climate, healthcare, and many others. In climate science, for instance, changepoint detection can help identify changes in climate patterns, such as shifts in temperature, precipitation, and extreme weather events, providing insights into the impacts of climate change on ecosystems and human societies [57]. Also, in the healthcare system, changepoint detection can be applied to identify critical changes in patient health status over time, which can aid in diagnosing and treating various diseases [58]. Overall, changepoint detection is a crucial tool for recognizing significant shifts in data patterns. By identifying these shifts, we gain a deeper understanding of how various systems change over time.

##### 4.1.1 Advancing from Single to Multiple Changepoint Detection

There is a myriad of changepoint detection methods available. Several studies have focused on detecting a single changepoint known as At Most One Changepoint (AMOC) detection, which assumes that a time series data contains at most one single changepoint. Various methods can be used to estimate a single changepoint in time series data. Shi et al [59] compared the effectiveness of several AMOC meth-

ods, including various CUSUM and likelihood ratio statistics tests (LRT). However, applications that involve longer time series may require more nuanced assumptions.

In these contexts, the assumption of AMOC may not be appropriate, as there may be multiple transitions or gradual changes over time [60] [8]. To address this issue, multiple changepoint detection methods have emerged. These methods are designed to detect multiple changepoints in a time series and estimate the time and location of each change. According to [60], the detection of multiple changepoints in time series analysis can be traced back to the 1980s. In the simplest piecewise stationary model, the time series is divided into  $m + 1$  distinct regimes by  $m$  unknown changepoints occurring at times  $1 < \tau_0 < \tau_1 < \dots < \tau_{m+1} \leq N + 1$ , with boundary conditions  $\tau_0 = 1$  and  $\tau_{m+1} = N + 1$ . Each regime has its own distinct mean and contains the data points  $X_{\tau_i+1}, \dots, X_{\tau_{i+1}}$ .

The model can be expressed as a simple piecewise function:

$$X_t = \mu_{r(t)} + \epsilon_t, \quad (4.1)$$

where  $X_t$  is the observed value at time  $t$ ,  $r(t)$  the regime index at time  $t$ ,  $\mu_r(t)$  denotes the regime mean, and  $\epsilon_t$  represents the stationary causal and invertible ARMA( $p, q$ ) noise term. The task of multiple changepoint detection is to estimate the number and locations of changepoints. This problem is particularly challenging when the number and locations of change points are unknown [61]. In the following section, we will present some popular techniques utilized to detect multiple changepoints.

#### 4.1.2 Multiple Changepoint Detection Methods

The various techniques for multiple changepoint detection can be categorized into two main groups: sequential binary segmentation methods and model selection methods [60]. The binary segmentation approach, first introduced by Scott and Knott

[62], uses any AMOC method to estimate multiple changepoint configurations. The process begins by testing the entire time series for a single changepoint. Once a changepoint is identified, the series is divided into two subsegments, and each subsegment is analyzed for additional changepoints using the AMOC strategy. This process is repeated until no subsegment shows any evidence of a changepoint or a stopping criterion is met. The binary segmentation method is particularly effective when the changepoints are well separated, and the means of each segment are distinct.

However, basic binary segmentation struggles with detecting short segments within long segments [59] [63]. Therefore, Olshen et al. [64] proposed Circular Binary Segmentation (CBS) to improve the basic binary segmentation by splicing the segment ends into a circle and searching for short segments within it. Another extension to binary segmentation is Wild Binary Segmentation (WBS) which was introduced by Fryzlewicz [65]. WBS works by randomly partitioning the data into smaller segments and applying binary segmentation to each segment independently. This allows WBS to detect changepoints at different scales and reduces the algorithm's computational complexity.

Another approach to multiple changepoint detection involves fitting a model to the data and then using the model to estimate the number and locations of changepoints. Since there are many possible changepoints in a time series, a penalty term can be added to the objective function to discourage the detection of too many changepoints. This is because too many changepoints can lead to overfitting of the model. The penalty term is added to the objective function to discourage the model from fitting the data too closely. This is done by adding a cost to the model for each detected changepoint. The cost of each changepoint is proportional to the penalty term. The penalty term is chosen so that the model is penalized more for detecting too many changepoints than for not detecting enough changepoints [59]:

$$F(\boldsymbol{\theta}) = C(\boldsymbol{\theta}) + \lambda P(\boldsymbol{\theta}), \quad (4.2)$$

where  $\boldsymbol{\theta}$  represents the model parameters, including  $m$  changepoints at the times  $\tau_1, \dots, \tau_m$  and other parameters,  $C(\boldsymbol{\theta})$  is the cost function,  $P(\boldsymbol{\theta})$  is the penalty function, and  $\lambda$  is the penalty parameter that controls the trade-off between goodness of fit and model complexity. The cost function  $C(\boldsymbol{\theta})$  quantifies the cost or loss associated with a particular model or set of model parameters. Where lower values of the cost function indicate better model fit. The penalty function  $P(\boldsymbol{\theta})$  is used to discourage the detection of too many changepoints, as this can lead to the overfitting of the model. It is typically a function of the number and locations of the changepoints detected by the model. The specific form of the penalty function depends on the problem being addressed. The goal of multiple changepoint detection is finding a configuration that minimizes the objective function  $F(\boldsymbol{\theta})$  with respect to the model parameters  $\boldsymbol{\theta}$ . This involves finding the optimal number and locations of the changepoints that best explain the data while balancing goodness of fit and model complexity [59].

Equation 4.2 incorporates a penalization term to the objective function in two different methods: penalized least-squares-based methods and penalized likelihood-based methods. The method of least squares and maximum likelihood are two approaches that can be used to estimate parameters in a statistical model. Least squares minimize the sum of the squares of the differences between the observed data and the model prediction, while maximum likelihood maximizes the likelihood function, which is the probability of observing the provided data given the parameters. Ridge Regression and Least Absolute Shrinkage and Selection Operator (LASSO) are popular penalized least-squared methods. On the other hand, common penalized likelihood methods involve the Akaike Information Criterion (AIC), the Modified Bayesian Infor-



mation Criterion (mBIC), and the standard BIC. Also, Minimum Description Length (MDL) is another popular penalty term [66] [59]. Bleakley et al. [67] proposed a multiple changepoint detection method using the group fused LASSO, a least squares penalized regression technique that promotes sparsity and clustering of the changepoints. The study demonstrated that the proposed method outperformed other existing methods when tested on different high-dimensional biological datasets. Also, Lu et al. [61] used likelihood penalization with an MDL as the penalization term to perform multiple changepoint detection. The method was evaluated on a simulated dataset and was found to be effective in estimating the number and location of changepoints. The method was also applied to a century of monthly temperatures from Tuscaloosa, Alabama, and was able to identify the changepoints that occurred in the data. In general, the likelihood-based approach is a more flexible and robust method for estimating model parameters than the least squares-based approach. This is because the likelihood-based approach does not carry any assumptions about the data distribution, while the least squares-based approach assumes that the data is normally distributed. On the other hand, the likelihood-based approach is more computationally expensive than the least squares-based approach. However, the extra computational cost of the likelihood-based approach is justified because it provides a more complete and accurate picture of the uncertainty in the model parameters [66].

### 4.1.3 Multiple Changepoints in Categorical Time Series

In many cases, categories are employed to represent various aspects, such as system states or user behaviors. The significance of analyzing categorical data is growing steadily within industrial settings. Detecting changes in categorical data enables us to detect shifts in the distribution of categories over time. By recognizing and understanding these changes, we gain valuable insights into the system, allowing us to

take corrective measures and prevent potential issues from arising [68]. If the categorical variables possess an inherent order or ranking, incorporating this ordinality into the analysis can provide valuable information and improve the accuracy of detecting meaningful changes [69]. Within the domain of changepoint detection, the significance of categorical data has yet to be considered by numerous studies, resulting in less accurate changepoint detection. However, there are studies that actively take into account the categorical features when conducting changepoint detection. For example, Wang et al. [70] used a modified log-linear model that incorporated a continuous latent variable to represent the underlying structure of the ordinal categorical data. This modification allowed them to effectively capture and analyze the ordinal information within their changepoint detection approach. Moreover, changepoint detection methods for categorical data often assume independent, identical, and distributed errors. However, such assumptions don't apply to hourly, daily, or monthly data with recurring patterns and strong autocorrelations. As stated in [8], ignoring these correlations can significantly distort changepoint findings, even confusing positive autocorrelation with a mean shift. Hence, Li and Lu [8] employed a CUSUM type test to detect a single changepoint within autocorrelated ordinal categorical time series data. Where an Autoregressive Ordered Probit (AOP) model is used as an underlying model to describe ordinal categorical time series. The study was demonstrated through a simulation and applied to real-world rainfall time series data categorized by location in Albuquerque, New Mexico. Although these studies have investigated changepoint detection for ordinal categorical time series data, a research gap exists in multiple changepoint detection over autocorrelated ordinal categorical time series data, based on the current extent of our knowledge. This unexplored domain constitutes the foundational emphasis of our ongoing study.

This chapter primarily focuses on detecting multiple changepoints within auto-

correlated ordinal time series data. Our methodology encompasses a comprehensive approach, leveraging the AOP model for a nuanced analysis of such data. Parameter estimation within the AOP model is achieved through the pairwise likelihood method, which supports our objective function and is further refined by a penalized likelihood incorporating a penalty term. We employ a genetic algorithm for efficient random walk searches to address the challenge of the extensive array of potential multiple changepoint configurations. Among varying patterns and trends, it becomes evident that not all potential changepoints significantly impact the data’s structural interpretation. This insight prompts the introduction of the “effective changepoint” concept, which distinguishes between statistical anomalies and meaningful shifts in the time series. For comparative analysis and methodological enhancement, binary segmentation, combined with the CUSUM type test, is utilized within the framework of the AOP model used in [8]. Finally, The method is utilized on Los Angeles’ Air Quality Index (AQI) data aiming to detect changes in air quality over daily data.

## 4.2 Methodology

### 4.2.1 AOP Model

In this section, we will adopt the AOP, which was introduced by Müller and Czado [71]. An AOP model can be seen as a dynamic expansion of the familiar ordered probit OP model. The AOP model keeps the conventional regression component of the OP model while incorporating an autoregressive part [72].

Consider a discrete response time series represented as  $\{X_t, t = 1, \dots, T\}$ . The values of  $X_t$  come from an ordered collection  $\{1, \dots, K\}$ , where  $K$  is an integer larger than 1. Within this framework, there exists an underlying latent continuous process  $\{Z_t, t = 1, \dots, T\}$  that generates  $X_t$  through a clipping mechanism. In other words,

$X_t$  is determined by which interval  $Z_t$  falls into:

$$X_t = k \iff Z_t \in \{c_{k-1}, c_k\}, \quad k \in \{1, \dots, K-1\}, \quad (4.3)$$

and

$$Z_t = \mu_t + \phi(Z_{t-1} - \mu_{t-1}) + \epsilon_t. \quad (4.4)$$

The set of unknown threshold parameters that determine the boundaries of the intervals denoted as  $\{c_1, c_2, \dots, c_{K-1}, c_K\}$ , satisfies the conditions:

$$-\infty = c_0 < c_1 < c_2 < \dots < c_{K-1} < c_K = \infty.$$

In Equation 4.4,  $Z_t$  represents the latent process at time  $t$ . Where  $\mu_t$  denotes the mean of  $Z_t$  at time  $t$ , which can vary with time. With  $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ , representing independent Gaussian distributed random variables. Additionally,  $\phi$  is a parameter associated with the autoregressive component. We assume that the autoregressive part of the latent process is stationary, i.e.,  $|\phi| < 1$ . It's important to recognize that the AOP model differs from the standard cumulative probit model solely due to the inclusion of the parameter  $\phi$ , which brings about dynamic behavior. We set  $\sigma^2$  to equal  $1 - \phi^2$  in our model to avoid the identifiability problem between parameters. By doing that, we ensure that the variability of the latent variable  $Z_t$  remains constant throughout the time series. In this context, the mean  $\mu_t$  is expressed as follows:

$$\mu_t = \alpha_0 + \Delta I_{(t \geq \tau)}, \quad (4.5)$$

where  $\alpha_0$  and  $\Delta I_{(t \geq \tau)}$  represent the intercept and a magnitude with a changepoint indicator, respectively. We adopt the binary segmentation discussed in Section 1.2 utilizing the CUSUM type test with an AOP model employed in [8]. The AOP model, tailored for ordered categorical responses, incorporates an autoregressive structure to

capture temporal dependencies. Once a single changepoint is identified using the CUSUM test, binary segmentation comes into play. This recursive application of binary segmentation on the subsegments allows for identifying additional changepoints. The iterative process continues until no further changepoints are detected or a predetermined stopping criterion is met. We will employ this combined binary segmentation method and the CUSUM test (BS-CUSUM) as a baseline approach for multiple changepoint detection, against which we will evaluate the performance of our proposed method. In [8], a reconstructed latent variable, denoted as  $\tilde{Z}_t$  was introduced as an estimate of the latent variable  $Z_t$  at time  $t$ , based on the observed categorical variable  $X_t$ . It is calculated as the conditional expectation  $E(Z_t|X_t)$ , representing the most likely value of  $Z_t$  given  $X_t$  alone.

