

# Contributions to Change Point and Functional Data Analysis

by

Jeremy VanderDoes

A thesis  
presented to the University of Waterloo  
in fulfillment of the  
thesis requirement for the degree of  
Doctor of Philosophy  
in  
Statistics

Waterloo, Ontario, Canada, 2025

© Jeremy VanderDoes 2025

## Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner: Rafal Kulik  
Professor, Dept. of Mathematics and Statistics  
University of Ottawa

Supervisors: Gregory Rice  
Professor, Dept. of Statistics and Actuarial Science  
University of Waterloo

Shojaeddin Chenouri  
Professor, Dept. of Statistics and Actuarial Science  
University of Waterloo

Internal Members: Paul Marriott  
Professor, Dept. of Statistics and Actuarial Science  
University of Waterloo

Alex Stringer  
Assistant Professor, Dept. of Statistics and Actuarial Science  
University of Waterloo

Internal-External Member: Hossein Abouee Mehrizi  
Professor, Dept. of Management Science and Engineering  
University of Waterloo

## Author's Declaration

This thesis consists of material all of which I authored or co-authored; see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

## Statement of Contributions

The work contained in this thesis has led to several published, submitted, and draft manuscripts. Chapter 2 is based on the published work in [Horváth et al. \(2025\)](#). Chapter 3 is based on a manuscript with Shojaeddin Chenouri that is submitted for publication. Chapter 4 is based on a manuscript with Gregory Rice and Martin Wendler that is currently being prepared for submission. Chapter 5 is based on the published work in [VanderDoes et al. \(2024\)](#). Chapter 6 is based on the published work in [Horváth et al. \(2024\)](#).

The contributions to the articles by Jeremy VanderDoes included deriving the methods, writing codes, conducting simulation studies and data analysis, and preparing the manuscripts. The codes are mainly written in R, but have some extensions to other languages, e.g. C++, for computational efficiency and speed. The codes are freely accessible at [gi thub. com/ j rvanderdoes](https://github.com/jrvanderdoes) and CRAN.

## Abstract

The advent and progression of computers has led to consideration of data previously considered too unwieldy. So called high-dimensional, or big, data can be considered large in both the size of observations and the number of observations. In this thesis, we consider such data which may be infinite dimensional and is often collected over some dimension, such as time. Methodology for detection of changes and exploration of this information-rich data is explored.

Chapter 1 provides a review of concepts and notation used throughout the thesis. Topics related to time series, functional data, and change point analysis are of particular interest and form the foundation of the thesis. The chapter concludes with an overview of the main contributions contained in the thesis.

An empirical characteristic functional-based method for detecting distributional change points in functional time series is presented in chapter 2. Although various methods exist to detect changes in functional time series, they typically require projection or are tuned to specific changes. The characteristic functional-based approach is fully functional and sensitive to general changes in the distribution of functional time series. Integrated- and supremum-type test statistics are proposed. Theoretical considerations for the test statistics are examined, including asymptotic distributions and the measure used to integrate the test statistic over the function space. Simulation, permutation, and approximation approaches to calibrate detection thresholds for the test statistics are investigated. Comparisons to existing methods are conducted via simulation experiments. The proposed methods are applied to continuous electricity prices and high-frequency asset returns.

Chapter 3 is devoted to graph-based change point detection. Graph-based approaches provide another method for detecting distributional changes in functional time series. Four test statistics and their theoretical properties are discussed. Extensive simulations provide context for graph-based tuning parameter choices and compare the approaches to other functional change point detection methods. The efficacy of graph-based change point detection is demonstrated on multi-year pedestrian counts, high-resolution stock returns, and continuous electricity prices.

Despite increased interest in functional time series, available implementations are largely missing. Practical considerations for applying functional change point detection are covered in chapter 4. We present fChange, a functional time series package in R. The package combines and expands functional time series and change point methods into an easy-to-use format. The package provides functionality to store and process data, summarize and validate assumptions, characterize and perform inference of change points, and provide visualizations. The data are stored as discretely observed observations, promoting usability

and accuracy. Applications to continuous electricity prices, cancer mortality, and long-term treasury rates are shown.

In chapter 5, we propose novel methodology for analyzing tumor microenvironments (TMEs) in cancer research. TMEs contain vast amounts of information on patient’s cancer through their cellular composition and the spatial distribution of tumor cells and immune cell populations. We present an approach to explore variation in TMEs, and determine the extent to which this information can predict outcomes such as patient survival or treatment success. Our approach can identify specific interactions which are useful in such predictions. We use spatial  $K$  functions to summarize interactions, and then apply a functional random forest-based model. This approach is shown to be effective in simulation experiments at identifying important spatial interactions while also controlling the false discovery rate. We use the proposed approach to interrogate two real data sets of Multiplexed Ion Beam Images of TMEs in triple negative breast cancer and lung cancer patients. The publicly available companion R package `funkycells` is discussed.

The random coefficient autoregressive model of order 1, RCA(1), is a model well-suited for volatile time series. Detection of changes between stable and explosive regimes of scalar data modeled with the RCA(1) is explored in chapter 6. We derive a (maximally selected) likelihood ratio statistic and show that it has power versus breaks occurring even as close as  $O(\log \log N)$  periods from the beginning/end of sample. Moreover, the use of quasi maximum likelihood-based estimates yields better power properties, with the added bonus of being nuisance-free. Our test statistic has the same distribution - of the Darling-Erdős type - irrespective of whether the data are stationary or not, and can therefore be applied with no prior knowledge on this. Our simulations show that the test has very good power and, when applying a suitable correction to the asymptotic critical values, the correct size. We illustrate the usefulness and generality of our approach through applications to economic and epidemiological time series.

Chapter 7 provides summaries and discussions on each chapter. Directions for future work are considered. These directions, with the provided commentary, extend the scope of the models and may behoove practitioners and researchers alike.

## Acknowledgements

I never imagined how quickly my time at the University of Waterloo would pass. My time has cultivated countless memories and life-long friends. It would require a thesis to simply thank everyone who made this experience possible. So although I name but a few specific people below, a sincere thank you for all who supported and enriched my journey.

I would like to begin by thanking my supervisors, Gregory Rice and Shojaeddin Chenouri. They have offered invaluable guidance, patience, and support. The completion of this thesis simply would not have been possible without their continual support and encouragement. I am so fortunate to have worked with them. Their wisdom has been invaluable in research and in life. Words are inadequate to convey my immense gratitude.

I also express my heartfelt gratitude to the members of my thesis committee—Paul Marriott, Alex Stringer, Hossein Abouee Mehrizi, and Rafak Kulik. Your time and insights have greatly improved this work, and my approach to research. Your questions have led me to consider topics previously overlooked.

I would have never studied statistics at the University of Waterloo if not for the guidance of Lajos Horváth. Thank you for all you taught about statistics and inspired me to achieve. Your effect is seen throughout this thesis.

Thank you to all those in the department who provided support, both seen and unseen, throughout this journey. A special thanks to Mary Lou Dufton for answering any questions and solving unsolvable problems. Profound thanks to Joslin Goh for her insight in all kinds of problems and connecting me to interesting projects and wonderful people. Thank you to the many friends I've made at Waterloo. I enjoyed the conversations with each of you. I express my deep appreciation for my office mates and collaborators.

My academic journey would have been impossible without the support of my family. Thank you for all your love, interest, and support. Special thanks to my parents, who also made sure I also enjoyed the beauty of Canada with regular visits. I can't express all my feelings for the support of my beautiful wife. The patience you have offered for all my late nights and long meetings is awe inspiring. Your many sacrifices have not been overlooked.

Again, thank you all. Your desire to learn and improve drives me when I am stuck.

## Dedication

This is dedicated to my wife, parents, and family.

# Table of Contents

|                                     |       |
|-------------------------------------|-------|
| Examining Committee                 | ii    |
| Author's Declaration                | iii   |
| Statement of Contributions          | iv    |
| Abstract                            | v     |
| Acknowledgements                    | vii   |
| Dedication                          | viii  |
| List of Figures                     | xiv   |
| List of Tables                      | xxvii |
| List of Abbreviations               | xxx   |
| 1 Introduction                      | 1     |
| 1.1 Motivation . . . . .            | 1     |
| 1.2 Time Series . . . . .           | 5     |
| 1.3 Functional data . . . . .       | 9     |
| 1.4 Change point analysis . . . . . | 14    |

|       |  |    |
|-------|--|----|
| 1.5   | Spatial statistics . . . . .   | 17 |
| 1.6   | Outline . . . . .  | 19 |
| 2     | Change Point Analysis for Functional Data Using Empirical Characteristic Functionals . . . . . | 21 |
| 2.1   | Introduction . . . . .   | 21 |
| 2.2   | Main results . . . . .   | 22 |
| 2.2.1 | Asymptotic results under $H_0$ . . . . .   | 25 |
| 2.2.2 | Consistency under $H_A$ and estimating $k$ . . . . .   | 27 |
| 2.3   | Estimating the distributions appearing in Theorem 2.2.1 . . . . .                              | 29 |
| 2.3.1 | Gaussian process simulation . . . . .  | 29 |
| 2.3.2 | Welch-style approximation ( $T_n$ only) . . . . .  | 31 |
| 2.3.3 | Permutation . . . . .  | 32 |
| 2.3.4 | Multiple change detection . . . . .  | 33 |
| 2.4   | Simulation study . . . . .   | 34 |
| 2.4.1 | Data generation . . . . .  | 34 |
| 2.4.2 | Null simulations . . . . .   | 37 |
| 2.4.3 | Single change point . . . . .  | 38 |
| 2.5   | Data applications . . . . .  | 42 |
| 2.5.1 | Spanish electricity data . . . . .   | 43 |
| 2.5.2 | S&P 500 data . . . . .   | 44 |
| 2.6   | Conclusion . . . . .   | 47 |
| 3     | Graph-Based Change Point Analysis for Functional Data . . . . .                                | 48 |
| 3.1   | Introduction . . . . .   | 48 |
| 3.2   | Functional graph-based change point analysis . . . . .   | 49 |
| 3.2.1 | Single change analysis . . . . .   | 50 |
| 3.2.2 | Multiple change analysis . . . . .   | 54 |

|         |  |    |
|---------|--|----|
| 3.2.3   | Graph construction . . . . .                             | 55 |
| 3.3     | Simulations . . . . .                                    | 57 |
| 3.3.1   | No change simulations . . . . .                          | 60 |
| 3.3.2   | Single change simulations . . . . .                      | 62 |
| 3.3.3   | Multiple change simulations . . . . .                    | 69 |
| 3.4     | Applications . . . . .                                   | 69 |
| 3.4.1   | Australian pedestrian counts . . . . .                   | 72 |
| 3.4.2   | Twitter stock returns . . . . .                          | 73 |
| 3.4.3   | Spanish electricity prices . . . . .                     | 75 |
| 3.5     | Discussion . . . . .                                     | 76 |
| 4       | fChange: Functional Change Point Analysis in R . . . . . | 78 |
| 4.1     | Introduction . . . . .                                   | 78 |
| 4.2     | Data representation . . . . .                            | 80 |
| 4.2.1   | The dfts class . . . . .                                 | 80 |
| 4.2.2   | Functions of the dfts class . . . . .                    | 81 |
| 4.3     | Functionality of fChange . . . . .                       | 83 |
| 4.3.1   | Data . . . . .   | 83 |
| 4.3.2   | Visualizations . . . . .                                 | 84 |
| 4.3.3   | Data processing and diagnostic tests . . . . .           | 86 |
| 4.3.3.1 | Projection . . . . .                                     | 89 |
| 4.3.3.2 | Stationarity test . . . . .                              | 90 |
| 4.3.3.3 | KPSS test . . . . .                                      | 91 |
| 4.3.3.4 | Mitigate stationarity violations . . . . .               | 93 |
| 4.3.3.5 | White noise tests . . . . .                              | 93 |
| 4.3.4   | Change point detection . . . . .                         | 94 |
| 4.3.4.1 | Mean change . . . . .                                    | 96 |
| 4.3.4.2 | Second-order change . . . . .                            | 98 |