To estimate the parameters of the AOP model, we assume that the number of changepoints and their locations are known. Let  $\boldsymbol{\theta} = (c_2, \dots, c_{k-1}, \alpha_0, \Delta, \phi)$  be a vector includes all the AOP model parameters. However, estimating parameters for the AOP model can present challenges, especially when dealing with a full likelihood function involving a high-dimensional Gaussian integral that's difficult to numerically assess [8]. As mentioned in [72], using a frequentist approach based on the likelihood for parameter estimation in the AOP model can lead to high computational costs. The estimation process of AOP models can be done using a range of methods. As an example, Niu et al [60] introduced the utilization of Markov chain and hidden Markov chain modeling for probit models. However, Varin and Vidoni [72] raised concerns about the hidden Markov approach, pointing out that the number of parameters exponentially increases with the chain's order. In response, they introduced an alternative estimation technique based on pairwise composite likelihood, leveraging bivariate Gaussian probabilities to address these issues. From a different perspective, Müller and Czado [71] employed a Bayesian inference approach to address parameter

estimation in the AOP model. Also, Li and Lu [8] considered a sequential parameter estimation method to estimate the parameters of the AOP model. In the next section, we will use the pairwise likelihood estimation method introduced in [72] for fitting the AOP model.

#### 4.2.2 Pairwise Likelihood Function for AOP Models

The AOP model assumes that the observed ordered categorical responses  $X_t$  at time points  $t = 1, 2, \dots, n$  are generated from a latent continuous variable  $Z_t$ . The likelihood is the following Gaussian integral with  $n$  dimensions:

$$L(\boldsymbol{\theta}, X_1, \dots, X_n) = P(X_1, \dots, X_n; \boldsymbol{\theta}) = \int_{B(X_1, \dots, X_n)} P(Z; \boldsymbol{\theta}) dZ_1 \cdots dZ_n, \quad (4.6)$$

where  $B(X_1, \dots, X_n) = \{Z = (Z_1, \dots, Z_n) : c_{X_{i-1}} < Z_i \leq c_{X_i}, i = 1, \dots, n\}$  and  $p(Z; \boldsymbol{\theta})$  is the joint Gaussian density of  $Z = (Z_1, \dots, Z_n)$ . Hence, when trying to assess the classical likelihood function and compute the corresponding maximum likelihood estimator, the process can become quite challenging due to the high dimensions of  $n$  [72]. To address this challenge, we will approach it using the pairwise likelihood function. The key advantage of using pairwise likelihood is that it significantly reduces the computational load, making it feasible to handle high dimensional datasets [73]. Here, we create bivariate marginal distributions known as pairwise functions. Here, we consider the probability of observing pairs of responses  $(X_i, X_j)$  for all combinations of time points  $i$  and  $j$  where  $i < j$ . The pairwise likelihood  $L_{PL}(\boldsymbol{\theta}, X_1, \dots, X_n)$  is the product of the bivariate probabilities for all unique pairs of observations:

$$L_{PL}(\boldsymbol{\theta}, X_1, \dots, X_n) = \prod_{i=1}^{n-1} \prod_{j=i+1}^n L_{ij}(\boldsymbol{\theta}; X_i, X_j), \quad (4.7)$$

where  $L_{ij}(\boldsymbol{\theta}, X_i, X_j) = P(X_i, X_j; \boldsymbol{\theta})$ ,  $i = 1, \dots, n-1$ ,  $j = i+1, \dots, n$  and its loglikelihood is  $\sum_{i=1}^{n-1} \sum_{j=i+1}^n \log(P(X_i, X_j; \boldsymbol{\theta}))$ .

Here, we will apply the pairwise likelihood method to the likelihood function of the AOP model in Equation 4.6:

$$L_{\text{PL}}(\theta; X) = \prod_{i=1}^{n-1} \prod_{j=i+1}^n P(X_i, X_j; \theta) \quad (4.8)$$

$$= \prod_{i=1}^{n-1} \prod_{j=i+1}^n \int_{B(X_i, X_j)} P(z_i, z_j; \theta) dz_i dz_j \quad (4.9)$$

$$= \prod_{i=1}^{n-1} \prod_{j=i+1}^n \left\{ F(c_{X_i} - \mu_i, c_{X_j} - \mu_j; \theta) - F(c_{X_i} - \mu_i, c_{X_{j-1}} - \mu_j; \theta) \right. \\ \left. - F(c_{X_{i-1}} - \mu_i, c_{X_j} - \mu_j; \theta) + F(c_{X_{i-1}} - \mu_i, c_{X_{j-1}} - \mu_j; \theta) \right\}, \quad (4.10)$$

where  $B(X_i, X_j) = \{(z_i, z_j) : c_{X_{i-1}} < z_i \leq c_{X_i}, c_{X_{j-1}} < z_j \leq c_{X_j}\}$ , and  $p(z_i, z_j, \theta)$  and  $F(\cdot, \cdot, \theta)$  indicate the bivariate Gaussian density and distribution function of the data pair  $(Z_i, Z_j)$ . Then, the log-likelihood function is:

$$\log(L_{\text{PL}}(\theta; X)) = \log \left\{ \prod_{i=1}^{n-1} \prod_{j=i+1}^n P(X_i, X_j; \theta) \right\} \quad (4.11)$$

$$= \sum_{i=1}^{n-1} \sum_{j=i+1}^n \log \{ P(X_i, X_j; \theta) \} \quad (4.12)$$

$$= \sum_{i=1}^{n-1} \sum_{j=i+1}^n \log \left\{ F(c_{X_i} - \mu_i, c_{X_j} - \mu_j; \theta) - F(c_{X_i} - \mu_i, c_{X_{j-1}} - \mu_j; \theta) \right. \\ \left. - F(c_{X_{i-1}} - \mu_i, c_{X_j} - \mu_j; \theta) + F(c_{X_{i-1}} - \mu_i, c_{X_{j-1}} - \mu_j; \theta) \right\}. \quad (4.13)$$

We aim to maximize this pairwise log-likelihood function in (4.13) with respect to the model parameters. The optimization process involves finding the values of the model parameters that maximize the log-likelihood function. Once optimized, these parameter values serve as the estimates that best explain the observed data within the AOP model framework.

### 4.2.3 Objective Function and Model Selection

As outlined in Section 4.1.2, we add a penalty term to the negative log-likelihood function in (4.6):

$$f_{\text{obj}} = -2 \log(L_{PL}) + \text{Penalty}, \quad (4.14)$$

where  $f_{\text{obj}}$  is the objective function, and  $-2 \log(L_{PL})$  is the negative log-likelihood pairwise function. In Equation (4.14), we need to estimate the model parameters in addition to different configurations of the number of changepoints and their locations; this process is known as a model selection problem [63]. Here, the objective function is a penalized likelihood function with two different penalty terms, MDL and mBIC:

$$\text{MDL} = -2 \log(L_{PL}) + \sum_{j=1}^{m+1} \log(\tau_j - \tau_{j-1}) + 2 \sum_{j=2}^m \log(\tau_j) + 2 \log(m), \quad (4.15)$$

and

$$\text{mBIC} = -2 \log(L_{PL}) + \frac{1}{2} [3m \log(N) + \sum_{j=1}^{m+1} \log(\frac{\tau_j - \tau_{j-1}}{N})], \quad (4.16)$$

where  $N$  is the length of the time series,  $m$  is the number of changepoints with locations  $1 = \tau_0 < \tau_1 < \dots < \tau_j < \dots < \tau_{m+1} = N + 1$ . For a comprehensive understanding of the derivation of the formulas for the aforementioned penalties, please refer to [63] and [61] for the MDL penalty. Additionally, [59] and [74] provide detailed explanations for the mBIC penalty.

The goal now is to find the value of  $m$  changepoints and locations  $\tau_1, \dots, \tau_m$  that minimize the objective function 4.14. One way to do this would be to check the optimal model for every possible combination of  $m$  and  $\tau_1, \dots, \tau_m$ . However, in a series of length  $N$ , there are  $2^{N-1}$  configurations of a number of changepoints and their locations. Since the search process is computationally intensive, the genetic algorithm offers a practical solution to expedite this process. The genetic algorithm efficiently



explores and refines potential solutions by employing principles of natural selection and evolution. This strategic exploration helps narrow the search space, enabling faster and more efficient identification of optimal configurations [75]. As a result, implementing the genetic algorithm becomes instrumental in mitigating the computational demands associated with multiple changepoint detection, making the overall process more efficient and feasible. genetic algorithm has proven effective in multiple changepoint detection, as demonstrated in various studies. For instance, Li and Lund [63] applied a genetic algorithm to detect changepoints of annual precipitation data from New Bedford, Massachusetts, and the tropical cyclone record in the North Atlantic basin. Also, Lu et al. [61] used the genetic algorithm to analyze a century of monthly temperatures from Tuscaloosa, Alabama. The following section provides a detailed exploration of the genetic algorithm process.

#### **4.2.4 Genetic Algorithm**

Genetic Algorithm (GA) begins with an initial set of individuals, each representing a potential solution to the problem at hand. Every individual, or chromosome, in this population is assessed based on the fitness function, which is a function that measures the performance of the represented solution. Individuals demonstrating higher fitness levels are more likely to be chosen as parents during the subsequent reproductive phase. During the crossover process, where genetic information is exchanged between parents, offspring (children), are generated. This exchange imparts certain characteristics of the parents to the offspring, yielding a combination of advantageous characteristics. Following this, a mutation stage introduces randomness to the population by applying random alterations to individuals with a small probability. The mutation process can enhance the overall diversity of the population and prevent the algorithm from converging to a local minimum [76] [77] [61].

The GA procedure unfolds through several key stages as the following:

**Initial Populatoin:** In the first batch, changepoint times were randomly and independently chosen for each year. The size of this initial set might vary and can be changed if necessary. We used 0.06 probability to match average changepoint numbers as stated in [63].

**Chromosome Representation:** Establishing the representation of the chromosome is an important step in the GA process. In this context, each individual is encoded as a set of parameters: the number of changepoints ( $m$ ) and the specific changepoint locations  $\tau_1, \dots, \tau_m$ . This results in a chromosome, denoted as  $u = (m, \tau_1, \dots, \tau_m)$ , forming an integer vector of length  $m + 1$ . We can see that the length of the chromosome relies on the number of changepoints in the population.

**Crossover:** The crossover process produces children in subsequent generations by combining the fitter individuals from the initial generation. Two parents, representing mother and father, are selected through a linear ranking and selection method, where selection probability is proportional to an individual's rank in optimizing the objective function. The selected parents contribute to creating the next generation's offspring through a crossover procedure, ensuring that a mother and father are distinct and not identical chromosomes.

**Mutation:** Mutation introduces chromosome diversity and prevents premature convergence. This mechanism ensures occasional exploration of genetic combinations different from the current generation, maintaining a diverse population and avoiding suboptimal solutions. Typically governed by a low constant probability, mutation involves modifying parameters, contributing to adaptability, and preventing zero probability scenarios in the admissible parameter space.

We can summarize the GA process as the following:

- Initialize a population with random solutions.
- Verify termination conditions: If the generation limit is reached, stop; otherwise, proceed.
- Assess individual fitness and select the most promising for reproduction.
- Employ crossover and mutation to create a new generation of offspring.
- Evaluate the fitness of the newly generated offspring.
- Replace the least fit individuals in the population with the offspring.
- Return to the termination check.

Common termination conditions are:

- Discovery of a solution that satisfies minimum criteria.
- Reach a fixed number of generations.
- The ranking of the generation is not improving anymore; it's reached its peak.
- Manual inspection and intervention.

Various improved versions of the GA exist. For instance, the GA process includes migration through islands. Here, the population is strategically divided into islands, enabling individuals with high fitness levels to migrate between them. Key parameters influencing this migration strategy include the number of islands (subpopulations), migration frequency, the number of migrants per event, and the method employed for selecting individuals for migration. This parallel approach facilitates a more comprehensive exploration of the solution space, ultimately enhancing the overall performance and efficiency of the GA.

### 4.2.5 Effective Number of Changepoints

In evaluating changepoint detection methods, it is crucial to compare the detected changepoints against the true number of changepoints, denoted by  $m$ . However, this direct comparison might only sometimes reflect the practical challenges encountered in various simulation settings, characterized by trends and influenced by different parameters. Recognizing that some changepoints may be inherently difficult to detect, we introduce the notion of the “effective number of changepoints.” This concept shifts the focus from the actual number of changepoints to those that are practically detectable, acknowledging the reality that certain changepoints, due to their subtlety, may not significantly impact the analysis.