|         |   |     |
|---------|---|-----|
| 4.3.4.3 | Distributional change   | 101 |
| 4.3.4.4 | Multiple changes  | 103 |
| 4.3.4.5 | Confidence intervals  | 106 |
| 4.4     | Summary and discussion  | 109 |
| 5       | Using Random Forests to Uncover the Predictive Power of Distance-Varying Cell Interactions in Tumor Microenvironments | 110 |
| 5.1     | Introduction  | 110 |
| 5.2     | Materials and methods   | 114 |
| 5.2.1   | Summarizing the image data using K functions  | 117 |
| 5.2.2   | Variable importance   | 120 |
| 5.2.3   | Variable importance comparison  | 121 |
| 5.2.4   | Comparing variable importance of spatial interactions and meta-variables to noise                                     | 123 |
| 5.2.5   | Choice of parameters $B$ and $H$  | 125 |
| 5.2.6   | Predictive accuracy estimates   | 126 |
| 5.2.7   | Variable importance plot  | 127 |
| 5.3     | Results   | 128 |
| 5.3.1   | Simulation study  | 129 |
| 5.3.2   | Applications to MIBIScope data  | 136 |
| 5.3.2.1 | Triple negative breast cancer   | 136 |
| 5.3.2.2 | Lung adenocarcinoma versus lung squamous cell carcinoma   | 138 |
| 5.4     | Discussion  | 141 |
| 6       | The Maximally Selected Likelihood Ratio Test in Random Coefficient Models   | 143 |
| 6.1     | Introduction  | 143 |
| 6.2     | The test statistics   | 147 |
| 6.2.1   | Assumptions   | 148 |

|         |  |     |
|---------|--|-----|
| 6.2.2   | Asymptotics . . . . .  | 149 |
| 6.2.2.1 | Consistency . . . . .  | 152 |
| 6.3     | Simulations . . . . .  | 153 |
| 6.4     | Empirical applications . . . . .   | 155 |
| 6.4.1   | The US housing market . . . . .  | 157 |
| 6.4.2   | Covid-19 UK hospitalisation data . . . . .                               | 158 |
| 6.5     | Concluding remarks . . . . .   | 160 |
| 7       | Conclusion . . . . .   | 161 |
| 7.1     | Discussions . . . . .  | 161 |
| 7.2     | Future work . . . . .  | 162 |
|         | References . . . . .   | 165 |
|         | APPENDICES . . . . .   | 195 |
| A       | Chapter 2 Appendix . . . . .   | 196 |
| A.1     | Proofs . . . . .   | 196 |
| A.2     | Selection of $B$ and $R$ . . . . .                                       | 202 |
| A.3     | Null tables . . . . .  | 205 |
| B       | Chapter 4 Appendix . . . . .   | 212 |
| B.1     | Data summary table . . . . .   | 212 |
| B.2     | Edge corrections . . . . .   | 215 |
| B.3     | Surrogate splitting . . . . .  | 215 |
| B.4     | Choosing $B$ and $H$ . . . . .   | 216 |
| B.5     | OOB accuracy and naïve model comparisons . . . . .                       | 217 |
| B.6     | Performance in the low sample, high interaction count scenario . . . . . | 218 |
| C       | Chapter 5 Appendix . . . . .   | 221 |
| C.1     | Technical lemmas . . . . .   | 221 |
| C.2     | Proofs . . . . .   | 230 |

# List of Figures

|     |   |    |
|-----|---|----|
| 1.1 | Spanish electricity prices. Hourly electricity prices in 2014 for Spain in the day-ahead market (MIBEL). Colors gradually change over time to aid the reader. . . . .   | 2  |
| 1.2 | Lung cancer biopsy. Point pattern resulting from the biopsy of a lung cancer patient and two related $\mathcal{K}$ functions of cell interactions. (a) Standardized cell pattern from a biopsy of a lung cancer patient using multiplexed ion beam imaging (MIBI) technology. Colored points indicate different cell types. For example, there are a large number of epithelial (green), stromal (purple), and B-cells (red) in the image. (b) $\mathcal{K}$ functions for the biopsy image. The red line is the B-cells $\mathcal{K}$ function and the green line is the epithelial cells $\mathcal{K}$ function. The dotted gray line is the theoretical $\mathcal{K}$ function when no interactions are present. . . . . | 3  |
| 1.3 | COVID-19 hospitalisations, Northern Ireland. Daily count of Covid-19 patients in Northern Ireland (UK) hospitals from March 2020 to December 2021. . . . .  | 4  |
| 2.1 | Data generating process power curves. Power curves as a function of the change magnitude based on mid-sample mean changes on data generated via different DGPs. $T_n$ simulation (red) and $M_n$ simulation (blue) are compared, along with ARS-18 (black) for two different data length settings— $n = 100$ (solid) and $n = 500$ (dashed). 1,000 simulations are computed at each point, with $h = 1$ . The characteristic approaches show good power in all cases. . . . .   | 39 |

|     |   |    |
|-----|---|----|
| 2.2 | Mean change power curves. Power curves based on mid-sample mean changes on data generated via ARKL DGP. $T_n$ simulation (red) and $M_n$ simulation (blue) are compared, along with ARS-18 (black) for length 100 data. 1;000 simulations are computed at each point, with independence assumed ( $h = 1$ ). (a) considers the mean change of the form $\mu_i(t) = c \sin(2\pi K t)$ for $c = -1, 1, 2, 4$ . Several choices for $K$ are considered. The characteristic approaches show loss of power while ARS-18 maintains good power for increasing values of $K$ . (b) considers a mean change, but with errors with distribution $t_1$ rather than normality. In this case the characteristic approaches show much higher power than ARS-18, which has no power. . . . . | 40 |
| 2.3 | Higher order change power curves. Power curves based on mid-sample changes on data generated via different the ARKL DGP. $T_n$ simulation (red) and $M_n$ simulation (blue) are compared, along with HRZ-22 (black) for two different data length settings—100 (solid) and 500 (dashed). 1;000 simulations are computed at each point, with independence assumed ( $h = 1$ ). The characteristic approaches show good power in all cases. . . . .   | 41 |
| 2.4 | Location power curves. Rejection rate for a minor mean change (0.15) from ARKL generated data of length 100 (solid) and 500 (dashed) in 1;000 simulations, with independence assumed ( $h = 1$ ). The $T_n$ (red), $M_n$ (blue), ARS-18 (black) and mARS-18 (purple) statistics are compared. $T_n$ outperforms the other methods which are generally comparable. . . . .   | 42 |
| 2.5 | Spanish electricity. Daily, hourly observed Spanish electricity spot prices in 2014. . . . .  | 43 |
| 2.6 | Spanish electricity changes. Changes detected on the residuals of the HU(7) ETS forecast for the 2014 Spanish electricity data, placed on the residual and original electricity spot price data. Colors indicate different segments of stationarity, showing three change points detected using the $T_n$ simulation approach. . . . .  | 45 |
| 2.7 | SPDR S&P 500. SPDR S&P 500 ETF OCIDRs based on one-minute resolution data from 2019 to 2023. . . . .  | 46 |
| 2.8 | SPDR S&P 500 changes. S&P 500 OCIDR curves based on one-minute resolution data from 2019 to 2023. Changing colors segments demarcated by five change points detected using the $T_n$ with binary segmentation. . . . .  | 47 |

|     |  |    |
|-----|--|----|
| 3.1 | Example graph constructions. Graphs constructed on 20 observations which were randomly placed according to a multivariate normal distribution. The graphs are constructed using the following trees: (a) minimal spanning tree (MST), (b) minimal distance pairing (MDP), and (c) nearest neighbour linking (NNL). No orthogonal trees were used. . . . .  | 56 |
| 3.2 | Example $\mathcal{K}$ -graph construction. A NNL-2 graph constructed on 20 observations which were randomly placed according to a multivariate normal distribution. The black lines indicate the first NNL tree; see Figure 3.1c. The red lines indicate the second NNL tree, which is orthogonal to the first tree. . . . .   | 57 |
| 3.3 | Mean power curves for normal errors. Power curves as a function of the change magnitude based on mid-sample mean changes. The change magnitude varied from $-1$ to $1$ with samples of length $n = 50$ . The number of orthogonal trees, $\mathcal{K}$ , are colored by increasing blue intensity. The dotted horizontal gray line is the nominal significance level. Power curves were constructed using the generalized test statistic as in Eq (3.6). Power curves by graph types are displayed in each figure: (a) MDP, (b) MST, and (c) NNL. Power tended to increase with a larger number of orthogonal trees. MST and NNL were marginally more powerful than MDP. . . . . | 61 |
| 3.4 | Mean power curves for skewed errors. Power curves as a function of the change magnitude based on mid-sample mean changes with normal and skewed-normal error distributions. The change magnitude varied from $0$ to $1$ with samples of length $n = 50$ . The top plot used data normal errors while the middle and bottom plots used increasingly skewed-normal errors. The methods are generally comparable. The weighted and original methods performed better than the generalized or max-type statistic, especially for the MDP graph. The original test statistic exhibited power comparable to ARS-18. . . . .  | 62 |

|     |   |    |
|-----|---|----|
| 3.5 | Mean power curves for $t$ -distribution errors. Power curves as a function of the change magnitude for a mid-sample mean change with errors distributed according to a $t$ -distribution. The change magnitude varied from 0 to 1 with samples of length $n = 50$ . The dotted horizontal gray line is the nominal significance level. The top figure was generated from data with $t_{10}$ errors, the middle with $t_3$ errors, and the bottom with $t_1$ errors; see also Figure 3.4 for power under Gaussian errors. As the normality assumption was increasingly violated, the graph-based methods increasingly outperformed ARS-18 in terms of size and power. The graph types performed similarly. The test statistics were also similar, although original and weighted typically performed better than generalized and max-type. . . . | 63 |
| 3.6 | Power curves for $t_1$ errors with varying data length. Power curves as a function of data length for a mid-sample mean change with $t_1$ errors. The change magnitude was set to 0.5 for data lengths of $n = 50; 100; 200$ . The dotted horizontal gray line is the nominal significance level. Improvement over increasing data length $n$ is observed for all graph-based approaches, but not for ARS-18. MST and NNL performed better than MDP. The original statistic performed better than the weighted and max-type. The weighted and max-type statistics performed better than the generalized statistic. . .  | 64 |
| 3.7 | Eigenvalue change power curves. Power curves as a function of the change magnitude based on mid-sample eigenvalue changes. The change magnitude was varied from 1 to 8 for sample length $n = 50$ . The dotted horizontal gray line is the nominal significance level. For MST and NNL graphs, the graph-based statistics quickly surpassed HRZ-22 in terms of power. MDP graphs performed worse than HRZ-22 for smaller eigenvalue changes and better for larger eigenvalue changes. Interestingly, the weighted statistic was not strongly influenced by the graph type, where the other graph-based test statistics exhibited increased power for MST and NNL graphs. . . . .  | 65 |
| 3.8 | Distributional change power curves. Power curves as a function of the change magnitude based on mid-sample distributional changes. The change magnitude was varied from 5 to 0.1 for sample length $n = 50$ . The dotted horizontal gray line is the nominal significance level. Both the graph-based and HRV-25 methods improved with increased distributional differences. The graph-based methods performed better for extreme cases and for the B-spline basis, but HRV-25 performed better for moderate changes with the Fourier basis. . . . .  | 66 |

|      |   |    |
|------|---|----|
| 3.9  | Mean change location power curves. Power curves as a function of the change location based on mean changes. Change magnitudes of $\delta = 0.25; 0.5$ for length $n = 100$ data was considered. The change location $k$ was varied from 5 to 95. The dotted horizontal gray line is the nominal significance level. The graph-based approaches exhibited more robustness to changes at the edges of the data than ARS-18. ARS-18 exhibited more power for mid-sample changes, particularly for mild changes. The weighted statistic performed best among graph-based methods, particularly with MST and NNL graph types. . . . .  | 67 |
| 3.10 | Size-adjusted mean change power curves under dependency. Power curves as a function of dependency magnitude based on mid-sample mean changes. Change magnitude $\delta = 0.5$ for $n = 50$ length data was considered. The dependency magnitude $\rho$ was varied from 0 to 0.9. The size-adjusted rejection rate is reported. Loss of power due to dependency is typically comparable between graph-based and ARS-18 methods, while MST and NNL performed better than MDP. . . . .   | 68 |
| 3.11 | Multiple mean change location estimates. Estimates of mean change point locations based on data with 2 changes. For data of length $n = 200$ , the magnitude change $\delta$ changed from 0 to 0.75 to 0 at locations $k_1 = 60$ and $k_2 = 170$ . The true mean model is given in the top plot. Histograms for the estimated change point locations from 100 simulations are given in the lower plots. Vertical red lines indicate the locations of the true changes. The graph-based methods and ARS-18 show robustness to the changed interval, or epidemic change, scenario. The graph-based methods had more early sample false detections while ARS-18 had more mid-sample false detections. The graph-based methods were generally comparable. . . . . | 70 |

|      |   |    |
|------|---|----|
| 3.12 | Multiple distributional change location estimates. Estimates of distributional change point locations based on data with 3 changes. For data of length $n = 250$ , the data changed at locations $k_1 = 30$ , $k_2 = 100$ , and $k_3 = 200$ . A mean change $\mu = 0.75$ occurred at $k_1$ . A distributional change for the errors $\epsilon$ 's changed from normal to exponential, both standardized, at $k_2$ . A covariance change $\Sigma = 3$ occurred at $k_3$ . A representation of the true model is given in the top plot. Histograms for the estimated change point locations from 100 simulations are given in the lower plots. Vertical red lines indicate the locations of the true changes. The graph-based methods and HRV-25 commonly detected the first change, with HRV-25 perhaps performing slightly better. However, the graph-based methods detected the distributional and variance changes far more commonly. . . . . | 71 |
| 3.13 | Melbourne pedestrian counts. Hourly pedestrian counts on a street in Melbourne Australia from 2019 to 2021. . . . .   | 72 |
| 3.14 | Segmented Melbourne pedestrian counts. Segmented hourly pedestrian counts on a street in Melbourne Australia (2019 - 2021). Colors and line types indicate approximately homogeneous regions defined by change points detected using MST-15 and the max-type test statistic in Eq (3.11) with $L^2$ distance. . . . .   | 73 |
| 3.15 | Twitter CIDRs. Twitter daily CIDRs (2019 - 2020), recorded at a one-minute resolution. . . . .  | 74 |
| 3.16 | Segmented Twitter CIDRs. Segmented daily Twitter CIDRs (2019 - 2020), recorded at one-minute resolution. Approximately homogeneous regions are indicated by colors and line types. The regions are demarcated by estimated changes from the max-type graph-based test statistic defined in Eq (3.11) using a MST-15 graph and $L^2$ distance. . . . .   | 75 |
| 3.17 | Electricity and its model residuals. Hourly electricity prices in 2014 for Spain and the $HU(7)$ electricity model residuals. (a) The hourly electricity prices. (b) The residuals of the electricity prices when modeled using the $HU(7)$ model with a Holt-Winters exponential smoothing model and weekly seasonality. . . . .   | 76 |
| 3.18 | Segmented electricity residuals. Segmentation for the residuals of the $HU(7)$ electricity model based on max-type graph-based change point detection using MST-15 and $L^2$ distance. Regions of approximate homogeneity are indicated by line type and line color. . . . .  | 77 |

|     |   |     |
|-----|---|-----|
| 4.1 | Electricity summary. Example view of electricity from the function <code>summary()</code> . The top plot is a spaghetti plot of the data. The middle left plot is an ACF plot with 20 lags, where the blue line is the threshold under the strong white noise assumption and the red line is the threshold under the weak white noise assumption. The middle right is a QQ plot for the first 2 functional principal components. The bottom plot displays the $p$ -values for white noise tests up to 20 lags. The red dotted line is the nominal size, $\alpha = 0.05$ .   | 82  |
| 4.2 | Spaghetti plots. Rainbow spaghetti plots for electricity, cancer, and rates.  | 85  |
| 4.3 | 2-dimensional rainbow plots. 2-dimensional rainbow plots for electricity and cancer. The thin, dotted, and colored lines are the observations. The thick gray lines are the pointwise means of the data.  | 87  |
| 4.4 | Autocorrelation functions. Functional autocorrelation functions computed from cancer and cancerresiduals. The threshold under strong white noise is given in blue and the threshold under weak white noise is given in red. (a) The cancer data exhibits strong serial correlation. A large number of violations are observed when compared to the assumption of strong and weak white noise. (b) The cancerresiduals exhibit no strong serial correlation. Only a lag 6 spike is observed, which is below the weak white noise threshold. The spike may be due to chance, or an indication that the weak white noise assumption is more appropriate. | 95  |
| 4.5 | Cancer segmentation. Segmentation for cancer and cancerresiduals using the fully functional mean change model with binary segmentation. (a) The 7 detected mean changes segment the cancer data. Colors indicate approximately homogeneous regions. (b) No mean changes were detected in cancerresiduals. A rainbow spaghetti plot is shown.  | 105 |
| 4.6 | Rates segmentation. Segmentation for rates using the elbow plot method. (a) Variability-change plot showing the total variability as a function of the number of changes. The red vertical line indicates 6 change points are algorithmically suggested. (b) Spaghetti plot for the rates data with 6 change points. Approximately homogeneous regions are indicated using the same color.  | 106 |

|     |   |     |
|-----|---|-----|
| 4.7 | Electricity segmentation and intervals. Segmentation for electricity from the fully functional distributional change point detection method based on characteristic functionals. Confidence intervals are computed. Colors indicate approximately homogeneous regions. The black curves indicate the observations immediately surrounding change points. The gray curves indicate the confidence intervals for change points. The observed electricity prices are displayed in light colors while the mean of the region is displayed with dark colors. . . . .   | 109 |
| 5.1 | An example of point pattern data and an associated $K$ function. (a) Point pattern data associated with a tumor imaged using MIBIScope from a triple negative breast cancer patient with multiple identified phenotypes. The $x$ - and $y$ -axes represent the spatial dimensions, with the points giving individual cell locations, and the colour of the points indicating one of 15 unique phenotypes, e.g. tumor (red), NK cells (purple), and monocytes/neutrophils (cyan). (b) The associated cross $K$ function (black) for two cell types in the image: tumor and monocytes/neutrophils. The $x$ -axis indicates the radius, $r$ , and the $y$ -axis gives the value of the $K$ function. The estimated $K$ function can be compared to $r^2$ (red dashed line), which is the theoretical $K$ function associated with complete spatial randomness. | 112 |
| 5.2 | Flow chart for data processing. The methods presented here begin with tabular data obtained after pre-processing multiplex images (steps that include cell segmentation, phenotyping, etc.). For a given image the tabular data consists of rows for each imaged cell, giving the associated $x$ - $y$ position, marker intensities, and cell phenotype. Next, the tabular data are converted into spatial $K$ functions for each interaction of interest (this can be exhaustive, and include all possible interactions between phenotypes, or selective, with only a subset of interactions analysed). Next, $K$ functions are converted into functional principal component scores. Patient meta-variables are added at this stage. The resulting data is then used in the statistical model, as described in Figure 5.3. . . . .                        | 116 |

|     |   |     |
|-----|---|-----|
| 5.3 | Flow chart of model. When modeling using <code>funkycells</code> , there are several major steps: organizing data, generating synthetic data, and modeling using random forests. The spatial data is organized into functional summaries ( $K$ functions) that are projected into finite dimensions (FPCA) and used with meta-variables to predict the outcome variable. The spatial data and meta-variables are permuted to create synthetic variables with similar properties but independent of the outcome. These synthetic variables are then added to the model, and used to quantify the strength of the relationships between the spatial and meta-data with the response. The model processes the data, employing cross-validation and permutation to return a variable importance plot (with predictive accuracy estimates) indicating spatial interactions and/or meta-variables which are significant in predicting the outcome $Z$ . . . . . | 117 |
| 5.4 | Sample variance importance plot. This sample variable importance plot uses simulated data with a binary outcome, two cell types, and two meta-variables. The data was simulated with significant differences between the outcomes in the $B\_B$ , $A\_B$ spatial interactions, and age meta-variable, but no significant difference across sex and the $A\_A$ spatial interaction. The point estimates of the variable importance values are the black dots, with accompanying intervals indicating the uncertainty. The red dotted straight line is the noise threshold and the orange dashed curved line is the interpolation threshold. Both thresholds are used to indicate if a variable is predictive of the outcome beyond that of random noise. The variable importance values of the known significant variables are shown to exceed that of the noise and interpolation thresholds. . . . .   | 128 |
| 5.5 | Comparison of TNBC and simulated data. (a) An image from the TNBC data and (b) an image from the simulated data. Different colors indicate one of the 16 different cell phenotypes, showing the comparability of the simulations and true data. . . . .   | 131 |
| 5.6 | No relationship simulation. Simulation of 16 cell types for 34 patients with meta-variable age. (a) Figure with the variable importance values for all variables. (b) Figure with only the top 25 largest variable importance values. All variables were generated with no-relationship to the outcome and all were determined to have no relation to the outcome beyond noise as the variable importance estimates are below noise and interpolation thresholds. . . . .   | 132 |

|     |   |     |
|-----|---|-----|
| 5.7 | <p>Relationship simulation. Simulation of 16 cell types for 34 patients with meta-variable age. (a) Figure with the variable importance values for all variables. (b) Figure with only the top 25 largest variable importance values. Most cell types were generated with no relationship to the outcome. However, age, <math>c1\_c2</math>, and <math>c1\_c3</math> were designed to have a relationship with the outcome (which naturally means <math>c2\_c2</math> and <math>c2\_c3</math> would also have relationships to the outcomes). These variables are seen with significantly larger variable importance values than the thresholds and other variable importance values. . . . .</p>   | 133 |
| 5.8 | <p>Power curves. Power curves showing the empirical rate, from 100 simulations, that the variable importance for the spatial interaction <math>c1\_c2</math> exceeded the 95% noise threshold, interpolation threshold and both the noise and interpolation thresholds, indicating a significant spatial interaction is detected. The curves indicate the threshold method: above both thresholds (teal), above the curved interpolation threshold (orange), and straight noise threshold (red) (the colours correspond with the noise curves used in the variable importance plots). The <math>x</math> axis gives the standard deviation parameter controlling the distribution of <math>c2</math> cells around <math>c1</math> cells. The vertical black dotted line is the base case, when both classes exhibit the same interactions across all cell types, including the <math>c1\_c2</math> interaction. To the right of this line, <math>c2</math> cells are less densely clustered around <math>c1</math> cells, and to the left of the line there is increased clustering around <math>c1</math> cells. The light horizontal, dotted gray line indicates the desired mis-classification rate when no signal is present (i.e. 0.05). (a) Spatial data with 4 cell types is used to create the curves. (b) Spatial data with 16 cell types is used to create the curves. Both images show the method is effective at correctly detecting when the important interaction does and does not differ between the patient outcomes. This is seen as all lines quickly climb to 1 (perfect detection of a signal) as the standard deviation parameter moves further from the no effect case (vertical line). The size and power is similar between the simulations despite the large increase in total interactions considered. . . . .</p> | 135 |

|      |   |     |
|------|---|-----|
| 5.9  | TNBC variable importance. Variable importance plot and random forest model summary for predicting “compartmentalized” versus “mixed” tumor types with the TNBC data. (a) Figure with the variable importance values for all variables. (b) Figure with only the top 25 largest variable importance values. The OOB far exceeds those of naïve models, and many of the spatial interactions between tumor cells and immune cell populations exhibited significant variable importance values, suggesting important interactions in the data (such as Tumor_Tumor). . . . .   | 138 |
| 5.10 | Example TNBC $K$ functions. The $K$ functions from the different outcomes are compared in these two plots. In both plots, the $x$ -axis indicates the radial distance, $r$ , in micrometers and the $y$ -axis is the value of the $K$ function. The lightly colored lines are the $K$ functions for each patient, while the bold lines indicates the average (point-wise mean). In the figures, red indicates the mixed tumors while blue indicates the compartmentalized tumors. The black dashed line indicates the curve of a totally spatially random process for reference. (a) Plots the $K$ functions for the Tumor_Tumor interaction, which was found to have significant differences in the outcomes. (b) Plots the $K$ functions for the CD4T_Endothelial interaction, which was found have no significant differences between the outcomes. In (a), as expected, the compartmentalized group has relatively larger $K$ functions—indicating increased clustering—and the functions are well grouped together. Conversely, (b) shows no clear differences between the $K$ functions of the two groups and $K$ functions are generally surrounded by $K$ functions from patients of an assortment of the groups. That is, $K$ function patterns vary widely even within the same outcome groups. . . . . | 139 |
| 5.11 | Lung cancer variable importances. Variable importance plot and random forest model summary for predicting LUAD versus LUSC tumor types. (a) Figure with the variable importance values for all variables. (b) Figure with only the top 25 largest variable importance values. The OOB is similar to a naïve model, and none of the measured variable importance value were statistically significant, indicating no significant variable interactions or meta-variables. . . . .  | 140 |
| 6.1  | Stationary region. Simulated stationary region using Gaussian errors based on $E \log j + 0.1j$ . . . . .   | 154 |