To quantify the effectiveness of changepoints, we propose a methodology based on confidence intervals for each segment of the time series. These confidence intervals are defined as follows:

$$CI_{\text{upper}} = \bar{X}_{\text{seg}}[i] + Z_{\alpha/2} \sqrt{\left(\frac{1+\phi}{1-\phi}\right) / n}, \quad (4.17)$$

and

$$CI_{\text{lower}} = \bar{X}_{\text{seg}}[i] - Z_{\alpha/2} \sqrt{\left(\frac{1+\phi}{1-\phi}\right) / n}, \quad (4.18)$$

where  $\bar{X}_{\text{seg}}[i]$  is the sample mean of the  $i^{\text{th}}$  segment,  $n$  is the size of  $i^{\text{th}}$  segment,  $\phi$  is the autocorrelation parameter, and  $Z_{\alpha/2}$  represents the significance level corresponding to the desired confidence interval. This method enhances traditional confidence interval calculations by accounting for the margin of error influenced by the series’ autocorrelation. A changepoint is considered effective if the confidence intervals on either side indicate a shift to different categories of ordinal data. In contrast, a changepoint is deemed ineffective if the adjacent confidence intervals encompass the same categories, suggesting that such a changepoint may not be significantly detectable. By adopting

this criterion, our evaluation emphasizes the changepoints that meaningfully alter the data’s structure and are critical for detection.

#### 4.2.6 Evaluation Methods

This section outlines our approach to comparing various methods of detecting multiple changepoints in time series data. Our comparison is grounded in a series of simulations designed to rigorously test each method’s performance against two critical criteria: the accuracy in detecting the number of changepoints and the precision in identifying their exact locations. Firstly, to assess each method’s ability to determine the number of changepoints correctly, we compute the proportion of simulations in which the technique accurately identifies the true number of changepoints.

Shi et al. [59] present a novel approach to address the complexity of comparing multiple changepoint configurations, especially when segmentations differ in the number and exact locations of changepoints. Their method introduces a changepoint-specific distance measure that simultaneously accounts for two crucial aspects of changepoint analysis: the number of changepoints and their precise locations. This approach provides a comprehensive and balanced metric for comparing changepoint configurations. It is particularly effective when configurations have many changepoints, as it emphasizes differences in their count. In their simulations, they employed this distance measure to compare estimated changepoint configurations against the true configuration, offering a robust framework for evaluating the accuracy of multiple changepoint detection methods.

In addition to the evaluation metrics mentioned above, we will use the concept of effective number of changepoints (discussed in Section 4.2.5) to assess the detection performance. This approach considers the differences between the true effective number of changepoints and the number detected by each proposed method

for each time series in each simulation. We calculate the proportions of these differences, where ideally, a zero difference indicates a perfect match between the detected and the true effective number of changepoints, thus signifying optimal performance. Proportions concentrated at zero would indicate higher accuracy in detecting the effective changepoints. Positive differences indicate an overestimation (overdetection) of changepoints, while negative values indicate an underestimation (underdetection) of the true effective changepoints. By evaluating the proportions of these differences, we gain additional insights into the performance of changepoint detection methods, especially in their ability to discern changepoints that are substantively significant within the data.

### 4.3 Simulation Studies

In this section, we will run different simulation studies to investigate the efficiency of the proposed methods. All the simulations involve five hundred series of length  $N = 500$ . We fixed the variance of the latent process to be  $\text{Var}(Z_t) = 1$ . The autocorrelation parameter selected was  $\phi = 0.5$ , and the white noise variance was  $\text{Var}(\epsilon_t) = 1 - \phi^2$ . In our model, we set  $\kappa = \frac{\Delta}{\sqrt{(\sigma^2)}}$ , which denotes the signal-to-noise ratio, and we introduce the same changing magnitude  $\Delta$  between the adjacent changepoints. The mean shifts are adjusted based on the signal-to-noise ratio  $\Delta = \kappa\sqrt{\sigma^2}$ . The magnitude of the mean shifts is critical. Whereas we increase  $\kappa$ , the magnitude of the mean shifts becomes larger. Therefore, we examine three different values of  $\kappa$ : 0.5, 1, and 2. Table 10 summarises the main parameter settings for the following methods:

- Binary Segmentation + CUSUM on TRUE  $Z_t$  (BS-CUSUM on  $Z_t$ ).
- Genetic Algorithm + MDL on TRUE  $Z_t$  (GA-MDL on  $Z_t$ ).

- Genetic Algorithm + mBIC on TRUE  $Z_t$  (GA-mBIC on  $Z_t$ ).
- Binary Segmentation + CUSUM on Reconstructed latent  $\tilde{Z}_t$  (BS-CUSUM  $\tilde{Z}_t$ ).
- Genetic Algorithm + MDL on Categorical  $X_t$  (GA-MDL on  $X_t$ ).
- Genetic Algorithm + mBIC on Categorical  $X_t$  (GA-mBIC on  $X_t$ ).

Parameters	Value
Number of categories ( $K$ )	3
Time series length ( $T_s$ )	500
Auto-correlation ( $\phi$ )	0.5
$\text{Var}(Z_t)$	1
$\sigma^2$	$1 - \phi^2$
signal-to-noise ratio ( $\kappa$ )	$\frac{\Delta}{\sqrt{\sigma^2}}$
Mean shift ( $\Delta$ )	$\kappa\sqrt{\sigma^2}$

Table 10.: Parameters of general setting.

For every simulation setting, we provide two plots. The upper plot displays the time series of the latent variable  $Z$ , while the lower plot focuses on the time series of the categorical variable  $X$ . In the plot corresponding to the latent variable, each segment includes a confidence interval, as elaborated in Section 4.2.5. The blue dashed line represents the upper confidence limit, and the red dashed line represents the lower confidence limit. In the plot for the categorical time series, the red dashed line signifies the sample mean for each segment. In addition, the positions of the changepoints are highlighted by vertical green dashed lines, offering a clear visual cue to identify significant transitions within the dataset.

#### 4.3.1 Three Changepoints Setting (Up Down Up, $\kappa = 2$ )

In this simulation set, we maintain the general specified parameters in Table 10 with the introduction of three changepoints. The changepoints are placed in times

$t = 125, 250,$  and  $375$ . The mean shifts follow a uniform distribution across time intervals. Specifically,  $\mu_t$  starts at  $-0.5$  for the first segment, then increases to  $1.23$  for the second segment in times  $126\text{--}250$ , then decreases to  $-0.5$  during times  $251\text{--}375$ , and rises to  $1.23$  for the last segment. With  $\kappa = 2$ , since  $\mu_t = \alpha_0 + \Delta$ , if we have  $\alpha_0 = -0.5$ , we will have the changepoint parameters, shown in Figure 16:

- $\alpha_0 = -0.5$ , mean of first segment:  $\mu_1 = \alpha_0 = -0.5$ ;
- $\Delta_1 = 1.732$ , mean of second segment:  $\mu_2 = \alpha_0 + \Delta_1 = 1.23$ ;
- $\Delta_2 = 0$ , mean of third segment:  $\mu_3 = \alpha_0 + \Delta_2 = -0.5$ ;
- $\Delta_3 = 1.732$ , mean of fourth segment:  $\mu_4 = \alpha_0 + \Delta_3 = 1.23$

	Effective Percentages of Detected Changepoints								Avg Dist
	Up Down Up, $\kappa = 2$								
	0	1	2	3	4	5	6	$m > 7$	
Effective Changepoints:	0	0	0	100	0	0	0	0	
BS+CUSUM: $Z_t$	0	10.05	4.39	13.08	17.48	15.52	14.06	26.2	2.52
GA+MDL: $Z_t$	0.78	0.19	1.17	96.97	0.87	0	0	0	0.05
GA+mBIC: $Z_t$	8.00	12.69	0.29	78.71	0.29	0	0	0	0.60
BS+CUSUM: $\tilde{Z}_t$	0	16.40	6.34	64.64	10.54	1.66	0.29	0.097	0.61
GA+MDL: $X_t$	0	0	0	99	1	0	0	0	0.03
GA+mBIC: $X_t$	0	0	0	77	20	32.3	0.33	0	0.28

Table 11.: Empirical proportions of the estimated number of changepoints, and the average distance (Up Down Up,  $\kappa = 2$ ). The true value of  $m$  is 3 in  $Z_t$ .

### 4.3.2 Three Changepoints Setting (Up Down Up, $\kappa = 1$ )

Similarly, we repeat the same pattern but with  $\kappa = 1$ . This simulation set keeps the same pattern with a smaller mean shift. The following is the calculation of each segment's mean, shown in Figure 17:



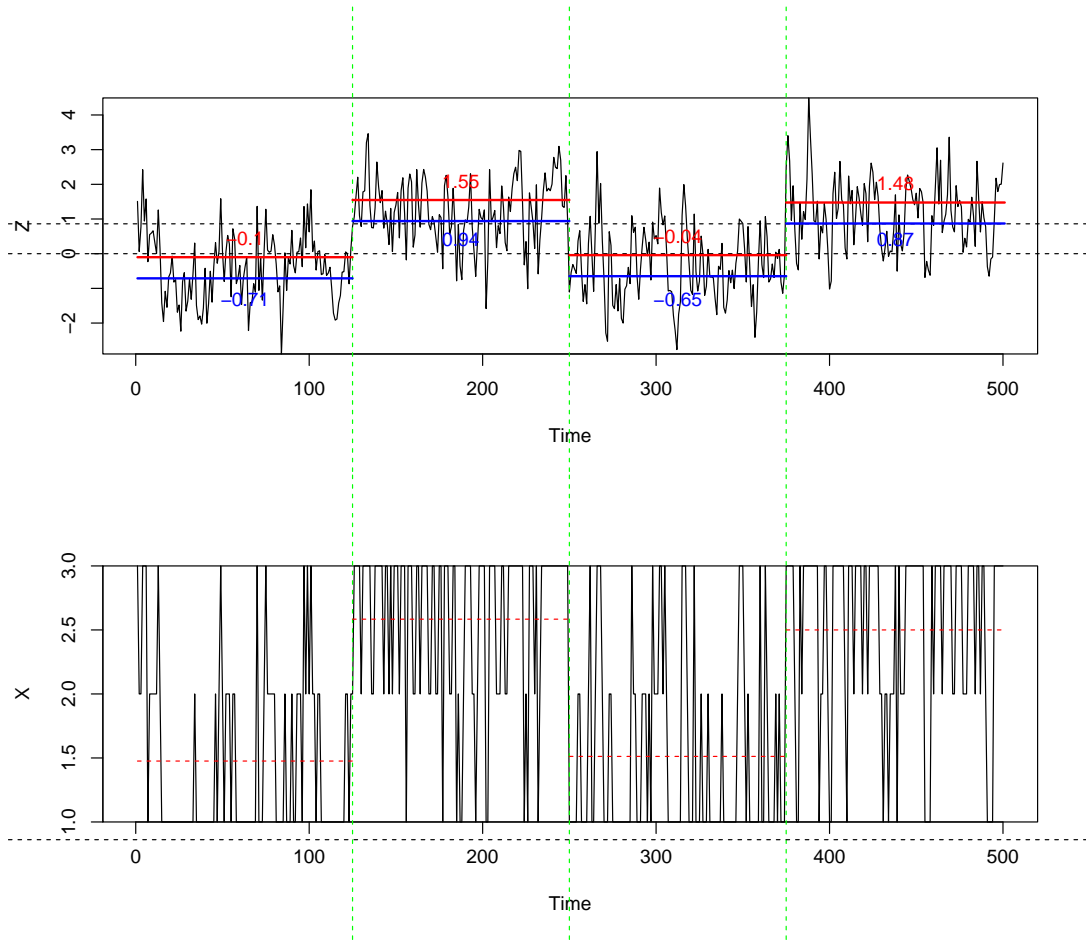


Fig. 16.: Time series plots of latent variable ( $Z$ ) and categorical variable ( $X$ ) for the Up Down Up,  $\kappa = 2$  setting.

- $\alpha_0 = -0.5$ , mean of first segment:  $\mu_1 = \alpha_0 = -0.5$ ;
- $\Delta_1 = 0.866$ , mean of second segment:  $\mu_2 = \alpha_0 + \Delta_1 = 0.366$ ;
- $\Delta_2 = 0$ , mean of third segment:  $\mu_3 = \alpha_0 + \Delta_2 = -0.5$ ;
- $\Delta_3 = 0.866$ , mean of forth segment:  $\mu_3 = \alpha_0 + \Delta_3 = 0.366$

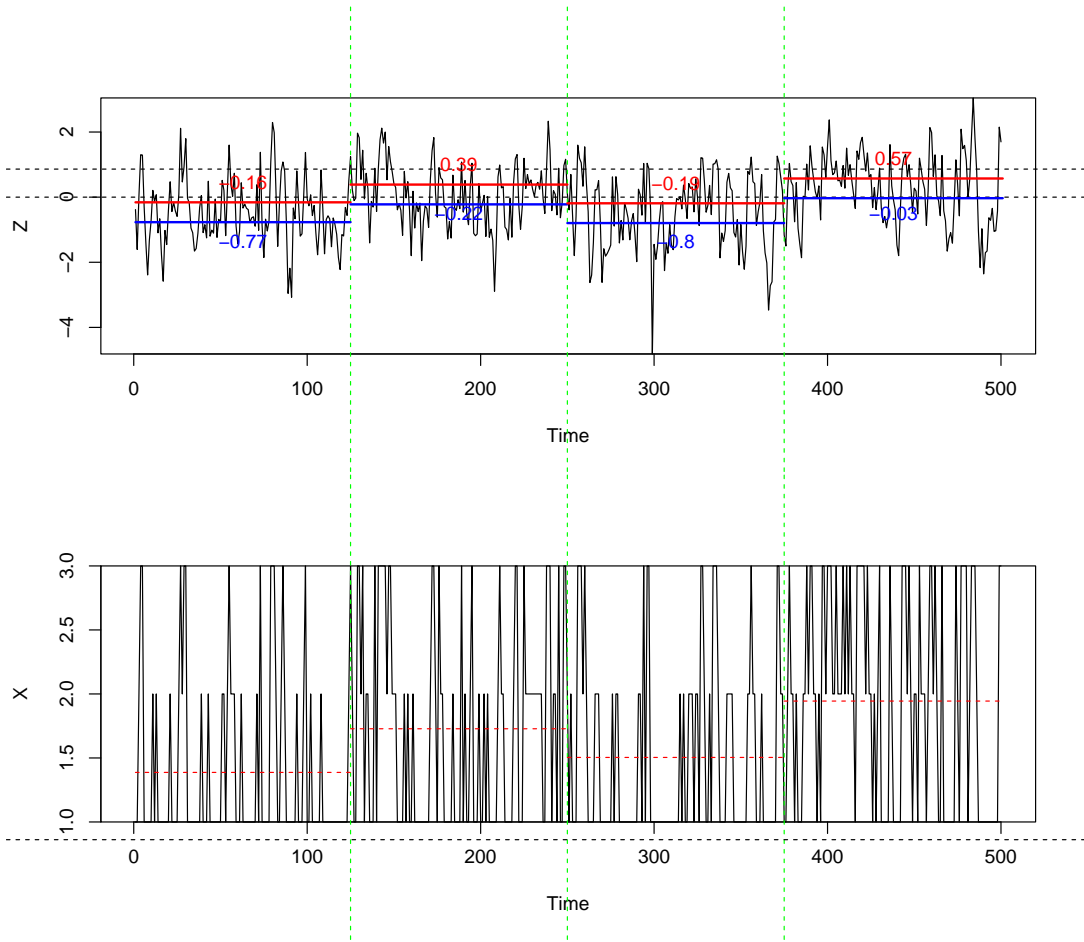


Fig. 17.: Time series plots of latent variable ( $Z$ ) and categorical variable ( $X$ ) for the Up Down Up,  $\kappa = 1$  setting.

### 4.3.3 Three Changepoints Setting (Up Down Up, $\kappa = 0.5$ )

Here, we set  $\kappa = 0.5$ , leading to smaller shifts in the mean. Below, we present the computation of each segment's mean, shown in Figure 18:

- $\alpha_0 = -0.5$ , mean of first segment:  $\mu_1 = \alpha_0 = -0.5$ ;
- $\Delta_1 = 0.433$ , mean of second segment:  $\mu_2 = \alpha_0 + \Delta_1 = -0.067$ ;
- $\Delta_2 = 0$ , mean of third segment:  $\mu_3 = \alpha_0 + \Delta_2 = -0.5$ ;

	Effective Percentages of Detected Changepoints								Avg Dist
	Up Down Up, $\kappa = 1$								
	0	1	2	3	4	5	6	$m > 7$	
Effective Changepoints:	0	0.6	8.2	91.2	0	0	0	0	
BS+CUSUM: $Z_t$	0.8	2.7	1	17.6	18.1	19.1	16.1	25.6	2.44
GA+MDL: $Z_t$	49.8	16.2	24.8	9.2	0	0	0	0	2.07
GA+mBIC: $Z_t$	80.6	18.0	0.5	0.9	0	0	0	0	2.77
BS+CUSUM: $\tilde{Z}_t$	4.68	12.5	3.12	50	20.3	7.81	1.56	0	0.93
GA+MDL: $X_t$	24	11	43	21	0	0	0	0	1.4
GA+mBIC: $X_t$	2.6	13	9.3	63	10.6	0.33	0.33	0	0.65

Table 12.: Empirical proportions of the estimated number of changepoints and the average distance for the setting Up Down Up,  $\kappa = 1$ . The true value of  $m$  is 3 in  $Z_t$ .

	$\leq -4$	-3	-2	-1	0	1	2	3	$\geq 4$	Effective Avg Dist
BS+CUSUM: $\tilde{Z}_t$	0	5.4	11.2	5.2	49.8	18.6	7.2	2.0	0.6	0.80
GA+MDL: $X_t$	0	18	11	42	28	1	0	0	0	1.02
GA+mBIC: $X_t$	0	1	12	10	65	10	1	0.3	0	0.54

Table 13.: Differences from an effective number of changepoints for the setting Up Down Up,  $\kappa = 1$ .

- $\Delta_3 = 0.433$ , mean of forth segment:  $\mu_3 = \alpha_0 + \Delta_3 = -0.067$

Table 11 shows the empirical proportions of estimated changepoints for the Up Down Up configuration with three changepoints; we observed distinct performance characteristics among the GA+MDL, GA+mBIC, and BS+CUSUM methods on the ordinal categorical variable  $X_t$ . GA+MDL demonstrated high precision in detecting the exact number of changepoints ( $m = 3$ ) in 99% of simulations, illustrating its effectiveness in detecting significant changes within the data with minimal overestimation, as indicated by an average distance of 0.03. This suggests that GA+MDL can accurately capture the underlying structure of categorical time series data. Conversely, the GA+mBIC on  $X_t$  method, while also showing a propensity for correct

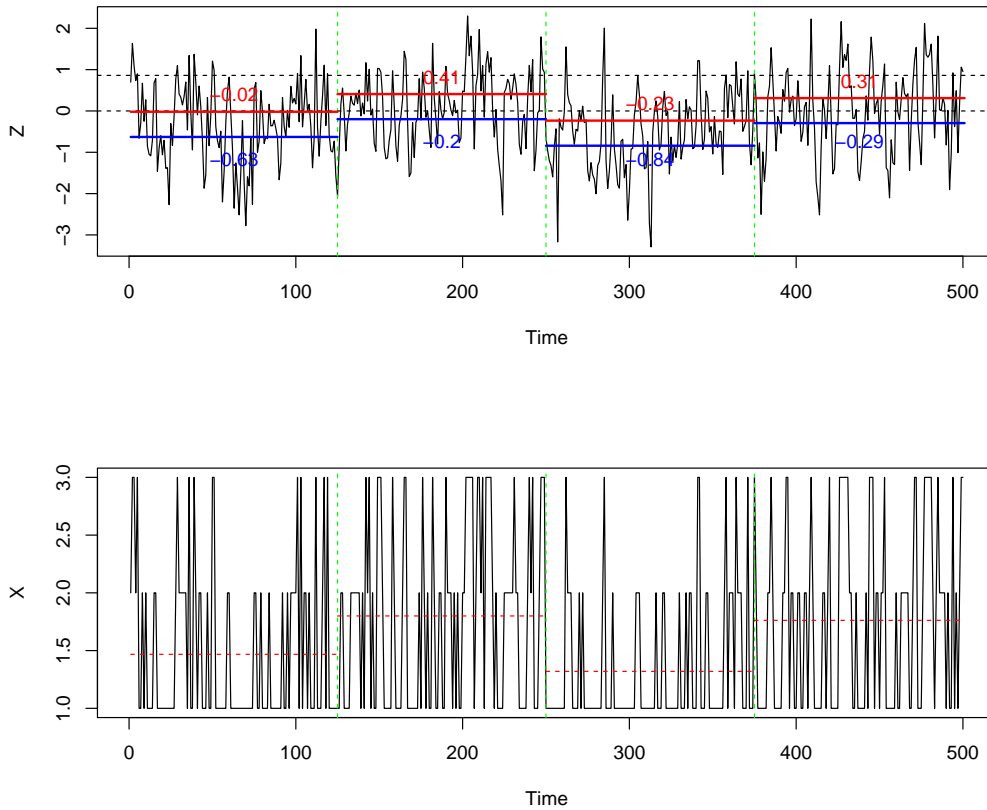


Fig. 18.: Time series plots of latent variable ( $Z$ ) and categorical variable ( $X$ ) for the Up Down Up,  $\kappa = 0.5$  setting.

change point detection in 77% of cases, exhibited a higher overestimation rate with an average distance of 0.28. This tendency to detect additional change points beyond the true count indicates a slight bias towards over-sensitivity in identifying changes. The BS+CUSUM on  $X_t$  method presented a more varied distribution of detected change points, correctly identifying the three change points in 64% of simulations but also showing instances of both underestimation and overestimation. This variability, along with an average distance of 0.61, indicates a less consistent performance in accurately detecting the exact number of change points than the GA-based methods.

	Effective Percentages of Detected Changepoints Up Down Up, $\kappa = 1$								Avg Dist
	0	1	2	3	4	5	6	$m > 7$	
Effective Changepoints:	34	24	31	11	0	0	0	0	
BS+CUSUM: $Z_t$	15.8	11.4	10.0	18.4	15.6	9.0	7.6	12.2	2.09
GA+MDL: $Z_t$	3	12	41	35	8	1	0	0	1.16
GA+mBIC: $Z_t$	1	18	40	34	6	0	0	0	1.16
BS+CUSUM: $\tilde{Z}_t$	55.0	25.8	8.8	7.8	2.0	0.6	0	0	2.38
GA+MDL: $X_t$	85	10	2	0	0	0	0	0	2.82
GA+mBIC: $X_t$	51	27	17	3	2	3	0	0	2.31

Table 14.: Empirical proportions of the estimated number of changepoints and average distance for the setting Up Down Up,  $\kappa = 0.5$ . The true value of  $m$  is 3 in  $Z_t$ .

	$\leq -4$	-3	-2	-1	0	1	2	3	$\geq 4$	Effective Avg Dist
BS+CUSUM: $\tilde{Z}_t$	0	5.67	17.3	13.3	46	9.67	5	2.34	0.6	1.02
GA+MDL: $X_t$	0	9	26	25	36	3	10	0	0	1.04
GA+mBIC: $X_t$	0	4	14	22	47	7	5	0	1	0.96

Table 15.: Differences from the effective number of changepoints for the setting Up Down Up,  $\kappa = 0.5$ .

In the second simulation setting, three changepoints (Up, Down, Up) with  $\kappa = 1$ , smaller mean shifts introduce a slightly more challenging detection environment compared to the first scenario with  $\kappa = 2$ . As indicated in the Table 12, not all simulations maintain three effective changepoints due to subtler shifts; 91.2% are identified with three effective changepoints, while 8.2% effectively have two, highlighting instances where the third changepoint does not significantly alter the data structure to be deemed effective. The comparison of GA+MDL, GA+mBIC, and BS+CUSUM methods when  $\kappa = 1$  reveals varying degrees of sensitivity and accuracy. Table 13 shows a reasonable performance for BS+CUSUM, exhibiting the highest concentration of zero difference (49.8%) from the effective number. This suggests a balanced detection ca-

pability, albeit with instances of both under and overestimation. However, GA+MDL, with 28% of simulations perfectly matching the effective number of changepoints, exhibits a tendency towards underdetection, as evidenced by a significant proportion of simulations indicating fewer detections than effective changepoints. Conversely, GA+mBIC shows an improved alignment with the effective changepoint model, with 65% of simulations achieving zero difference. This indicates a more accurate detection capability that aligns well with the nuanced reality of effective changepoints, though it still encounters challenges with overestimation. Moving to the scenario with  $\kappa = 0.5$ , we observe reduced mean shifts, thereby escalating the detection challenge. As illustrated in Table 14, the distribution of the effective number of changepoints shows less concentration around three changepoints in comparison to earlier settings: 100% for  $\kappa = 2$ , 91% for  $\kappa = 1$ , and decreasing to 11% for  $\kappa = 0.5$ . This reduction leads to most methods underestimating detecting the true number of changepoints, and some methods, like GA, detected zero changepoints. Table 15 indicates that the proposed methods have a lower rate of exact matches between the detected changepoints and the true effective changepoints. This highlights the challenges in accurately detecting the true number of changepoints. Additionally, the increase in average distances in this scenario, compared to previous settings, points out the challenge of correctly detecting both the number and locations of the changepoints. However, the GA+mBIC method accurately matches the true number of changepoints in approximately half of the cases (47%) with a lower average distance than the remaining methods. Overall, the decrease in mean shift magnitude at  $\kappa = 0.5$  accentuates the necessity for careful consideration in selecting changepoint detection methodologies. While GA+MDL and GA+mBIC present promising solutions for accurately detecting significant shifts in data with subtler variances with the mean shifts.