|     |   |     |
|-----|---|-----|
| 6.2 | Power curves. Power curve as a function of the change in the RCA coefficient constructed from the percentage of rejections of $H_0$ with nominal level $\alpha = 0.05$ . (upper left panel) $\beta = 0$ , change point is at $0.5n$ , using Gaussian errors. (upper right panel) $\beta = 0$ , change point is at $0.9n$ , using Gaussian errors. (lower left panel) $\beta = 0.5$ , change point is at $0.5n$ , using Gaussian errors. (lower right panel) $\beta = 0.5$ , change point is at $0.5n$ , using Bernoulli errors. . . . . | 156 |
| 6.3 | Housing changes. Los Angeles (left) and Boston (right) daily housing price indices (top plots) and the estimated beta values (bottom plots) with detected change points between January 4th, 1995, and 23rd October, 2012. . . . .  | 157 |
| 6.4 | Covid-19 hospitalisation changes. Daily Covid-19 patients hospitalised for the four UK nations (top plot) and the estimated beta value (bottom plot). The England (upper left), Northern Ireland (upper right), Scotland (lower left), and Wales (lower right) data starts in March 2020 and ends on December 31, 2021. . . . .   | 159 |
| A.1 | Parameter choices. Investigations into the selection and effects of $B$ and $R$ . Box-plots, labeled with the respective sizes, are shown for the test statistics $T_n$ and $M_n$ —Figures (a) and (b)—and their logged values—Figures (c) and (d). All data is length 100 generated from ARKL DGP given in section 3.3. Each setting is run for 1,000 iterations to generate the given box-plots. When $B$ or $R$ are being investigated, we respectively set $R = 50$ and $B = 20$ . . . . .  | 203 |
| A.2 | $B$ power curves. Power curves based on mid-sample mean changes for varying choices of $B$ . All data is length 100 generated from ARKL DGP given in section 3.3. $T_n$ simulation (solid) and $M_n$ simulation (dot-dash) are compared for moderate changes. 1,000 simulations are computed at each point. Beyond $B = 1$ , there seems to be negligible power gain for a given test statistic, with $T_n$ performing better than $M_n$ . Each setting displayed approximately nominal empirical size. . . . .                         | 204 |

|     |  |     |
|-----|--|-----|
| B.1 | Selection of $\mathcal{B}$ . Simulations of into the number of synthetics, $B$ , for (a) 4 and (b) 16 cell scenarios. In both, only one interaction, $c2\_c2$ , is significant, and all other interactions are non-significant. Per Figure 8 in the main manuscript, the expected effect is mild (standard deviation in placement of the cells from the Thomas process changing from 1=25 to 1=60). Running 50 trials for significance of 0.05, the power is reasonably level across selections of $B$ , excluding (expected) low power when the number of synthetics is far too low to accurately approximate the desired quantile. . . . . | 217 |
| B.1 | Detection rate. Mean detection rate of the significant interaction (red) and mean detection rate of all non-significant interactions (blue) in a small sample scenario with an increasing number of cell types. In each scenario, 100 iterations per scenario, all interactions were considered ( $T(T + 1)=2$ interactions). As the number of interactions increase, the method suffers only minor power loss while controlling for false discovery rate. . . . .   | 220 |

# List of Tables

|     |  |    |
|-----|--|----|
| 2.1 | $T_n$ size table. Empirical size for $T_n$ test statistic using simulation, permutation, and approximation based on 1,000 simulations from a ARKL DGP with Gaussian errors. Data lengths of 100, 250, and 500 are considered under varying degrees of dependence. Several bandwidths used to compute the long-run variance estimators and permutation block sizes, both denoted $h$ , are also considered. The nominal level is set at 0.05. . . . .     | 36 |
| 2.2 | $M_n$ size table Empirical size for $M_n$ test statistic using simulation and permutation based on 1,000 simulations from a ARKL DGP with Gaussian errors. Data lengths of 100, 250, and 500 are considered under varying degrees of dependence. Several window and block sizes, both denoted $h$ , are also considered. The nominal level is set at 0.05. . . . .   | 37 |
| 3.1 | Size table under $L^2$ distance. Table for the empirical sizes of the graph-based change point detection methods. Combinations of test statistics, tree types, and orthogonal tree counts tuning parameters were examined. Each graph-based test statistic was computed using the $L^2$ distance. Several data lengths were considered, $n = 15; 25; 50; 100; 200$ . In general, the graph-based change point detection methods were well-sized. . . . . | 59 |
| 3.2 | Size table under varying distance metrics. Table for the empirical sizes of the graph-based change point detection statistics with varying tree types and orthogonal tree counts using $L^2$ and $L^1$ distance metrics. Data were length $n = 50$ . Both $L^2$ and $L^1$ were typically well-sized, with the exception of MDP-1 which was very conservative. . . . .  | 60 |

|     |  |     |
|-----|--|-----|
| 5.1 | False discovery table. Table showing empirical false discovery rate based on 100 independent simulations for both $T = 4$ and $T = 16$ cell types with 34 patients split into positive and negative groups. The false discovery rate is computed as the percentage of simulations for which any VI exceeded the corresponding threshold or both thresholds (Variable Importance Only), any of the cross-validated one standard error intervals for the VI lies to the right of the corresponding threshold (Variable Importance-1 CVSD), and whether the largest VI computed exceeded the corresponding threshold (Largest Variable Importance). . . . . | 134 |
| 6.1 | Size table. Empirical rejection frequencies under the null . . . . .   | 155 |
| A.1 | $T_n$ size table at <b>0:01</b> significance. Empirical size for $T_n$ test statistic using simulation, permutation, and approximation based on 1,000 simulations from a ARKL DGP with Gaussian errors. Data lengths of 100, 250, and 500 are considered under varying degrees of dependence. Several bandwidths used to compute the long-run variance estimators and permutation block sizes, both denoted $h$ , are also considered. The nominal level is set at 0.01. . . . .   | 206 |
| A.2 | $M_n$ size table at <b>0:01</b> significance. Empirical size for $M_n$ test statistic using simulation and permutation based on 1,000 simulations from a ARKL DGP with Gaussian errors. Data lengths of 100, 250, and 500 are considered under varying degrees of dependence. Several bandwidths used to compute the long-run variance estimators and permutation block sizes, both denoted $h$ , are also considered. The nominal level is set at 0.01. . . . .   | 207 |
| A.3 | $T_n$ AROU size table. Empirical size for $T_n$ test statistic using simulation, permutation, and approximation based on 1,000 simulations from a AROU DGP. Data lengths of 100, 250, and 500 are considered under varying degrees of dependence. Several bandwidths used to compute the long-run variance estimators and permutation block sizes, both denoted $h$ , are also considered. The nominal level is set at 0.05. . . . .   | 208 |
| A.4 | $M_n$ AROU size table. Empirical size for $M_n$ test statistic using simulation and permutation based on 1,000 simulations from a AROU DGP. Data lengths of 100, 250, and 500 are considered under varying degrees of dependence. Several bandwidths used to compute the long-run variance estimators and permutation block sizes, both denoted $h$ , are also considered. The nominal level is set at 0.05. . . . .   | 209 |

|     |   |     |
|-----|---|-----|
| A.5 | $T_n$ FAR(1) size table. Empirical size for $T_n$ test statistic using simulation, permutation, and approximation based on 1,000 simulations from a FAR(1) DGP. Data lengths of 100, 250, and 500 are considered under varying degrees of dependence. Several bandwidths used to compute the long-run variance estimators and permutation block sizes, both denoted $h$ , are also considered. The nominal level is set at 0.05. . . . .  | 210 |
| A.6 | $M_n$ FAR(1) size table. Empirical size for $M_n$ test statistic using simulation and permutation based on 1,000 simulations from a FAR(1) DGP. Data lengths of 100, 250, and 500 are considered under varying degrees of dependence. Several bandwidths used to compute the long-run variance estimators and permutation block sizes, both denoted $h$ , are also considered. The nominal level is set at 0.05. . . . .  | 211 |
| B.1 | Lung cancer summary table. Table summarizing the patients, number of images, type of lung cancer, and the cell count for each type. . . . .   | 212 |
| B.2 | Null summary metrics table. In 100 simulations without differences between the outcome groups (i.e., the null scenario), we compare the out-of-bag (OOB) accuracy of the true model to the naïve approaches of always selecting the most common outcome (BIAS) and selecting a random outcome based on frequency (GUESS)—see Predictive accuracy estimates and Simulation study sections in the main manuscript for more details. The table shows the methods are unskewed and all randomly report the largest values about an equal number of times (about one third for each metric). Compare this table to Table B.3. . . . .  | 218 |
| B.3 | Alternative summary metrics table. In 100 simulations with a difference between the outcome groups (i.e., the “alternative” scenario), we compare the out-of-bag (OOB) accuracy of the true model to the naïve approaches of always selecting the most common outcome (BIAS) and selecting a random outcome based on frequency (GUESS)—see Predictive accuracy estimates and Simulation study sections in the main manuscript for more details. The table shows the methods are unskewed and OOB consistently reports the largest values, showing the model does capture the existing information—that is the model does seem to recognize informative interactions and/or meta-variables. Compare this table to Table B.2. . . . . | 219 |

# List of Abbreviations

- ACF autocorrelation function 6, 7
- AMOC at most one change 14–16, 23, 28, 32–34, 50, 62, 96, 143, 152, 153
- ARIMA autoregressive integrated moving average 9, 79
- ARMA autoregressive moving average 8, 53, 163
- BS binary segmentation xv, xx, 16, 34, 42, 44, 46, 47, 49, 54, 69, 79, 103–105, 108, 156
- CSR complete spatial randomness 17, 18
- CUSUM cumulative sum 15, 16, 22, 24, 41, 48, 87, 94, 96, 97, 99, 100, 102, 145, 146, 153, 154, 160
- CV cross-validation 14, 122, 126, 127
- ETS error-trend-seasonality xv, 9, 45, 93
- FD functional data 2, 9–11, 19, 47, 48, 58, 78, 79, 81, 86, 118, 161, 163
- FDA functional data analysis 9, 13, 19, 78, 112, 119, 141
- FPC functional principal component xx, 12, 13, 82, 89
- FPCA functional principal component analysis xxii, 11, 13, 75, 79, 89, 117, 119, 163
- FTS functional time series 10, 11, 19, 21, 22, 25, 26, 30, 32–35, 43, 48–51, 53, 73, 75, 79, 80, 83, 85, 86, 89, 91, 93, 94, 102, 103, 109, 161, 163
- GARCH generalized autoregressive conditional heteroskedasticity 8

IID independent and identically distributed 5, 7, 9, 25, 27, 33, 36, 84, 92, 100, 102, 146, 147, 154

KL Karhunen-Loève 12, 13, 58

LRV long-run variance 7, 8, 99, 107

MIBI multiplexed ion beam imaging xiv, 3

PCA principal component analysis 13

PELT pruned exact linear time 16

RCA(1) order one random coefficient additive 4, 20, 143–145, 153, 156–158, 164

SARIMA seasonal autoregressive integrated moving average 9

TSA time series analysis 5

TVE total variance explained 13, 75, 76, 90, 93, 98, 163

# Chapter 1

## Introduction

### 1.1 Motivation

This thesis is focused on models for data that may be infinite dimensional and exhibit patterns of dependence. In many cases, such data also exhibit structural changes. In each case, the goal remains to understand the data enough to apply and understand an appropriate model. To motivate our work, consider the following examples.