#### 4.3.4 Three Changepoints Cetting (Up Up Up, $\kappa = 2$ )

This simulation differs from the previous setting by presenting a scenario where mean shifts consistently increase, each with equal magnitude. Changepoints are also placed at  $t = 125, 250,$  and  $375$ , indicating transitions across segments with uniform time distributions. Similarly, the mean of the first segment starts at  $-0.5$  for the first segment, and then the second segment will increase with  $\mu = 1.23$ . Further, the third segment increases to  $2.964$ , and finally, the last segment increases to  $4.696$ , shown in Figure 19:

- $\alpha_0 = -0.5$ , mean of first segment:  $\mu_1 = \alpha_0 = -0.5$ ;
- $\Delta_1 = 1.732$ , mean of second segment:  $\mu_2 = \alpha_0 + \Delta_1 = 1.23$ ;
- $\Delta_2 = 3.464$ , mean of third segment:  $\mu_3 = \alpha_0 + \Delta_2 = 2.964$ ;
- $\Delta_3 = 5.196$ , mean of forth segment:  $\mu_3 = \alpha_0 + \Delta_3 = 4.696$

	Effective Percentages of Detected Changepoints								Avg Dist
	Up Down Up, $\kappa = 1$								
	0	1	2	3	4	5	6	$m > 7$	
Effective Changepoints:	0	42.4	57.6	0	0	0	0	0	
BS+CUSUM: $Z_t$	0	0	5	17	14	19	17	28	2.51
GA+MDL: $Z_t$	0	0	0.2	99.2	0.6	0	0	0	0.027
GA+mBIC: $Z_t$	0	0	0.1	99.7	0.2	0	0	0	0.022
BS+CUSUM: $\tilde{Z}_t$	0	29	57	13	7	1	0	0	1.27
GA+MDL: $X_t$	0	0	99	1	0	0	0	0	1.01
GA+mBIC: $X_t$	0	0	88	11	1		0	0	0.97

Table 16.: Empirical proportions of the estimated number of changepoints and the average distance in the setting Up Up Up,  $\kappa = 2$ . The true value of  $m$  is 3 in  $Z_t$ .

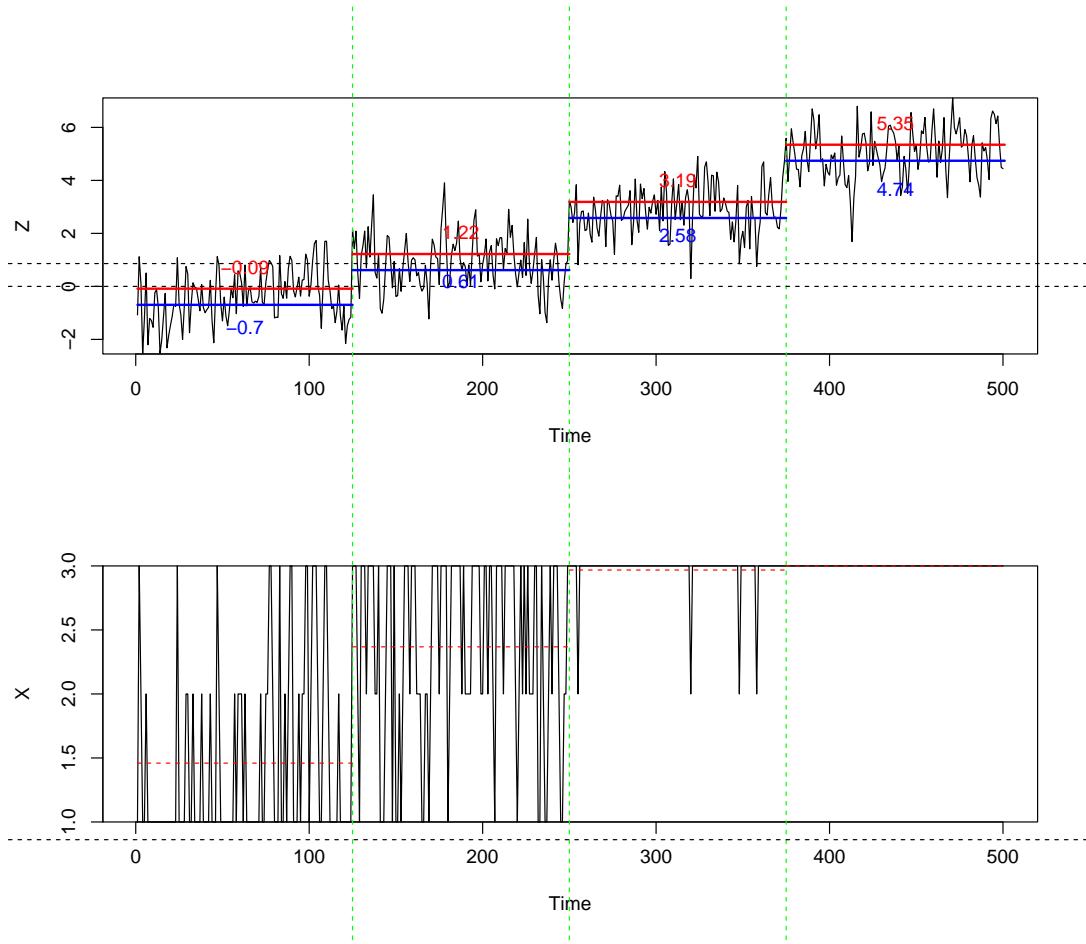


Fig. 19.: Time series plots of latent variable ( $Z$ ) and categorical variable ( $X$ ) for the Up Up Up,  $\kappa = 2$  setting.

	$\leq -4$	-3	-2	-1	0	1	2	3	$\geq 4$	Effective Avg Dist
BS+CUSUM: $\tilde{Z}_t$	0	0	0	8.6	59.8	27.8	3.4	0.2	0.2	0.46
GA+MDL: $X_t$	0	0	0	0	61	39	0	0	0	0.32
GA+mBIC: $X_t$	0	1	0	0	50	45	5	0	0	0.56

Table 17.: Differences from the effective number of changepoints for the setting Up Up Up,  $\kappa = 2$ .



### 4.3.5 Three Changepoints Setting (Up Up Up, $\kappa = 1$ )

This simulation set keeps the same increasing pattern with a smaller mean shift. The mean shifts should be smaller after we reduce  $\kappa$  since  $\Delta = \kappa\sqrt{1 - \phi^2}$ . The changepoints are placed in times  $t= 125, 250,$  and  $375$ . The mean shifts follow a uniform distribution across time intervals. Starting with  $\mu_t$  at  $-0.5$  for time steps 1–125, then escalates to  $0.366$  for times 126–250. Further, it increases to  $1.232$  during times 251–375 and undergoes another increment to  $2.098$  for times 376–500. This configuration illustrates mean shifts in a consistent direction (Up Up Up), with all shifts characterized by identical magnitudes, shown in Figure 20:

- $\alpha_0 = -0.5$ , mean of first segment:  $\mu_1 = \alpha_0 = -0.5$ ;
- $\Delta_1 = 0.866$ , mean of second segment:  $\mu_2 = \alpha_0 + \Delta_1 = 0.366$ ;
- $\Delta_2 = 1.732$ , mean of third segment:  $\mu_3 = \alpha_0 + \Delta_2 = 1.232$ ;
- $\Delta_3 = 2.598$ , mean of forth segment:  $\mu_3 = \alpha_0 + \Delta_3 = 2.098$

	Effective Percentages of Detected Changepoints								Avg Dist
	Up Down Up, $\kappa = 1$								
	0	1	2	3	4	5	6	$m > 7$	
Effective Changepoints:	0	2.6	66.8	30.6	0	0	0	0	
BS+CUSUM: $Z_t$	1	4	0	23	19	21	12	20	2.29
GA+MDL: $Z_t$	0	0.1	92.9	6.9	0.1	0	0	0	0.99
GA+mBIC: $Z_t$	0	29.0	65.9	5.1	0	0	0	0	1.28
BS+CUSUM: $\tilde{Z}_t$	0	2	45	44	8	1	0	0	0.68
GA+MDL: $X_t$	0	0	98.3	1.6	0	0	0	0	2.03
GA+mBIC: $X_t$	0	1.6	47.3	42.3	7	2	0	0	0.28

Table 18.: Empirical proportions of the estimated number of changepoints and average distance for the setting Up Up Up,  $\kappa = 1$ . The true value of  $m$  is 3 in  $Z_t$ .

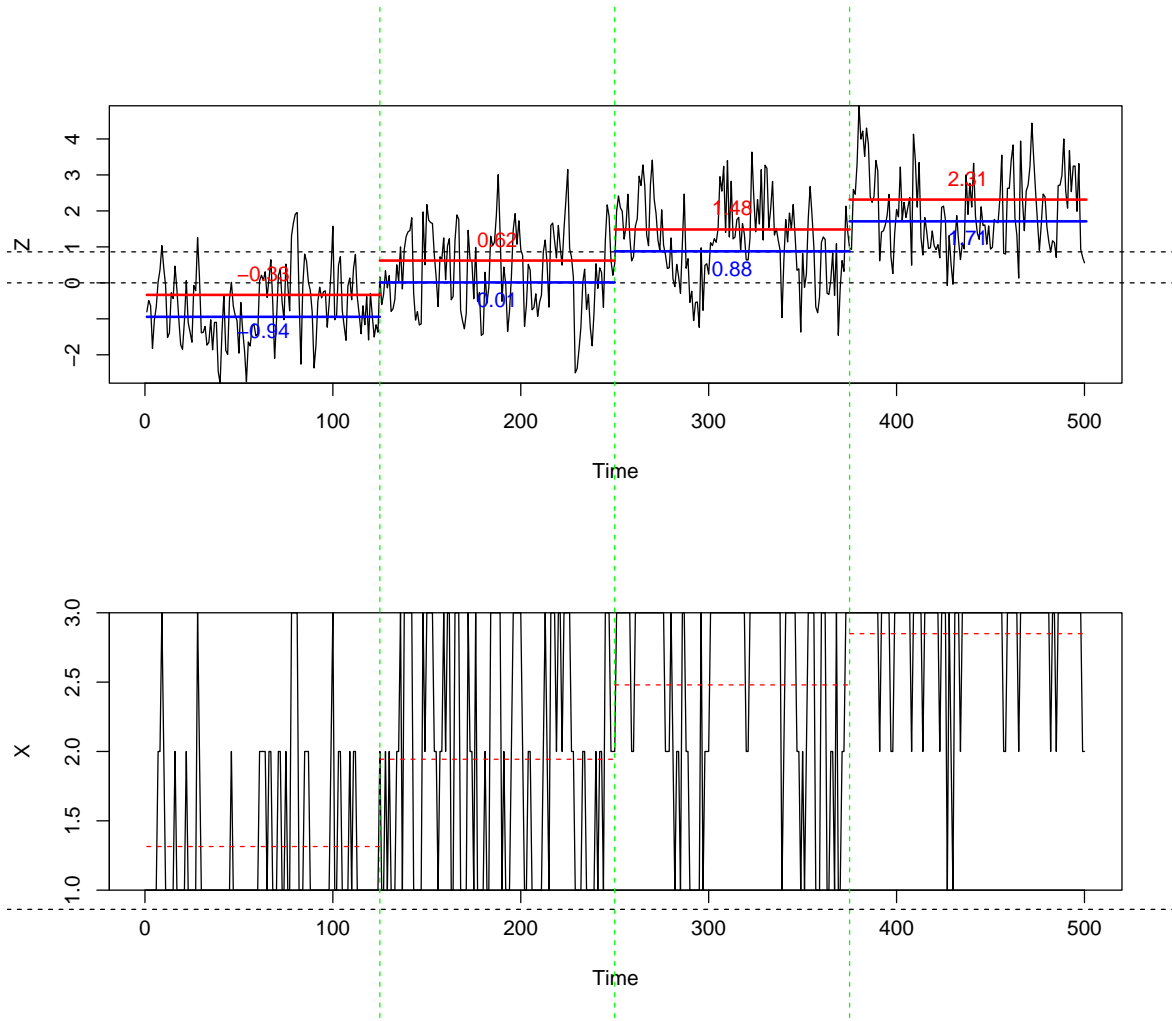


Fig. 20.: Time series plots of latent variable ( $Z$ ) and categorical variable ( $X$ ) for Up Up,  $\kappa = 1$  setting.

#### 4.3.6 Three Changepoints Setting (Up Up Up, $\kappa = 0.5$ )

In this simulation set, we use the same pattern as the previous simulation but with  $\kappa = 0.5$ . This reduction in the mean shift should make the detection more challenging than the previous setting. Initially, the mean  $\mu_t$  is set at  $-0.5$  for the first 125 time steps, then increases slightly to  $-0.067$  for the interval 126–250. It elevates

	$\leq -4$	-3	-2	-1	0	1	2	3	$\geq 4$	Effective Avg Dist
BS+CUSUM: $\tilde{Z}_t$	0	0	0.2	13.4	41.2	34.8	9.8	0.6	0	0.61
GA+MDL: $X_t$	0	0	0	35	60.6	4.3	0	0	0	0.43
GA+mBIC: $X_t$	0	1	0	15	50	28	6	0	0	0.52

Table 19.: Differences from the effective number of changepoints. For the setting Up Up Up,  $\kappa = 1$ .

again to 0.366 in the span of 251–375, and experiences a further rise to 0.799 for the final segment 376–500. This configuration delineates a sequence of mean shifts moving in an upward direction, maintaining uniform magnitude throughout. With  $\kappa = 0.5$  and an initial setting of  $\alpha_0 = -0.5$ , the specified changepoint parameters highlight the increase in mean values across the series, shown in Figure 21).

- $\alpha_0 = -0.5$ , mean of first segment:  $\mu_1 = \alpha_0 = -0.5$ ;
- $\Delta_1 = 0.433$ , mean of second segment:  $\mu_2 = \alpha_0 + \Delta_1 = -0.067$ ;
- $\Delta_2 = 0.866$ , mean of third segment:  $\mu_3 = \alpha_0 + \Delta_2 = 0.366$ ;
- $\Delta_3 = 0.433$ , mean of fourth segment:  $\mu_4 = \alpha_0 + \Delta_3 = 0.799$

Moving to the three changepoints setting (Up Up Up) under varying mean shift magnitudes, denoted by  $\kappa$ . Our findings reveal a nuanced landscape of detection challenges and methodological effectiveness as influenced by the magnitude of  $\kappa$ . This divergence primarily stems from the subtler mean shifts, especially for the last changepoint, which tends to merge with the categories of preceding segments. Observations from the analysis of the true effective number of changepoints (Tables 16, 18 and 20) reveal a notable pattern: the proportion of the third (last) changepoint decreases as the mean shift is increased by increasing  $\kappa$ . Where the proportions of the last changepoint as an effective changepoint are as the following: 46% when  $\kappa = 0.5$ ,

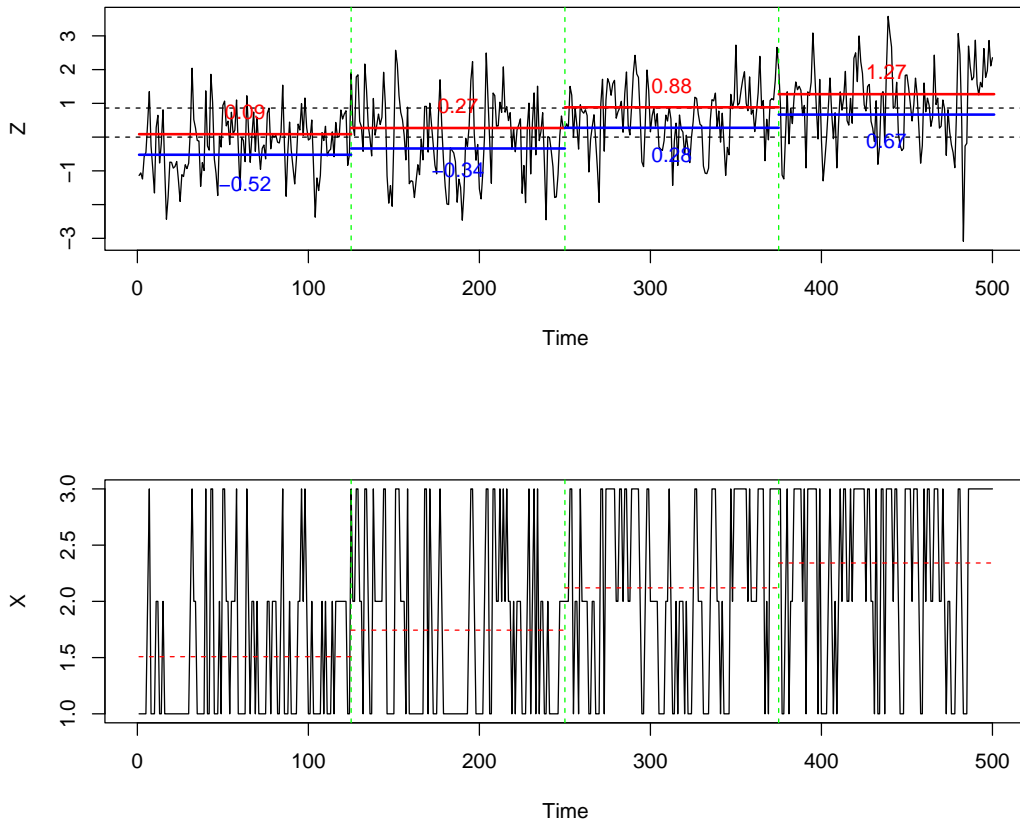


Fig. 21.: Time series plots of latent variable ( $Z$ ) and categorical variable ( $X$ ) for Up Up,  $\kappa = 0.5$  setting.

30% when  $\kappa = 1$ , and 0% when  $\kappa = 2$ . This trend underscores the nuanced impact that increasing mean shifts have on the detectability of subsequent changepoints within the series. This scenario highlights the necessity of the “effective number of changepoints” concept, illustrating that the assumption of three changepoints may not always align with the practical detectability within the data. In assessing the performance of the detection methods, it’s evident that the traditional approach of assuming a fixed number of changepoints across all simulations does not provide a fair comparison. The discrepancies between the expected and effective number of

	Effective Percentages of Detected Changepoints								Avg Dist
	Up Down Up, $\kappa = 1$								
	0	1	2	3	4	5	6	$m > 7$	
Effective Changepoints:	0	5.2	48.6	46.8	0	0	0	0	
BS+CUSUM: $Z_t$	0	7.0	15.8	19.8	14.8	16.6	10.0	26	2.10
GA+MDL: $Z_t$	1.6	22.6	75.8	0	0	0	0	0	1.37
GA+mBIC: $Z_t$	3.8	91.4	4.8	0	0	0	0	0	2.03
BS+CUSUM: $\tilde{Z}_t$	0	32.2	35.8	20.8	7.8	1.6	1.4	0.6	1.29
GA+MDL: $X_t$	0	25	75	0	0	0	0	0	1.36
GA+mBIC: $X_t$	0	49.5	39.5	10	1.0	3	0	0	1.5

Table 20.: Empirical proportions of the estimated number of changepoints and average distance for the setting Up Up Up,  $\kappa = 0.5$ . The true value of  $m$  is 3 in  $Z_t$ .

	$\leq -4$	-3	-2	-1	0	1	2	3	$\geq 4$	Effective Avg Dist
BS+CUSUM: $\tilde{Z}_t$	0	0	14	33.6	28.8	15.8	4.2	2.2	1.4	1.05
GA+MDL: $X_t$	0	0	12	42	43	3	0	0	0	0.82
GA+mBIC: $X_t$	0	0	20	44	28.5	6.5	1	0	0	1.03

Table 21.: Differences from the effective number of changepoints for the setting Up Up Up,  $\kappa = 0.5$ .

changepoints necessitate reevaluating how detection accuracy is measured.

Starting with the setting at  $\kappa = 2$ , this configuration offers an insightful starting point. With the largest mean shifts among our test cases, it presents a seemingly advantageous scenario for changepoint detection. Table 16 illustrates that the proposed methods predominantly identified two changepoints, aligning closely with the proportions observed for the effective number of changepoints. The empirical evidence strongly supports the efficacy of the proposed methods, showcasing a remarkable adaptability in accurately identifying the effective number of changepoints. Table 17 demonstrates that GA+MDL matches the true number of effective changepoints with a 61% accuracy rate. Additionally, GA+MDL records the smallest effective average

distance at 0.32, indicating its effectiveness in accurately detecting the number and locations of changepoints.

Transitioning to  $\kappa = 1$ , this intermediate scenario maintains the positive detection outcomes observed with  $\kappa = 2$ , albeit with slightly reduced mean shift magnitudes. Despite this adjustment, the GA+MDL and GA+mBIC methods outperformed the BS+CUSUM baseline method with proportions of 60% and 50% for GA+MDL and GA+mBIC respectively as shown in Table 19. This consistency highlights their robustness across varying levels of data variability, affirming their role as promising approaches for changepoint analysis in ordinal time series. Introducing the final setting at  $\kappa = 0.5$  escalates the detection challenge considerably due to the reduced mean shift magnitude. This scenario unveils a divided landscape of effective changepoints, with empirical proportions balancing between two and three changepoints. Such a distribution indicates the heightened complexity in discerning the subtler shifts within the data, particularly affecting the detectability of the final changepoint. The empirical proportions presented in Table 20, revealing an underestimation of the number of changepoints, clearly highlight the challenges detection methods encounter in this intricate setting. This underscores the importance of methodological flexibility in response to smaller mean shifts. However, GA+MDL outperformed all other methods with a 43% accuracy rate in matching the true effective number of changepoints and attained the smallest effective average distance of 0.82.

The progression from  $\kappa = 2$  to  $\kappa = 0.5$  delineates a clear trajectory of increasing detection difficulty, inversely related to the mean shift magnitude. Through this exploration, the indispensable role of considering the effective number of changepoints becomes evident, challenging traditional notions of changepoint detection and urging a nuanced understanding of method performance across different dynamical

landscapes within time series data. Table 21 depicts that the GA-based method outperformed the BS+CUSUM method with lower average distances. The GA+MDL method, in particular, emerges as a notably adaptable tool capable of navigating the complexities introduced by varying  $\kappa$  values with commendable success.

#### 4.3.7 Comparing Different Values of the Auto-correlation Parameter

In exploring changepoint detection within time series data, particularly under the Up Down Up configuration with a signal-to-noise ratio of  $\kappa = 2$ , we extend our investigation to examine the influence of varying autocorrelation levels. Autocorrelation, denoted by  $\phi$ , represents the correlation between values in a time series separated by a specific lag. It is a critical factor in the temporal structure of time series data, affecting the detection and interpretation of changepoints. To this end, we systematically assess the impact of different  $\phi$  values on detecting changepoints. The chosen values,  $\phi = \{-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75\}$ , span both negative and positive correlations, offering a comprehensive view of how autocorrelation influences the detection process. This range allows us to observe the effects of both inverse and direct correlations between successive observations in the series, providing insights into how these dynamics affect the detection of true changepoints.

This simulation setting aims to illuminate the role of autocorrelation parameter  $\phi$  in changepoint detection, enhancing our understanding of its implications for accurately identifying shifts in complex time series data. By varying  $\phi$ , we delve into how temporal dependence within the data influences the effectiveness of changepoint detection methodologies, thereby refining our approach to analyzing time series with diverse autocorrelation characteristics.

Table 22 and Figure 22 depict the effects of varying autocorrelation levels on changepoint detection within time series data, revealing insights into how these

	$\phi$						
	0.75	-0.5	0.25	0	0.25	0.5	0.75
BS+CUSUM: $Z_t$	0.72	0.79	0.73	0.70	1.08	2.6	6.46
GA+MDL: $Z_t$	0.006	0.007	0.01	0.012	0.012	0.05	2.23
GA+mBIC: $Z_t$	0.001	0.002	0.01	0.013	0.023	0.33	2.74
BS+CUSUM: $\tilde{Z}_t$	0.63	0.73	0.57	0.65	0.51	0.72	1.86
GA+MDL: $X_t$	0.005	0.005	0.006	0.007	0.01	0.03	0.37
GA+mBIC: $X_t$	0.005	0.005	0.01	0.01	0.03	0.28	1.88

Table 22.: Average Distances for varies values of  $\phi$ , for the setting Up Down Up,  $\kappa = 2$ .

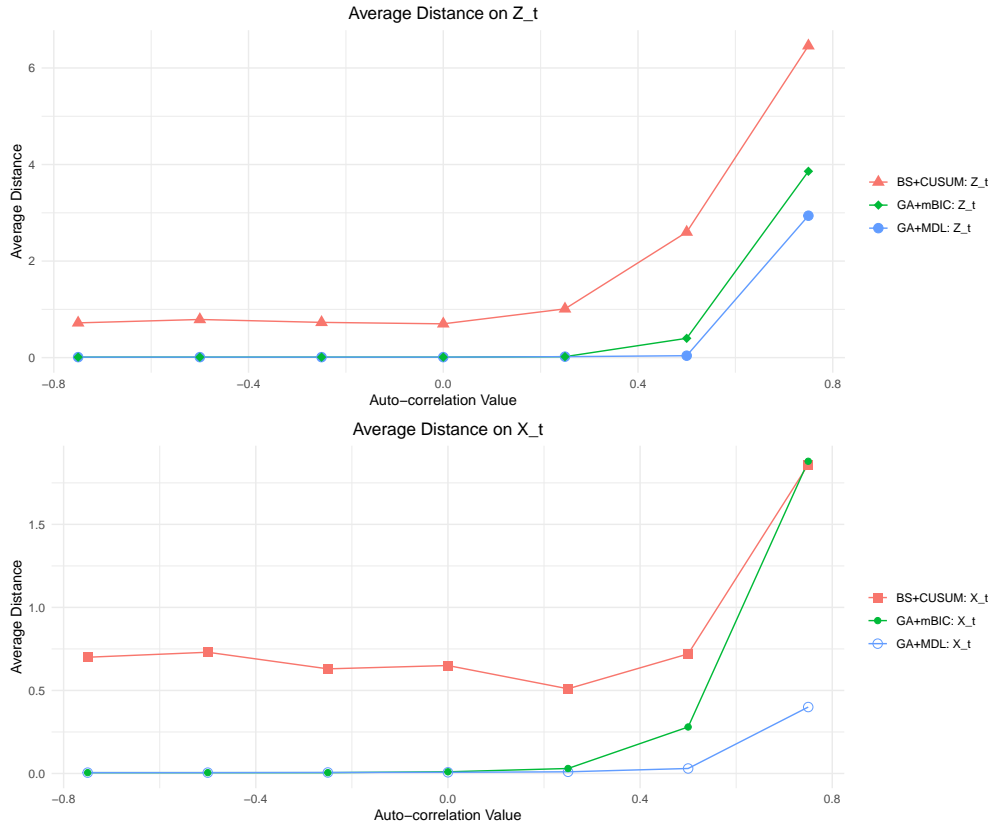


Fig. 22.: Average distance with varies values of  $\phi$  for the setting Up Down Up,  $\kappa = 2$ . The true number of changepoints  $m = 3$  in  $Z_t$

methodologies adapt to temporal dependencies. As we systematically adjust  $\phi$  across a spectrum from negative to positive, we observe a consistent pattern: an increase in



$\phi$  correlates with heightened challenges in changepoint detection. This phenomenon is quantitatively captured through the average distance metric, which escalates as  $\phi$  progresses towards stronger positive autocorrelation. This pattern confirms that higher autocorrelation complicates the discernment of true changepoints due to the enhanced similarity between sequential observations.