Example 1.1.1 (Electricity Usage). *Consider the one day-ahead market prices for Spanish electricity in 2014, measured hourly. The data are presented in Figure 1.1. The daily prices may be thought of as observations of functions of the form*

$$fX_i(t) : i = 1, \dots, 365 ; t \in [0, 1]g;$$

*denoting the electricity price on day  $i$  at the (standardized) intraday time  $t$ .*

*A natural use for this data is to forecast the price. In order to do so, it is useful to understand the structural patterns of the data. Approximately homogeneous regions in the data can often be demarcated via change point detection. Such regions show patterns or can be used so estimates only depend on relevant data. Several clear patterns can be observed in Figure 1.1. The electricity prices are related to usage, and exhibit daily and seasonal fluctuations. This dependence extends beyond the expected usage to varying usage distribution. There are consistently high prices in the summer—perhaps due to regular use of air conditioners—while in colder seasons high prices are volatile and more concentrated to daytime hours.*

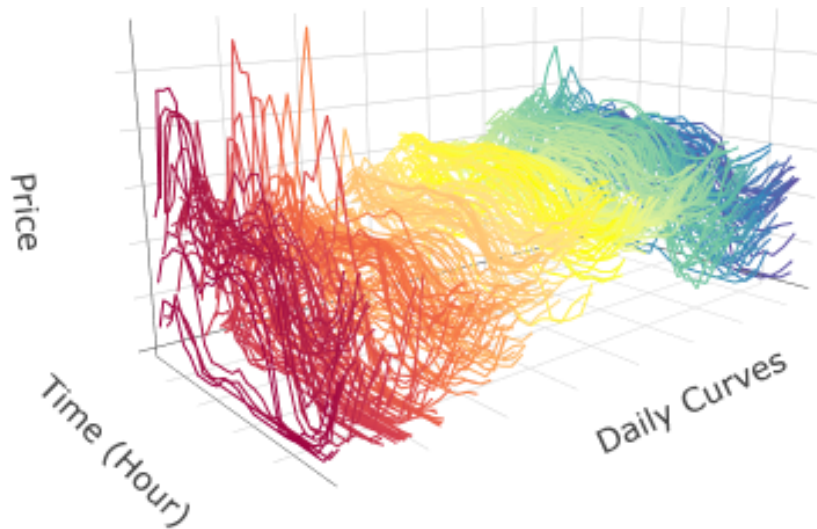


Figure 1.1: Spanish electricity prices. Hourly electricity prices in 2014 for Spain in the day-ahead market (MIBEL). Colors gradually change over time to aid the reader.

*Detecting a change point requires a method that is sensitive to the potential change. Chapters 2 and 3 present approaches to detect distributional changes in [functional data](#). Comparisons to other functional change point detection methods are also discussed. Yet for practitioners, selecting and implementing a model may be prohibitively complex. Chapter 4 presents a freely accessible **R** package which provides methodology for functional change point detection, **fChange**.*

Example 1.1.2 (Lung Cancer). *Consider the 2-dimensional diagram of cells and their locations retrieved from a biopsy of a lung cancer patient presented in Figure 1.2a. The image, or point pattern, represents each cell by a colored point indicating its type. The spatial data can be represented in the form*

$$fZ(s; m) : s \in D; m \in M; g;$$

*where  $M$  is a space containing the cell types and  $D$  is the 2-dimensional space.*

*Cell interactions can often be detected as clustering or repulsion between cell types in point processes. These interactions can indicate key traits such as a treatment's viability. One measure of interactions that can simplify data-dense images are  $K$  functions. For a point process with  $n$  cells on region  $D \subset \mathbb{R}^2$ , the  $K$  function is defined*

$$K(r) = \frac{1}{D} E[\text{Number of (extra) cells within distance } r];$$

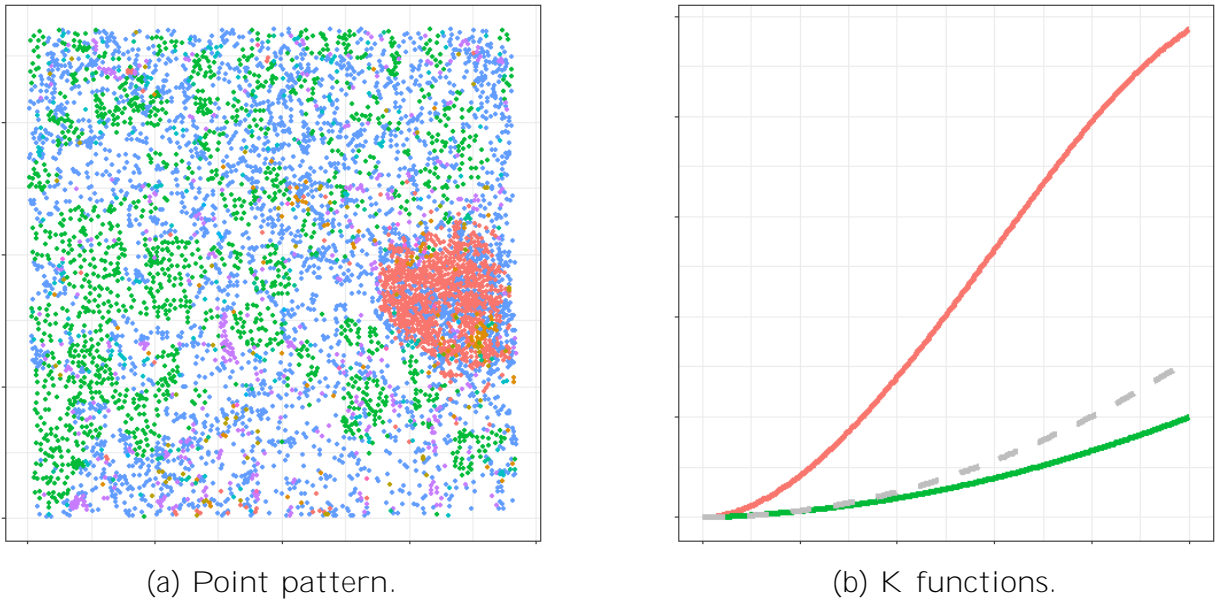


Figure 1.2: Lung cancer biopsy. Point pattern resulting from the biopsy of a lung cancer patient and two related  $K$  functions of cell interactions. (a) Standardized cell pattern from a biopsy of a lung cancer patient using [multiplexed ion beam imaging \(MIBI\)](#) technology. Colored points indicate different cell types. For example, there are a large number of epithelial (green), stromal (purple), and B-cells (red) in the image. (b)  $K$  functions for the biopsy image. The red line is the B-cells  $K$  function and the green line is the epithelial cells  $K$  function. The dotted gray line is the theoretical  $K$  function when no interactions are present.

where  $D = \frac{n}{jD_j}$ . When no interactions are present,  $K(r) = r^2$ . Regularity between cells tends to reduce  $K(r)$  while clustering tends to increase it. An extension to the classic  $K$  function that allows examination of interactions between cell types is discussed in section 1.5. Example  $K$  functions are given in Figure 1.2b.

Studying cancer biopsy images using  $K$  functions is the main goal of Chapter 5. A model for prediction using cell interactions, and supporting data, from biopsies of different patients is provided. The model aims to differentiate between significant and insignificant interactions while accounting for false detection.

Example 1.1.3 (Covid-19 Pandemic). Figure 1.3 presents the daily hospitalisations due to Covid-19, as given by the UK government, for Northern Ireland from March 2020 through December 2021. There is clear temporal dependence between observations as well as explo-

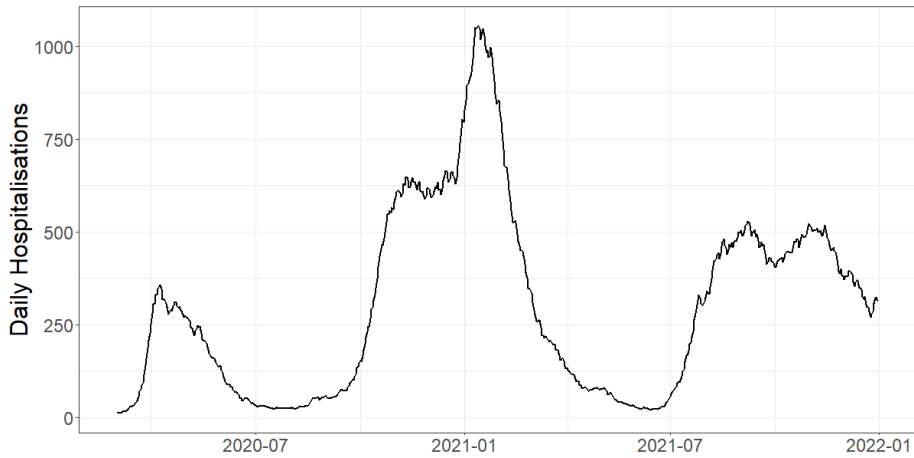


Figure 1.3: COVID-19 hospitalisations, Northern Ireland. Daily count of Covid-19 patients in Northern Ireland (UK) hospitals from March 2020 to December 2021.

*sive and stable periods. Such data volatility is typical of epidemiological data, as well as data such as housing or stock market indices. Determining the points at which the data changes from explosive to stable regimes can provide valuable information, e.g. indicate the effectiveness of governmental policies or highlight natural patterns. Such change points can also suggest times when preparation is crucial.*

*The daily hospitalisation data can be thought of as observations of the form*

$$fX_t : t = 1; 2; \dots; ng;$$

*where  $t$  indicates the day of observation, up to day  $n$ . Data with periods of stability and explosions or crashes are well modeled by the [order one random coefficient additive \(RCA\(1\)\) model](#). Yet there is limited literature on detecting regime changes in the [RCA\(1\) model](#). [Chapter 6](#) provides a framework for change point detection based on a maximally selected likelihood ratio statistic.*

To address these and other examples, several methodologies have been developed in this thesis. The remainder of this chapter reviews foundational information relevant for exploring these contributions. Some more specific details will be left to each chapter. This chapter concludes with an outline of our contributions.

## 1.2 Time Series

Many data are collected sequentially over time. Examples can be found in nearly every field, e.g. daily temperature, hourly radiation emitted, or movements of a robot. Such data can be referred to as time series data. Although time series data and its analysis are prevalent throughout history, formalization of [time series analysis \(TSA\)](#) is often first attributed to works such as [Yule \(1926\)](#) and [Yule \(1927\)](#). The publication of [Box and Jenkins \(1970\)](#) popularized [TSA](#) among many general practitioners. Continued development has resulted in more modern monographs, such as [Brockwell and Davis \(1987\)](#), [Hamilton \(1994\)](#), and [Shumway and Stoffer \(2017\)](#).

Theoretically, time series may be thought of as a realization of random variables indexed according to the order in which they were obtained. A collection of indexed random variables is often referred to as a stochastic process, and hence a time series process is a particular type of stochastic process. Let a time series process be denoted  $fX_t : t \in \mathbb{Z}$ . The parameter  $t$  is often thought of as “time”, although in many cases time series arise from other, sequentially observed processes. An *observed* time series is a particular realization of the process. The realization is typically observed at discrete, equally-spaced points, say  $fX_t : t = 0; 1; \dots; ng$ . Even when the process is naturally continuous, the data will typically be collected discretely due to technological constraints and storage. It is common to use the term “time series” to refer to both the data and the process from which it arises. This approach is taken in the current thesis when the meaning is clear from context.

Due to the temporal collection, an intrinsic property of time series is that adjacent or nearly adjacent observations are often neither independent nor identically distributed. Rather a complete description at time points  $t_1; \dots; t_d$  requires the joint finite dimensional distributions, evaluated as the probability that the values of series are jointly less than constants  $c_1; \dots; c_d$ ,

$$F_{t_1; \dots; t_d}(c_1; \dots; c_d) = P(X_{t_1} < c_1; \dots; X_{t_d} < c_d) :$$

Because of this, traditional statistical methods that rely on the assumption that the data are [independent and identically distributed \(IID\)](#) must be adapted.

A strong, or strictly, stationary time series is one where the probabilistic behavior of every finite collection of variables is invariant to time shifts. In other words, for all  $k = 1; 2; \dots$ , time points  $t_1; \dots; t_k \in \mathbb{Z}$ , constants  $c_1; \dots; c_k \in \mathbb{R}$ , and time shifts  $h = 1; 2; \dots$ , then

$$P(X_{t_1} < c_1; \dots; X_{t_k} < c_k) = P(X_{t_1+h} < c_1; \dots; X_{t_k+h} < c_k) :$$

Strong stationarity is a very desirable property, although most observed time series do not appear strictly stationary. Moreover, it is generally difficult to assess whether such an assumption is plausible from a single data set.

Weak stationarity assumes a constant mean, and covariance that is only dependent on time separation across the entire data set. For some time series  $\{X_t; t \in \mathbb{Z}\}$  with finite variance,  $X_t$  is weakly stationary if

$$E(X_t) = \mu; \quad \mu \in \mathbb{R};$$

and

$$\text{Cov}(X_t, X_{t+h}) = \gamma(h);$$

for all  $t$  and  $h$ , where  $\gamma$  is a nonnegative definite function on the positive integers. When the covariance function acts on the same series at two different time points  $t$  and  $s$ , it is frequently referred to as autocovariance and denoted  $\gamma(t; s)$ .

Unless otherwise specified, throughout this thesis, the term “stationarity” refers to strict stationarity. To check if the data are stationary, often the [autocorrelation function \(ACF\)](#) is employed. The [ACF](#) is defined as

$$\rho(s; t) = \frac{\text{Cov}(X_s, X_t)}{\sqrt{\text{Var}(X_s)\text{Var}(X_t)}};$$

Since all stationary processes only depend on the time separation of  $X_s$  and  $X_t$ ,  $h = |s - t|$ , and not on the locations in time,  $\rho(s; t) = \rho(h; 0) = \rho(h)$ .