Among the proposed methods, GA+MDL and GA+mBIC, when applied to both latent and categorical processes, demonstrate remarkable robustness against the variations in  $\phi$ . They maintain relatively low average distances across the range, showcasing an adeptness at navigating the complexities introduced by autocorrelation. Notably, GA+MDL on  $X_t$  stands out for its consistently low average distances across nearly all values of  $\phi$  and outperforming all the remaining methods. GA+MDL strikes a good balance in changepoint analysis, effectively detecting the number and exact locations of changepoints despite the challenges posed by higher autocorrelation. In contrast, conventional approaches like BS+CUSUM reveal a pronounced susceptibility to higher levels of autocorrelation, evidenced by a significant uptick in average distances as  $\phi$  increases.

#### 4.4 Los Angeles City AQI Data

The time series plot in Figure 23 (top panel) illustrates daily AQI measurements in Los Angeles, USA. The dataset extends from January 1, 2020, to June 18, 2022, encapsulating a period critical for understanding air pollution trends in a major urban center. Monitoring air quality is vital for public health, environmental policy-making, and urban planning, particularly in densely populated areas such as Los Angeles. The city's complex topography and meteorological conditions contribute to variable air quality levels, making it an important case study for environmental and health research.

The categorized AQI time series is plotted in the bottom panel of Figure 23. The data groups daily AQI readings into three distinct categories based on the AQI value: Category 1 signifies 'Good' air quality with 258 observed instances, Category 2 denotes 'Moderate' air quality with 596 instances, and Category 3 represents 'Unhealthy' air quality with 46 instances. These categories are defined by cutoff points on the continuous AQI scale, where the cutoff for 'Good' category is 0-50, the cutoff for 'Moderate' category is 51-100, and the cutoff for 'Unhealthy' category is greater than 100. The dataset, comprising 900 daily observations, is a pivotal resource for studying the impact of urbanization, climate change, and regulatory measures on air quality.

In fitting our models with the GA, we set a population size of 30 individuals, structured across 5 islands. Crossover and mutation were applied with probabilities of 0.95 and 0.1, respectively, while the probability assigned for changepoints in each time series was set at 0.06. The fitted GA+MDL model detected two changepoints at 116 April 2020 and 847 April 2022. While GA+mBIC model detected four changepoints at 108 Apr 2020, 508 Jun 2021, 586 Aug 2021, and 839 Apr 2022. These periods may reflect changes in environmental regulations, seasonal variations in air quality, or other socio-economic activities affecting the urban atmosphere. For example, both GA-based models detected a changepoint at the time (April 2020). This detection aligns with the onset of global changes in human activity due to the COVID-19 pandemic, which caused an environmental impact of pandemic-related lockdowns, manifesting in significant air quality improvements. On the other hand, the changepoint detected in April 2022 by the GA+mBIC model corresponds with the publication of the American Lung Association's "State of the Air" report in 2022, which highlighted Los Angeles and several other California cities as having the country's most polluted air during the period of April 2022. While the GA+mBIC model

exhibited tendencies of over-detection in certain simulation settings, the GA+MDL model accurately detected the most plausible changepoints within this time series data, demonstrating its effectiveness in detecting significant shifts.

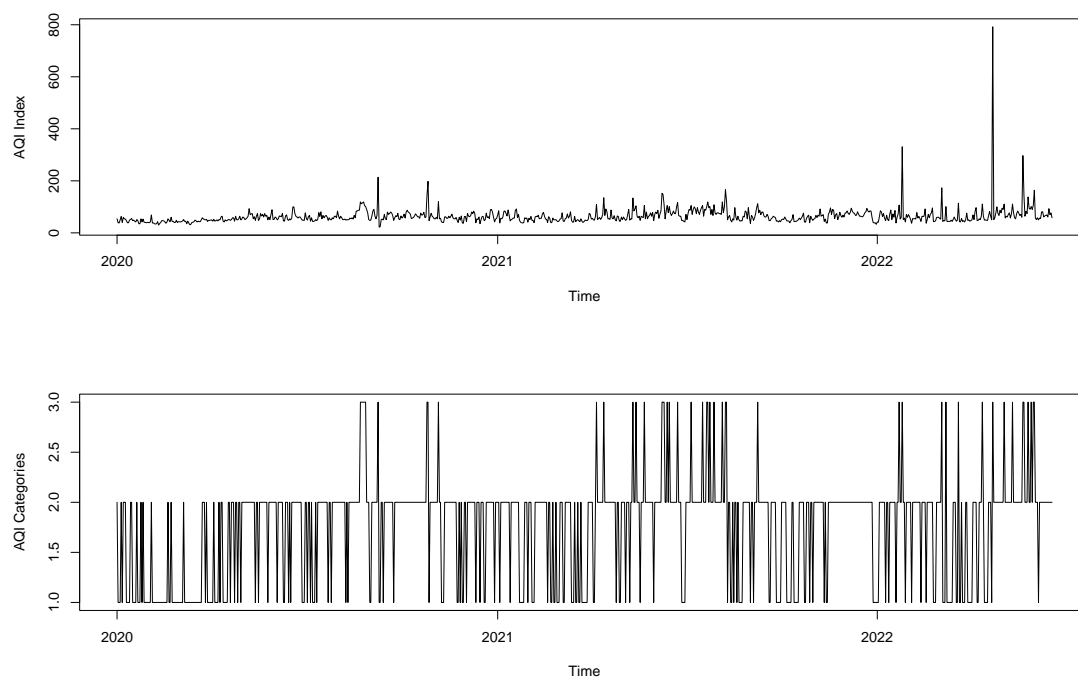


Fig. 23.: Continuous ( $Z$ ) (Top) and the categorical ( $X$ ) daily AQI time series in Los Angeles from 2020 to 2022.

## 4.5 Conclusions

This project proposes GA-based methods, namely GA+MDL and GA+mBIC, for changepoint detection in ordinal categorical time series data, utilizing the AOP model to effectively handle the data’s inherent ordinal categorical characteristics. The study refines changepoint analysis by introducing the concept of the “effective number of changepoints” offering a more nuanced assessment of changepoints’ significance. This adjustment addresses the limitations of evaluating detection methods based solely on the number of changepoints, ensuring a fairer comparison across diverse configurations. Applied to simulated studies with different configurations and parameters, the proposed methods showcased robustness and superior performance compared to traditional changepoint detection techniques. This approach is applied to AQI data from Los Angeles to detect changepoints, indicating significant shifts in air quality over time.

## CHAPTER 5

### SUMMARY AND FUTURE WORK

The main objective of this dissertation is to leverage ML and time series analysis methods to address challenges across different domains. Each project contributes to the overarching theme of enhancing analytical methodologies and their application in real-world scenarios.

The first project aims to develop an RA CUSUM control chart that integrates ML models for more precise probability estimation. This method expands on traditional techniques by applying advanced ML to improve the RA CUSUM chart's capability to monitor small changes in the process. The project uses data from hospital readmissions to accurately assess patients' risk based on their preoperative characteristics and to observe changes in readmission rates within healthcare applications. Using tree-based ML algorithms, the project shows how these analytical tools can enhance the creation and implementation of RA CUSUM charts in healthcare settings, presenting a new solution for process monitoring. Furthermore, it introduces a strategy for comparing predictive models with different levels of complexity, offering a balanced approach to model comparison.

An intriguing direction for future work is suggested by the present project's limitation to a binary classification of the response variable. The methodology employed thus far has only considered two classes for the response variable in constructing the control chart. However, extending this work to include multiple classes for the response variable would considerably expand the RA CUSUM charts' applicability and analytical depth. Adopting a multi-class approach would enable the control chart

to encompass a broader spectrum of information, rendering it a more intricate and thorough monitoring instrument.

In the second project, the focus shifts to the field of chemical analysis, addressing the challenges of high dimensionality, multicollinearity, and non-linearity in spectroscopic data. It introduces the DLR and CLR methods, integrated with PLS and PCR, to adeptly manage the complexities of predicting chemical concentrations. The DLR and CLR showcase outperformed standard PLS and PCR models in accurately predicting concentrations across various chemical compounds, marking a significant advance in chemometric methodologies.

The third project in the dissertation delves into detecting multiple changepoints in autocorrelated ordinal time series data. It introduces models based on the GA (GA+MDL and GA+mBIC), demonstrating superior performance compared to the traditional BS+CUSUM technique. Utilizing the “effective changepoints” concept, these models enhance the precision of detecting significant shifts in data. Their efficacy is corroborated through a range of simulation studies. The application of these models is further illustrated by detecting multiple changepoints in the AQI data from Los Angeles, showcasing their practical utility in environmental data analysis.

A notable limitation arises with the handling of larger time series. The computational time required by the GA-based methods may become burdensome as the length of the time series increases. In addition, future work could delve more profoundly into the concept of “effective number of changepoints”, focusing on its definition and utility in scenarios with diverse magnitudes of mean shifts. Such an investigation would not only refine the methodologies for detecting changepoints but also enhance the interpretability and applicability of the results, offering a more nuanced understanding of the dynamics within the data. This future direction promises to augment the precision and relevance of changepoint analysis in complex time series data.

## REFERENCES

- [1] C. Nicolay, S. Purkayastha, A. Greenhalgh, *et al.*, “Systematic review of the application of quality improvement methodologies from the manufacturing industry to surgical healthcare,” *Journal of British Surgery*, vol. 99, no. 3, pp. 324–335, 2012.
- [2] J. C. Benneyan, “The design, selection, and performance of statistical control charts for healthcare process improvement,” *International Journal of Six Sigma and Competitive Advantage*, vol. 4, no. 3, pp. 209–239, 2008.
- [3] S. H. Steiner, R. J. Cook, V. T. Farewell, and T. Treasure, “Monitoring surgical performance using risk-adjusted cumulative sum charts,” *Biostatistics*, vol. 1, no. 4, pp. 441–452, 2000.
- [4] G. Rossi, S. D. Sarto, and M. Marchi, “A new risk-adjusted bernoulli cumulative sum chart for monitoring binary health data,” *Statistical Methods in Medical Research*, vol. 25, no. 6, pp. 2704–2713, 2016.
- [5] M. A. Farag, M. Sheashea, C. Zhao, and A. A. Maamoun, “Uv fingerprinting approaches for quality control analyses of food and functional food coupled to chemometrics: A comprehensive analysis of novel trends and applications,” *Foods*, vol. 11, no. 18, p. 2867, 2022.
- [6] S. Hossain, C. W. Chow, G. A. Hewa, D. Cook, and M. Harris, “Spectrophotometric online detection of drinking water disinfectant: A machine learning approach,” *Sensors*, vol. 20, no. 22, p. 6671, 2020.

- [7] D. Suhandy and M. Yulia, "The use of partial least square regression and spectral data in uv-visible region for quantification of adulteration in indonesian palm civet coffee," *International Journal of Food Science*, vol. 2017, 2017.
- [8] M. Li and Q. Lu, "Changepoint detection in autocorrelated ordinal categorical time series," *Environmetrics*, vol. 33, no. 7, e2752, 2022.
- [9] S. Upadhyay, A. L. Stephenson, and D. G. Smith, "Readmission rates and their impact on hospital financial performance: A study of washington hospitals," *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, vol. 56, 2019.
- [10] *Centers for Medicare and Medicaid Services. readmissions reduction program*, <http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html>, Accessed: 2014-05-26, 2014.
- [11] A. Sarwar, C. A. Hostage Jr, J. L. Weinstein, *et al.*, "Causes and rates of 30-day readmissions after percutaneous transhepatic biliary drainage procedure," *Radiology*, vol. 290, no. 3, pp. 722–729, 2019.
- [12] N. Pathak, C. A. Kahlenberg, H. G. Moore, P. K. Sculco, and J. N. Grauer, "Thirty-day readmissions after aseptic revision total hip arthroplasty: Rates, predictors, and reasons vary by surgical indication," *The Journal of Arthroplasty*, vol. 35, no. 12, pp. 3673–3678, 2020.
- [13] A. Sarwar, L. Zhou, N. Chakrala, *et al.*, "The relevance of readmissions after common ir procedures: Readmission rates and association with early mortality," *Journal of Vascular and Interventional Radiology*, vol. 28, no. 5, pp. 629–636, 2017.