Because the mean functions  $E(X_t)$  for stationary time series are not dependent on time  $t$ , the sample mean is an appropriate estimate

$$\bar{X} = \frac{1}{n} \sum_{t=1}^n X_t;$$

Likewise, it is appropriate to use the sample autocovariance function as an estimate for the autocovariance function

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-h} (X_t - \bar{X})(X_{t+h} - \bar{X});$$

for  $h = 1, \dots, n-1$ .

Under mild assumptions that  $X_t$  is IID and  $EjX_t^4 < 1$ , the asymptotic distribution of the ACF is given as

$$\sqrt{n} \hat{\rho}(h) \xrightarrow{D} N(0, 1);$$

where  $\xrightarrow{D}$  denotes convergence in distribution; see Billingsley (1968). Technical details on the asymptotic distribution of the ACF can be found in Shumway and Stoffer (2017). The ACF has benefits which include ease of interpretability, clear visual representation, and connection to the Pearson correlation function. Yet the ACF only considers relationships between each lag individually so combination effects may be missed. Such limitations are addressed in portmanteau tests, e.g. see Ljung and Box (1978).

If  $X_1; \dots; X_n$  is a realization of a stationary process, the variance of the sample mean should be calculated

$$\text{Var}(X_n) = \frac{1}{n^2} \text{Var} \left( \sum_{i=1}^n X_i \right) = \frac{1}{n^2} \sum_{i,j=1}^n \text{Cov}(X_i; X_j);$$

Yet due to stationarity, the covariances only depend on the differences  $h = i - j$ . Thus,

$$\text{Var}(X_n) = \frac{1}{n^2} \sum_{h=-n+1}^{n-1} (n - |h|) \gamma_h = \frac{1}{n} \sum_{h=-n+1}^{n-1} \left( 1 - \frac{|h|}{n} \right) \gamma_h;$$

The dominated convergence theorem implies that if  $\sum_{h=-\infty}^{\infty} |h| \gamma_h < 1$ ,

$$\lim_{n \rightarrow \infty} n \text{Var}(X_n) = \sum_{h=-\infty}^{\infty} \gamma_h;$$

This quantity defines the long-run variance (LRV),

$$\sigma_{LRV}^2 = \sum_{h=-\infty}^{\infty} \gamma_h = \gamma_0 + 2 \sum_{h=1}^{\infty} \gamma_h;$$

See Parzen (1957) and Grenander and Rosenblatt (1957).

The LRV is often estimated using a kernel bandwidth long-run variance estimator

$$\hat{\sigma}_{LRV}^2 = \hat{\sigma}_{LRV}^2(K; q) = \sum_{h=-n+1}^{n-1} K \left( \frac{h}{q} \right) \hat{\gamma}_h$$

with the sample autocovariance function  $\hat{\gamma}_h$ . The kernel  $K$  is continuous and has the properties  $K(-x) = K(x)$  and  $K(0) = 1$ . The parameter  $q$  is the bandwidth. The choice of kernel and bandwidth has spawned many discussions; e.g. see [Newey and West \(1987\)](#), [Politis and Romano \(1995\)](#), and [Sun et al. \(2008\)](#). Selection of the bandwidths are often more important than selection of the kernels, as any reasonable kernel can achieve close to optimal results ([Epanechnikov, 1969](#)). Additional details can be found in the seminal bandwidth selection papers by [Bartlett \(1950\)](#) and [Parzen \(1957\)](#). Data-driven selection methods have also been explored; see [Horváth et al. \(2016\)](#) and [Rice and Shang \(2017\)](#). The [long-run variance](#) was extended to long-run covariance matrices in [Newey and West \(1987\)](#), [Andrews \(1991\)](#), and [Andrews and Monahan \(1992\)](#).

Perhaps the most basic time series process is the white noise process which we denote as  $w_t \sim WN(0; \frac{2}{w})$ . A process  $w_t$  is a (weak) white noise if  $E(w_t) = 0$  and

$$\text{Cov}(w_t; w_{t-h}) = \begin{cases} \frac{2}{w}; & h = 0 \\ 0; & h \neq 0 \end{cases}.$$

Many time series models attempt to capture the innate autocorrelation and serial dependence between observations by assuming the series is generated from a driving white noise innovation sequence  $\{w_t; t \in \mathbb{Z}\}$ . These models take the form

$$X_t = g(w_t; w_{t-1}; \dots);$$

and are often called Bernoulli shifts ([Wu, 2005](#)). Bernoulli shifts are stationary processes. They are also causal, meaning they do not depend on any future innovations. Many common time series models fit within the (linear) Bernoulli-shift framework, e.g. the [autoregressive moving average \(ARMA\)](#) model ([Whittle, 1951](#)). A common (nonlinear) Bernoulli-shift process is the [generalized autoregressive conditional heteroskedasticity \(GARCH\)](#) model ([Engle, 1982](#); [Bollerslev, 1986](#)).

Unfortunately, observed time series are often nonstationary. If the data are thought to be nonstationary, one may attempt to convert the data to a stationary form by using a method such as detrending or differencing.

Detrending requires modeling the data's dependence and using this model to remove trends so that only stationary residuals remain. Many time series models have been developed to capture such dependency; e.g. see [Whittle \(1951\)](#), [Box and Jenkins \(1970\)](#), and [Shumway and Stoffer \(2017\)](#).

Differencing directly uses the data to remove trends. Let  $B$  be defined as the backshift operator,  $B^d X_t = X_{t-d}$  for  $d = 0; 1; 2; \dots$ . An  $d$ 'th differenced time series is defined

$$Y_t = (1 - B)^d X_t;$$

Often incorporating differencing or detrending, models have been proposed to mitigate nonstationarity; e.g. autoregressive integrated moving average (ARIMA) (Box and Jenkins, 1970), seasonal autoregressive integrated moving average (SARIMA) (Box et al., 2008), and error-trend-seasonality (ETS) models (Holt, 1957; Winters, 1960; Brown, 1963; Gardner, 2006; Hyndman et al., 2008).

Perhaps the most common nonstationary time series process is the random walk. A random walk is obtained by cumulatively combining IID random variables,  $Z_t$ , such that

$$S_t = Z_1 + Z_2 + \dots + Z_t; \quad \text{for } t = 1; 2; \dots;$$

It is easy to show  $E(S_t) = 0$  and  $\text{var}(S_t) = t$ .

A, in a sense limiting, version of the random walk is a Wiener process, or standard Brownian motion. We say

$$fW(x); x \geq 0g$$

is such a process if it is a continuous Gaussian process defined for  $x \geq 0$  with  $E[W(x)] = 0$  and  $E[W(x)W(y)] = \min(x; y)$ . See Breiman (1968) or Csörgö and Révész (1981) for construction and properties of the Wiener process. The continuous Gaussian process

$$fB(t) = W(t) - tW(1); 0 \leq t \leq 1g$$

is called a Brownian bridge. It follows that  $E[B(t)] = 0$  and  $E[B(t)B(s)] = \min(t; s) - ts$ .

## 1.3 Functional data

Though recent technological advances have drawn statisticians to so-called “big data”, these data are still typically thought of as taking values in some finite dimensional space. The fundamental philosophy of functional data analysis (FDA) is considering observed data as infinite dimensional objects. Originally coined in Ramsay (1982), FDA fits a finite number of discrete observations to functions and uses these, in principle, infinite dimensional objects throughout the analysis. Although traditionally associated with curves, any data varying over a domain can be viewed as functional data (FD), including pictures or sound waves (Ramsay and Silverman, 2002, 2005). Additional features are often ascribed to FD objects, e.g. continuity or differentiability. Such concepts are difficult to apply to standard multivariate data, but are easily accommodated in FDA.

FD objects are typically framed as realizations of stochastic processes indexed by some set. These processes are defined on a general probability space  $(\Omega; \mathcal{F}; \mathbf{P})$  with index set  $E$ ,

$$fX(t; \omega) : t \in E; \omega \in \Omega$$

for  $\mathcal{F}$ -measurable  $X(t; \cdot)$ . For a given  $\omega \in \Omega$ , the function  $X(\cdot; \omega)$  is called a realization of the process  $X$ , or a sample path. For convenience,  $\omega$  is generally suppressed.

A sample  $X_1(\cdot); \dots; X_n(\cdot)$  of such realizations constitute a realization of FD. When sequentially observed, the data are often referred to as **functional time series (FTS)**. FTS are commonly assumed to take value in a separable Hilbert space  $H$ , with inner product denoted  $\langle x; y \rangle$ , for  $x; y \in H$ , and corresponding norm  $\|x\|^2 = \langle x; x \rangle$ . A typical space for FTS are  $L^2([0; 1]; \mathbf{R})$ , denoting the space of square integrable real valued functions on the unit interval. This space is so common for FTS, it is often (interchangeably) denoted  $L^2[0; 1]$  or  $L^2$ .

There are functional summary statistics for FTS that correspond to traditional summary statistics. The covariance operator of some stochastic process  $X(t)$  is defined as

$$C : H \rightarrow H; C(y) = E[\langle X(t) - E(X(t)), y \rangle (X(t) - E(X(t)))] :$$

When  $H = L^2([0; 1]; \mathbf{R})$ , then

$$C(y)(t) = E\left[\left(\int_0^1 (X(s) - E(X(s))) y(s) ds\right) (X(t) - E(X(t)))\right] = \int_0^1 c(t; s) y(s) ds;$$

where

$$c(t; s) = \text{Cov}[X(t); X(s)]; \quad t; s \in [0; 1] :$$

Thus the covariance operator  $C$  is an integral operator with kernel  $c(t; s)$ , called covariance function or covariance kernel. The operator  $C$  is symmetric, nonnegative-definite, and the eigenvalues satisfy  $\sum_{j=1}^{\infty} \lambda_j < \infty$ . Empirically, the covariance operator can be estimated

$$\hat{c}(t; s) = \frac{1}{n} \sum_{i=1}^n [X_i(t) - \bar{X}(t)][X_i(s) - \bar{X}(s)];$$

where  $\bar{X}(t) = \frac{1}{n} \sum_{i=1}^n X_i(t)$ .

Since the seminal paper of [Ramsay \(1982\)](#), many methods have been developed for use with FD. For example see the monographs [Ramsay and Silverman \(2002\)](#), [Ramsay](#)

and Silverman (2005), Ferraty and Vieu (2006), Horváth and Kokoszka (2012), Hsing and Eubank (2015), and Kokoszka and Reimherr (2017). Ullah and Finch (2013) review applications in various fields; see also Wang et al. (2016) and Koner and Staicu (2023). Likewise, FTS has experienced upsurge in exposure since the seminal monograph of Bosq (2000); e.g. Hörmann and Kokoszka (2012), Horváth and Kokoszka (2012), and Kokoszka and Reimherr (2017).

Scalar models often consider a response  $Y$  based on covariates  $X$  and take the form

$$Y = f(X) + \epsilon;$$

where both  $Y$  and  $X$  are finite dimensional. FD considers models where (i)  $Y$  is infinite dimensional and  $X$  is finite dimensional, (ii)  $Y$  is finite dimensional and  $X$  is infinite dimensional, and (iii)  $Y$  and  $X$  are infinite dimensional (Horváth and Kokoszka, 2012). The term fully functional model is used when both  $Y$  and  $X$  are infinite dimensional. This model, originally presented as the varying coefficient model by Hastie and Tibshirani (1993), can encompass the mixed dimensional models.

A common approach to model FD is to project it into some finite dimensional space through a dimension reduction technique; e.g. using the widely popular functional principal component analysis, which will be discussed later in this chapter. While such projections were once essential, growth in functional literature is quickly diminishing such requirements. For example, see Cardot et al. (1999), Cardot et al. (2003), and Chen et al. (2011) for discussions on functional linear models and see Hall and Van Keilegom (2007) and Fremdt et al. (2013) for discussions on functional hypothesis testing. See Aue et al. (2018), Aue et al. (2020), and Wegner and Wendler (2024) for fully functional change point detection models.

It is important to note that because the data are observed discretely, FD are based on the *idea* that there exists some underlying process giving rise to the observed data. If the sampling rate of the discrete observations is dense enough, or the function is sufficiently smooth, reasonable approximation by a function is feasible. Linear combinations of basis functions is one method of fitting and presenting such functions.

Basis functions are computationally well adapted for storing and manipulating functional information. A basis function system of a real separable Hilbert space  $H$  is a set of known, usually orthogonal, functions  $\{f_1, f_2, \dots, g\}$  that span the space in the sense that for each  $X \in H$ ,

$$X = \sum_{i=1}^{\infty} c_i f_i;$$

where  $c \in \mathbb{R}$  and  $\ell \in \mathbb{N}$ . Often we consider finite dimensional subspaces of the form

$$K = \text{span}\{g_1, \dots, g_K\}$$

Hence by linear expansion, these basis functions can approximate arbitrarily well (for a large  $K$ ) any function  $X$  with a function in  $K$ ,

$$X(t) \approx \sum_{k=1}^K c_k g_k(t) = c^\top g(t);$$

where  $c = (c_1, \dots, c_K)^\top \in \mathbb{R}^K$  and  $g(t) = (g_1(t), \dots, g_K(t))^\top$ .

In practice, numerous options exist to select the basis functions. Common options include the B-spline basis and the Fourier basis. [Chui \(1992\)](#), [Silverman et al. \(2001\)](#), [Ramsay and Silverman \(2005\)](#), and [Hastie et al. \(2009\)](#) conduct more complete discussions on basis systems and various applications; see also [Friedman and Silverman \(1989\)](#) and [De Boor \(2001\)](#). Fitting data to the selected basis functions can be done by methods such as least squares or regularization, sometimes termed roughness penalty ([Ramsay and Silverman, 2005](#)).

Let  $X_i(t)$ ,  $i = 1, \dots, n$  be a realization of a square integrable process with mean  $\mu(t)$  and covariance function  $c(t; s)$  for  $t, s \in [0; 1]$ . [Mercer \(1909\)](#) states that  $c(\cdot; \cdot)$  has an orthogonal expansion

$$c(t; s) = \sum_{j=1}^{\infty} \lambda_j \phi_j(t) \phi_j(s); \quad t, s \in [0; 1];$$

where the  $\phi_j(\cdot)$ 's and  $\lambda_j$ 's are the respective eigenfunctions and eigenvalues of  $c(\cdot; \cdot)$ . Note that  $\lambda_1 \geq \lambda_2 \geq \dots$  and the eigenfunctions are orthonormal such that  $\int \phi_k \phi_l = 1$  when  $k = l$ , and 0 otherwise. In this case,

$$X_i(t) - \mu(t) = \sum_{j=1}^{\infty} u_{ij} \phi_j(t); \quad i = 1, \dots, n;$$

where  $u_{ij} = \int (X_i(t) - \mu(t)) \phi_j(t) dt$ . The term  $u_{ij}$  is called a functional principal component score and  $\phi_j$  is called the [functional principal component \(FPC\)](#). See discussion by [Kuhfeld \(1990\)](#) and [Dawkins \(1990\)](#) for synonymous terminology in the finite dimensional case. The scores are mean 0 with variance  $\lambda_j$  and are uncorrelated across  $j$ . Owing to [Karhunen \(1946\)](#) and [Loève \(1946\)](#), this decomposition is often referred to as the [Karhunen-Loève](#)

(KL) decomposition. The KL decomposition states that the eigenfunctions can be used to optimally (in  $L^2$ -sense) represent the process  $X$ .

Principal component analysis (PCA) is a data-adapted, dimension reduction technique employed throughout statistics and related disciplines. Early forms can be traced to Pearson (1901), but PCA was formalized by Hotelling (1933) who coined the term “principal component”. Originally computationally prohibitive, PCA has become a standard tool. See Wold et al. (1987), Abdi and Williams (2010), and Jolliffe and Cadima (2016) for article reviews and Mardia et al. (1979), Jolliffe (1986), or Gnanadesikan (1997) for textbook length discussions.

In FDA, functional principal component analysis (FPCA) has been widely used. This extension provides a way to look at the covariance structure rather than analyzing the variance-covariance function and can also be used as a dimension reduction method (Dauxois et al., 1982; Rice and Silverman, 1991; Silverman, 1996; Jacques and Preda, 2014). See Ramsay and Silverman (2005), Horváth and Kokoszka (2012), Jolliffe and Cadima (2016), and Kokoszka and Reimherr (2017). FPCA is used in areas such as data exploration (Hyndman and Shang, 2010), modeling (Ramsay and Dalzell, 1991; Ramsay and Silverman, 2005), forecasting (Hyndman and Ullah, 2007; Aue et al., 2015; Shang, 2017), and change point detection (Aue et al., 2009).

Intuitively, FPCA iteratively projects infinite-dimensional sample curves onto directions that contain the most variability within the data. Consequently, the leading FPCs can capture a significant portion of the “information” from the data, achieving dimension reduction. In practice, the first  $K$  FPCs explain the majority of the variability in the function. Hence, a reasonable approximation can be made by truncating the infinite sum such that

$$X_i(t) \approx \sum_{j=1}^K \hat{\alpha}_{ij} \hat{\psi}_j(t); \quad i = 1; \dots; n;$$

The estimates for the components and scores can be obtained by solving

$$\int_0^1 \hat{\psi}(t; s) \hat{\psi}_j(s) ds = \hat{\alpha}_j \hat{\psi}_j(t); \quad j = 1; 2; \dots; \quad t \in [0; 1];$$

Although we note that the components are only unique up to sign. A common approach to select the value for  $K$  is to use total variance explained (TVE),

$$TVE(K) = \frac{\sum_{j=1}^K \hat{\alpha}_j^2}{\sum_{j=1}^{\infty} \hat{\alpha}_j^2};$$

where  $\hat{\lambda}_i$ 's are the empirical eigenvalues. Other selection methods include information criteria (Shibata, 1981; Yao, 2007), resampling (Hall and Vial, 2006; Bathia et al., 2010), and cross-validation (Leng and Müller, 2006).

## 1.4 Change point analysis

Change point analysis is generally concerned with identifying significant changes in the distribution of some ordered data. For many problems, change point detection is the primary goal. For other problems, detection of change points is secondary and used to strengthen the plausibility of the primary model assumptions. The broad applications of change point analysis have resulted in continued growth in the field with entire monographs devoted to it, such as Csörgő and Horváth (1997), Chen and Gupta (2012), and Horváth and Rice (2024). Although traditional methods in change point analysis are concerned with scalar observations observed over time, methods have been adapted for objects such as vectors and functions over time or space.

Approaches in change point analysis fall into two general categories: sequential or online methods and retrospective or offline methods (Aminikhanghahi and Cook, 2017). The sequential methodology is generally dated to have been formalized by Page (1954) and Page (1955). The methodology aims to detect change points as quickly as possible in an incoming data stream; see also Barnard (1959) and Shiryaev (1963). The retrospective methodology was formalized by Page (1957) and Picard (1985). It is concerned with detecting changes in data that have been wholly observed. Recent advances have also entwined the frameworks and some methods can address both problems. This thesis is concerned with retrospective detection.

The fundamental hypothesis test for change point analysis is called the **at most one change (AMOC)** hypothesis test. Consider a sequence of  $n$  ordered random variables  $X_1, \dots, X_n$  with probability distributions  $P_1, \dots, P_n$ . The problem of detecting changes in the sequence of probability functions can be framed as a hypothesis test of the null hypothesis

$$H_0 : P_1 = \dots = P_n$$

versus the **AMOC** alternative

$$H_A : P_1 = \dots = P_k \notin P_{k+1} = \dots = P_n$$

with the unknown location  $k = 1; \dots; n$ . When the general distribution of the observations has a parametric form, this hypothesis can be defined in terms of the parameters. Let the general distribution function be denoted  $P$  and let each probability function  $P_i = P(j_i)$  for parameter vector  $\theta_i$ . The null hypothesis for no change in the parameters is defined as

$$H_0: \theta_1 = \dots = \theta_n$$

and is often compared against the alternative,

$$H_A: \theta_1 = \dots = \theta_k \neq \theta_{k+1} = \dots = \theta_n$$

for the unknown location  $k = 1; \dots; n$ .

Many change point detection methods exclusively focus on potential mean change point detection and assume other distributional properties remain unchanged, e.g. see [Aue and Horváth \(2013\)](#), [Aue et al. \(2018\)](#), and [Wegner and Wendler \(2024\)](#). There are relatively fewer methods to consider changes in other distributional properties, such as variance ([Gombay et al., 1996](#)) or covariance ([Aue et al., 2009](#); [Avanesov and Buzun, 2018](#); [Wang et al., 2021](#)). [Brodsky and Darkhovsky \(2000\)](#) show that under the Bernoulli shift framework from section 1.2, any distributional change can be considered as a change in the mean values of some new sequence constructed from the initial one.

Methods to test the AMOC hypothesis come in different forms. Often a first approach is using penalized least squares to develop a test statistic. There are also many test statistics based on the likelihood ratio test. Yet perhaps the most popular approach is to develop test statistics using the cumulative sum (CUSUM) process ([Page, 1954](#)). This process is shown to be identical to the likelihood ratio test in the univariate setting ([Wang et al., 2020](#)). The CUSUM process is constructed based on partial sums of the observations  $X_1; \dots; X_n$ . The CUSUM process, sometimes called the centered CUSUM process, to detect mean changes can be defined as

$$Z_n(x) = \frac{1}{n^{1/2}} \left( \sum_{i=1}^{bnc} X_i - \frac{bnc}{n} \sum_{i=1}^n X_i \right);$$

where  $x \in (0; 1)$  and  $bnc$  indicates the integer part of  $yx$ . Under the null,  $Z_n(x)$  can be shown to converge to a Brownian bridge ([Billingsley, 1968](#)). If a change occurs at location  $bnc$ , the value of  $Z_n(x)$  would be large. Since  $w(x) = [x(1-x)]^{1/2}$  is related to the variance of  $Z_n(x)$ ,  $Z_n(x)/w(x)$  is sometimes called the standardized CUSUM process and this process

has improved power, particularly near the endpoints, when compared to  $Z_n(x)$ . Thus one potential test statistic is defined as

$$M_n = \sup_{x \in (0,1)} \frac{Z_n(x)}{[x(1-x)]^{1/2}};$$

though many other test statistics exist. Often the limiting distributions of such statistics, with proper normalization, are studied using Gaussian approximations as introduced in [Darling and Erdős \(1956\)](#). The monograph by [Horváth and Rice \(2024\)](#) derive important asymptotic properties for the [CUSUM](#) process.

Although presented under the [AMOC](#) hypothesis, the hypotheses and related test statistics can be extended to multiple change point detection. Extensions from single to multiple change point detection take two forms: local or global searches. Local searches are greedy and sequentially select what appears to be the most likely change point. Global searches attempt to minimize some global risk formulation with a penalty term to select the number and placement of the change points.

The first and perhaps most widely used extension method is [binary segmentation \(BS\)](#) ([Scott and Knott, 1974](#); [Vostrikova, 1981](#)). [BS](#) is a forward-selection procedure that finds all change points through recursively segmenting the data at each detected change point until no additional change points are detected. Hence it is classified as a local search method. It has been shown to be consistent, yet not optimal, based on the standard [CUSUM](#) process in the case of independent, homoscedastic (scalar) Gaussian sequences ([Venkatraman, 1992](#)). However due to sequentially searching each segment for individual change points, [BS](#) can have poor performance if the change points are too close together. Such shortcomings have been mitigated in extensions of [BS](#) such as circular binary segmentation ([Olshen et al., 2004](#)) and wild binary segmentation ([Fryzlewicz, 2014](#); [Wang et al., 2020](#)).

A popular global search method is the [pruned exact linear time \(PELT\)](#) algorithm ([Killick et al., 2012](#)). [PELT](#) penalizes the inclusion of each additional change point by adding a penalty value to the cost of each segment and finding the segmentation whose penalized cost is smallest among all possible segmentations. Further discussion on extension methods, e.g specific penalty metric selections or methods such as the group fused lasso is left to other works ([Yao, 1988](#); [Pan and Chen, 2006](#); [Zhang and Siegmund, 2007](#); [Wu, 2008](#); [Bleakley and Vert, 2011](#); [Ciuperca, 2011, 2012](#)).

## 1.5 Spatial statistics

Spatial statistics is concerned with data indexed with some geometric or spatial dimension. [Cressie \(1993\)](#) considers there to be three main branches of spatial statistics: geostatistical data, lattice data, and point patterns. Geostatistical data refers to data comprised of multiple quantities in a region of interest. Lattice data refers to data measured in a region rather than exact locations. Point patterns refers to study of spatial locations of an event of interest. Geostatistical data and point patterns differ in that the former treats locations as an explanatory variable and the values as a response while the latter treats both as responses. Our interests in this work are related to point patterns. Point patterns, or spatial processes, handle dependence from all spatial-directions and this dependence generally weakens as the locations become more dispersed. The locations do not need to be regularly located.