- [14] O. A. Panagiotou, A. Kumar, R. Gutman, *et al.*, “Hospital readmission rates in medicare advantage and traditional medicare: A retrospective population-based analysis,” *Annals of Internal Medicine*, vol. 171, no. 2, pp. 99–106, 2019.
- [15] R. J. Novick, S. A. Fox, L. W. Stitt, T. L. Forbes, and S. Steiner, “Direct comparison of risk-adjusted and non–risk-adjusted cusum analyses of coronary artery bypass surgery outcomes,” *The Journal of Thoracic and Cardiovascular Surgery*, vol. 132, no. 2, pp. 386–391, 2006.
- [16] J. C. Benneyan, “Statistical quality control methods in infection control and hospital epidemiology, part i introduction and basic theory,” *Infection Control & Hospital Epidemiology*, vol. 19, no. 3, pp. 194–214, 1998.
- [17] W. A. Shewhart, *Economic control of quality of manufactured product*. Macmillan and Co Ltd, London, 1931.
- [18] M. A. Mohammed, K. Cheng, A. Rouse, and T. Marshall, “Bristol, shipman, and clinical governance: Shewhart’s forgotten lessons,” *The Lancet*, vol. 357, no. 9254, pp. 463–467, 2001.
- [19] J. M. Lucas and M. S. Saccucci, “Exponentially weighted moving average control schemes: Properties and enhancements,” *Technometrics*, vol. 32, no. 1, pp. 1–12, 1990.
- [20] D. M. Hawkins and D. H. Olwell, *Cumulative sum charts and charting for quality improvement*. Springer Science & Business Media, 1998.
- [21] M. Aslam, A. Shafqat, M. Albassam, J.-C. Malela-Majika, and S. C. Shongwe, “A new cusum control chart under uncertainty with applications in petroleum and meteorology,” *Plos One*, vol. 16, no. 2, 2021.

- [22] T. B. Rasmussen, S. P. Ulrichsen, and M. Nørgaard, “Use of risk-adjusted cusum charts to monitor 30-day mortality in danish hospitals,” *Clinical Epidemiology*, vol. 10, p. 445, 2018.
- [23] E. S. Page, “Continuous inspection schemes,” *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.
- [24] M. R. De Leval, K. François, C. Bull, W. Brawn, and D. Spiegelhalter, “Analysis of a cluster of surgical failures: Application to a series of neonatal arterial switch operations,” *The Journal of Thoracic and Cardiovascular Surgery*, vol. 107, no. 3, pp. 914–924, 1994.
- [25] J. Neuburger, K. Walker, C. Sherlaw-Johnson, J. van der Meulen, and D. A. Cromwell, “Comparison of control charts for monitoring clinical performance using binary data,” *BMJ Quality & Safety*, vol. 26, no. 11, pp. 919–928, 2017.
- [26] O. Grigg and V. Farewell, “An overview of risk-adjusted charts,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 167, no. 3, pp. 523–539, 2004.
- [27] L. H. Sego, M. R. Reynolds Jr, and W. H. Woodall, “Risk-adjusted monitoring of survival times,” *Statistics in Medicine*, vol. 28, no. 9, pp. 1386–1401, 2009.
- [28] J. Li, J. Jiang, X. Jiang, and L. Liu, “Risk-adjusted monitoring of surgical performance,” *Plos One*, vol. 13, no. 8, 2018.
- [29] V. Parsonnet, D. Dean, and A. D. Bernstein, “A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease.,” *Circulation*, vol. 79, no. 6 Pt 2, pp. I3–12, 1989.

- [30] Y. Huang, A. Talwar, S. Chatterjee, and R. R. Aparasu, “Application of machine learning in predicting hospital readmissions: A scoping review of the literature,” *BMC Medical Research Methodology*, vol. 21, no. 1, pp. 1–14, 2021.
- [31] S. Shin, P. C. Austin, H. J. Ross, *et al.*, “Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality,” *ESC Heart Failure*, vol. 8, no. 1, pp. 106–115, 2021.
- [32] J. Futoma, J. Morris, and J. Lucas, “A comparison of models for predicting early hospital readmissions,” *Journal of Biomedical Informatics*, vol. 56, pp. 229–238, 2015.
- [33] Y. Huang, A. Talwar, Y. Lin, and R. R. Aparasu, “Machine learning methods to predict 30-day hospital readmission outcome among us adults with pneumonia: Analysis of the national readmission database,” *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1–14, 2022.
- [34] J. C. Stoltzfus, “Logistic regression: A brief primer,” *Academic Emergency Medicine*, vol. 18, no. 10, pp. 1099–1104, 2011.
- [35] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [36] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, pp. 1189–1232, 2001.
- [37] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.
- [38] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, “Variable selection using random forests,” *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

- [39] S. C. Larson, "The shrinkage of the coefficient of multiple correlation.," *Journal of Educational Psychology*, vol. 22, no. 1, p. 45, 1931.
- [40] M. R. Reynolds Jr and Z. G. Stoumbos, "A cusum chart for monitoring a proportion when inspecting continuously," *Journal of Quality Technology*, vol. 31, no. 1, pp. 87–108, 1999.
- [41] M. A. Jones and S. H. Steiner, "Assessing the effect of estimation error on risk-adjusted cusum chart performance," *International Journal for Quality in Health Care*, vol. 24, no. 2, pp. 176–181, 2012.
- [42] J. S. Ribeiro, T. d. J. G. Salva, and M. B. Silvarolla, "Prediction of a wide range of compounds concentration in raw coffee beans using nirs, pls and variable selection," *Food Control*, vol. 125, 2021.
- [43] Z. Shi, C. W. Chow, R. Fabris, J. Liu, and B. Jin, "Applications of online uv-vis spectrophotometer for drinking water quality monitoring and process control: A review," *Sensors*, vol. 22, no. 8, p. 2987, 2022.
- [44] T. Togkalidou, M. Fujiwara, S. Patel, and R. D. Braatz, "Solute concentration prediction using chemometrics and atr-ftir spectroscopy," *Journal of Crystal Growth*, vol. 231, no. 4, pp. 534–543, 2001.
- [45] L. Jiao, S. Bing, X. Zhang, and H. Li, "Interval partial least squares and moving window partial least squares in determining the enantiomeric composition of tryptophan by using uv-vis spectroscopy," *Journal of the Serbian Chemical Society*, vol. 81, no. 2, pp. 209–218, 2016.
- [46] K. Chawla, A. Bankapur, M. Acharya, J. S. D'Souza, and S. Chidangil, "A micro-raman and chemometric study of urinary tract infection-causing bac-

- terial pathogens in mixed cultures,” *Analytical and Bioanalytical Chemistry*, vol. 411, pp. 3165–3177, 2019.
- [47] M. B. Takahashi, J. Leme, C. P. Caricati, A. Tonso, E. G. Fernández Núñez, and J. C. Rocha, “Artificial neural network associated to uv/vis spectroscopy for monitoring bioreactions in biopharmaceutical processes,” *Bioprocess and Biosystems Engineering*, vol. 38, pp. 1045–1054, 2015.
- [48] J. S. Torrecilla, R. Aroca-Santos, J. C. Cancilla, and G. Matute, “Linear and non-linear modeling to identify vinegars in blends through spectroscopic data,” *LWT-Food Science and Technology*, vol. 65, pp. 565–571, 2016.
- [49] P. Díaz-Rodríguez, J. C. Cancilla, K. Wierzchoś, and J. S. Torrecilla, “Non-linear models applied to experimental spectroscopical quantitative analysis of aqueous ternary mixtures of imidazolium and pyridinium-based ionic liquids,” *Sensors and Actuators B: Chemical*, vol. 206, pp. 139–145, 2015.
- [50] O. M. Kvalheim, “Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots,” *Journal of Chemometrics*, vol. 24, no. 7-8, pp. 496–504, 2010.
- [51] R. D. Tobias *et al.*, “An introduction to partial least squares regression,” in *Proceedings of the Twentieth Annual SAS users Group International Conference*, Citeseer, vol. 20, 1995, pp. 1250–1257.
- [52] J. Gareth, W. Daniela, H. Trevor, and T. Robert, *An introduction to statistical learning: with applications in R*. Springer, 2013.
- [53] K. Kumar, “Discrete wavelet transform (dwt) assisted partial least square (pls) analysis of excitation-emission matrix fluorescence (eemf) spectroscopic data

- sets: Improving the quantification accuracy of eemf technique,” *Journal of Fluorescence*, vol. 29, pp. 185–193, 2019.
- [54] Ö. Yeniay and A. GÖKTAŞ, “A comparison of partial least squares regression with other prediction methods,” *Hacettepe Journal of Mathematics and Statistics*, vol. 31, pp. 99–111, 2002.
- [55] A. De Girolamo, V. Lippolis, E. Nordkvist, and A. Visconti, “Rapid and non-invasive analysis of deoxynivalenol in durum and common wheat by fourier-transform near infrared (ft-nir) spectroscopy,” *Food Additives and Contaminants*, vol. 26, no. 6, pp. 907–917, 2009.
- [56] Y. M. Fayez, S. M. Tawakkol, N. M. Fahmy, H. M. Lotfy, and M. A.-A. Shehata, “Comparative study of the efficiency of computed univariate and multivariate methods for the estimation of the binary mixture of clotrimazole and dexamethasone using two different spectral regions,” *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 194, pp. 126–135, 2018.
- [57] M. Zarenistanak, A. G. Dhorde, and R. Kripalani, “Trend analysis and change point detection of annual and seasonal precipitation and temperature series over southwest iran,” *Journal of Earth System Science*, vol. 123, pp. 281–295, 2014.
- [58] H. Assareh, I. Smith, and K. Mengersen, “Bayesian change point detection in monitoring cardiac surgery outcomes,” *Quality Management in Healthcare*, vol. 20, no. 3, pp. 207–222, 2011.
- [59] X. Shi, C. Gallagher, R. Lund, and R. Killick, “A comparison of single and multiple changepoint techniques for time series data,” *Computational Statistics & Data Analysis*, vol. 170, 2022.

- [60] Y. S. Niu, N. Hao, and H. Zhang, “Multiple change-point detection: A selective overview,” *Statistical Science*, pp. 611–623, 2016.
- [61] Q. Lu, R. Lund, and T. C. Lee, “An mdl approach to the climate segmentation problem,” 2010.
- [62] A. J. Scott and M. Knott, “A cluster analysis method for grouping means in the analysis of variance,” *Biometrics*, pp. 507–512, 1974.
- [63] S. Li and R. Lund, “Multiple changepoint detection via genetic algorithms,” *Journal of Climate*, vol. 25, no. 2, pp. 674–686, 2012.
- [64] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler, “Circular binary segmentation for the analysis of array-based dna copy number data,” *Biostatistics*, vol. 5, no. 4, pp. 557–572, 2004.
- [65] P. Fryzlewicz, “Wild binary segmentation for multiple change-point detection,” 2014.
- [66] S. Sankararaman and S. Mahadevan, “Model parameter estimation with imprecise and unpaired data,” *Inverse Problems in Science and Engineering*, vol. 20, no. 7, pp. 1017–1041, 2012.
- [67] K. Bleakley and J.-P. Vert, “The group fused lasso for multiple change-point detection,” *ArXiv Preprint ArXiv:1106.4199*, 2011.
- [68] M. Leyli-Abadi, A. Same, L. Oukhellou, *et al.*, “Online common change-point detection in a set of nonstationary categorical time series,” *Neurocomputing*, vol. 439, pp. 176–196, 2021.
- [69] A. Agresti, *Analysis of Ordinal Categorical Data*. John Wiley & Sons, 2010, vol. 656.

- [70] J. Wang, D. Ding, and Q. Su, “Latent change-point detection in ordinal categorical data,” *Quality and Reliability Engineering International*, vol. 35, no. 2, pp. 504–516, 2019.
- [71] G. Müller and C. Czado, “An autoregressive ordered probit model with application to high-frequency financial data,” *Journal of Computational and Graphical Statistics*, vol. 14, no. 2, pp. 320–338, 2005.
- [72] C. Varin and P. Vidoni, “Pairwise likelihood inference for ordinal categorical time series,” *Computational Statistics & Data Analysis*, vol. 51, no. 4, pp. 2365–2373, 2006.
- [73] C. Varin and P. Vidoni, “Pairwise likelihood inference for general state space models,” *Econometric Reviews*, vol. 28, no. 1-3, pp. 170–185, 2008.
- [74] N. R. Zhang and D. O. Siegmund, “A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data,” *Biometrics*, vol. 63, no. 1, pp. 22–32, 2007.
- [75] R. B. Lund, C. Beaulieu, R. Killick, Q. Lu, and X. Shi, “Good practices and common pitfalls in climate time series changepoint techniques: A review,” *Journal of Climate*, vol. 36, no. 23, pp. 8041–8057, 2023.
- [76] N. Khan, S. McClean, S. Zhang, and C. Nugent, “Optimal parameter exploration for online change-point detection in activity monitoring using genetic algorithms,” *Sensors*, vol. 16, no. 11, p. 1784, 2016.
- [77] T. Polushina and G. Sofronov, “Change-point detection in biological sequences via genetic algorithm,” in *2011 IEEE Congress of Evolutionary Computation (CEC)*, IEEE, 2011, pp. 1966–1971.