Early work on spatial processes aggregated data into counts per area and compared the number of observations to a Poisson distribution ([Student, 1907](#)). Later procedures neutralized, but did not remove, the effects of spatial correlation ([Fisher, 1935](#)). Eventually models to address such spatial phenomena began to appear ([Whittle, 1954](#)). However, development of the field accelerated with the presentation of Ripley's  $K$  function ([Bartlett, 1964](#); [Ripley, 1976, 1977](#)).

A spatial process can generally be written as a stochastic process of the form

$$fZ(s) : s \in D; g;$$

where both  $Z(\cdot)$  and  $D$  are random. A random set, such as  $D$ , is a measurable mapping from a probability space onto a measure space of (usually closed) subsets of  $\mathbb{R}^d$ . Although,  $D$  is contained in any  $d$ -dimensional Euclidean space, applications are usually for  $d = 2$  or  $d = 3$ . This process can be expanded by adding additional properties, such as size, creating a marked point process,

$$fZ(s; m) : s \in D; m \in \mathcal{M}; g;$$

where  $\mathcal{M}$  is a space containing the marking attributes. Time can also be added to models to create space-time models.

Point patterns arise when the important variable is the location of some “agents”. Generally the first question for such a pattern is whether it exhibits [complete spatial randomness \(CSR\)](#) or interactions such as clustering.

[CSR](#) is the “white noise” of spatial point processes, and characterizes the absence of structure in the data. Hence, it is often the null hypothesis in statistical tests. Since

spatial processes were originally aggregated into counts, early tests for CSR compared departures from a Poisson distribution (Neyman, 1939). Consequently, the term CSR can be synonymous with a (homogeneous) Poisson process. A Poisson process  $N$  with mean  $\lambda(R)$  for region  $R$  is described as

$$P(N(R) = n) = \frac{\lambda(R)^n e^{-\lambda(R)}}{n!}; \quad n = 0; 1; \dots$$

The process is a homogeneous Poisson process if  $\lambda(R) = \lambda |R|$  for some  $\lambda > 0$ . Such a process states that conditional on the number of events  $n$ , the events  $s_1; \dots; s_n$  satisfy

$$P(s_1 \in B_1; \dots; s_n \in B_n) = \prod_{i=1}^n \frac{|B_i|}{|R|}; \quad B_1; \dots; B_n \subset R;$$

where  $|B| = \int_B ds$ . This means events are equally likely to occur anywhere in  $R$  and do not interact.

There are many metrics to measure the degree that a spatial process exhibits CSR. Originally conceptualized in Bartlett (1964) and formalized in Ripley (1976),  $K$  functions capture spatial dependence of the data using all interpoint distances.  $K$  functions have many benefits over other measures, such as nearest neighbor: they summarize point patterns over wider ranges since they account for all possible pairs, they have high power detecting both clustering and regularization with the same test, and they are independent of the study area shape (Ripley, 1977; Diggle, 1979).

For a spatial process with  $n$  events, or agents, on region  $R$ , the  $K$  function is defined

$$K(r) = \frac{1}{|R|} E(\text{Number of (extra) agents within distance } r);$$

where  $|R| = \frac{n}{\lambda}$ . Under the assumption of CSR in  $\mathbb{R}^2$ ,  $K(r) = \frac{1}{2} r^2$ . Regularity tends to reduce  $K(r)$  while clustering tends to increase it.

The classic  $K$  function definition can be expanded to cross-over  $K$  functions of marked point processes by considering only the distances to agents of type  $j$  from agents of type  $i$ ,

$$K_{ij}(r) = \frac{1}{|R|} E(\text{Number of (extra) agents of type } j \text{ within distance } r \text{ of agents of type } i);$$

See Hywood et al. (2021) and VanderDoes et al. (2024). Derivatives of  $K$  functions also exist (Besag, 1977).

Another consideration of point processes is the potential truncation error incurred due to the edges of observed region  $R$ . Test statistics, such as the  $K$  function, do not account for the agents outside  $R$ . Thus, agents near the edge of the region (spuriously) seem to be further from other agents. Various edge-effect correlation methods exist, such as toroidal edge corrections (Yamada and Rogerson, 2003).

## 1.6 Outline

The work contained in this thesis is related to the fields of time series, functional data, and change point analysis. The presentation of the work in the remainder of this thesis is organized as follows.

Changes for FTS are considered in chapter 2. A novel approach for detecting distributional changes in FTS based on empirical characteristic functionals is proposed. Several test statistics and methods for thresholding the statistics are discussed. Theoretical and practical considerations for tuning parameters are also explored. Applications on continuous electricity and high-frequency asset prices are given. This work has been published in Horváth et al. (2025).

Chapter 3 also investigates distributional changes in FTS. Graph-based change point detection is investigated. Several test statistics are discussed. Simulations to compare statistics and determine recommendations for tuning parameters are conducted. The graph-based methods are used on multi-year pedestrian counts, high-resolution stock returns, and continuous electricity prices.

The R package fChange is the focus of chapter 4. The package offers numerous methods for analyzing FTS with a particular emphasis on functional change point detection and supporting methodology. Details of the package are discussed and examples are given in a variety of scenarios, e.g. visualizations, tests for stationarity and white noise, simulations, and change point detection methods. Continuous electricity prices, cancer mortality, and long-term treasury rates are used for demonstration of the proposed functions.

FDA is used in chapter 5 to model point patterns related to cancer data. Data-rich images are summarized using cross-over  $K$  functions. The resultant functions, and any meta-data, are modeled using a functional random forest model. Modifications to mitigate overfitting and classify significant variables are discussed. Extensive simulations provide context for parameter choices. Real lung cancer and breast cancer point patterns are examined. This work has been published in VanderDoes et al. (2024).

Scalar data with stationary and nonstationary regimes is explored in chapter 6. A maximum likelihood-based change point detection method for [RCA\(1\)](#) models is derived to determine points of regime change. Theoretical and practical considerations are explored. COVID-19 hospitalisations and housing prices indices are studied under the [RCA\(1\)](#) model with change points. This work has been published in [Horváth et al. \(2024\)](#).

The thesis is concluded in chapter 7. Discussions on the preceding chapters are conducted. Potential avenues for future research and concluding remarks are also given.

Additional details for several chapters can be found in the accompanying appendices.

## Chapter 2

# Change Point Analysis for Functional Data Using Empirical Characteristic Functionals

### 2.1 Introduction

It is natural to consider general change point methods that are powerful against arbitrary changes in the underlying distribution of the series under study. For scalar time series, such methods have been developed based on sequentially estimated empirical processes, and are related to rank-based methods for change point analysis; see e.g. [Gombay and Hušková \(1998\)](#) and section 2 of [Horváth and Rice \(2014\)](#). A popular framework to construct similar methods for multivariate data is to employ empirical characteristic functions. To our knowledge the first paper to consider such an approach was [Hušková and Meintanis \(2006\)](#), who considered empirical characteristic function based change point detection for serially independent, scalar sequences. These results were generalized to multivariate and certain classes of serially dependent time series in [Hušková and Meintanis \(2008\)](#), [Steland and Rafajłowicz \(2014\)](#), and [Hlávka et al. \(2017\)](#). The idea was further popularized and extended to serially independent multivariate data in [Matteson and James \(2014\)](#) using “Energy distances”, which are specialized metrics between empirical characteristic functions allowing for efficient computation; see [Székely and Rizzo \(2013\)](#). The recent paper [Dette and Tang \(2024\)](#) studies energy distances in the setting of infinite dimensional Hilbert spaces.

In this chapter, we develop general change point test statistics for [functional time series](#)

(FTS) based on [cumulative sum \(CUSUM\)](#) processes of empirical characteristic functionals defined over  $L^2[0;1]$ . A key consideration in constructing such statistics is the metric used to evaluate the magnitude of such empirical characteristic processes. We consider integrated metrics against probability measures on  $L^2[0;1]$ , as a functional generalization of the approach introduced in [Hušková and Meintanis \(2006\)](#). The limiting null-distribution of the proposed test statistics is established under mild conditions allowing for low-order moments and generic serial dependence in the data, and we also show that they may be used to consistently detect and localize change points in the distribution of functional data. In order to approximate the no-change null-distribution of functionals of such [CUSUM](#) processes, we consider block-permutation methods, Welch-style approximations, and Gaussian process simulation techniques, and we study their consistency properties. We show in numerous simulation experiments that these methods can be tuned to have approximately nominal empirical size even for strongly serially dependent [FTS](#), and are capable of detecting general changes in the distribution of such a series. The proposed methods are studied in two data applications to Spanish electricity price data, and an electronically traded fund (ETF) tracking the US S&P 500 stock market index.

The remainder of this chapter is organized as follows. Several characteristic function-based test statistics and estimators are presented in [section 2.2](#), along with results establishing their asymptotic properties. In [section 2.3](#), we introduce and study several computational methods to estimate critical values needed to carry out significance tests based on the proposed statistics. These approaches are investigated via simulations in [section 2.4](#), along with comparisons to other change point methods for [FTS](#). Application of the method to the electricity and return curves is discussed in [section 2.5](#). [section 2.6](#) concludes. All proofs and additional simulation evidence are contained within [Appendices A.1-A.3](#).

## 2.2 Main results

The following notation is used below. We consider a [functional time series](#)  $fX_i g_{i2Z} := fX_i(t) : i \in Z; t \in [0;1]g$ , from which we assume that we have observed a stretch of length  $n$ ,  $X_1; \dots; X_n$ . Here each variable  $X_i$  is viewed as a stochastic process whose sample paths lie in the Hilbert space  $H = L^2[0;1]$  of measurable, real valued, square integrable functions equipped with the inner product

$$\langle f; g \rangle = \int_0^1 f(t)g(t) dt;$$

and corresponding norm  $\| \cdot \|_k = \sqrt{h} \| \cdot \|_i$ . Although the methods and results below hold for any separable Hilbert space  $H$ , we take  $H = L^2[0; 1]$ , since this covers the applications that we wish to consider.

Let the distribution of each process  $X_j$  viewed as an element of  $L^2[0; 1]$  be denoted by  $F_j$ , i.e.  $F_j$  is the probability measure defined on the standard Borel  $\sigma$ -algebra  $\mathcal{B}(L^2[0; 1])$  of subsets of  $L^2[0; 1]$ , defined by  $F_j(A) = P(X_j \in A)$  for each  $A \in \mathcal{B}(L^2[0; 1])$ .

We are interested in detecting and estimating change points in the sequence of distributions  $F_j$ . In the framework of [at most one change \(AMOC\)](#), the problem of detecting such a change point can be framed as a hypothesis test in which we wish to test the null hypothesis

$$H_0 : F_1 = \dots = F_n \tag{2.1}$$

versus the alternative

$$H_A : \text{there exists a } k = bn + c \text{ so that } F_1 = \dots = F_k \neq F_{k+1} = \dots = F_n; \tag{2.2}$$

for a break fraction  $b \in (0; 1)$ .

In the spirit of [Hušková and Meintanis \(2006\)](#), we propose using empirical characteristic functions to carry out this test. The characteristic function of a random element of  $L^2[0; 1]$  is defined as follows:

*Definition 1. For a random element  $X \in L^2[0; 1]$ , the characteristic function or characteristic functional of  $X$  is*

$$f_X(v) = E \exp(\mathbf{i}hX; v); \quad v \in L^2[0; 1];$$

where  $\mathbf{i}$  is the imaginary unit. See [Bosq \(2000\)](#), pg. 32.

As with scalar and vector valued random variables, the characteristic function of  $X_j$  determines the distribution  $F_j$ , so that

$$F_1 = \dots = F_n \text{ if and only if } f_{X_1} = \dots = f_{X_n}; \tag{2.3}$$

Our approach is based on partial sample estimates of  $f_X$ . Let

$$\hat{f}_n(v; x) = \frac{1}{n} \sum_{j=1}^{bn+c} \exp(\mathbf{i}hX_j; v) \quad v \in L^2[0; 1]; \quad x \in [0; 1];$$

When  $H_A$  of (2.2) holds, one expects due to Eq (2.3) a large difference between  $\hat{f}_n(v; x)$  and  $x\hat{f}_n(v; 1)$  for some  $v \in L^2[0; 1]$  when  $x$  takes the value near  $\cdot$ . A CUSUM statistic, as introduced originally in Page (1954), can be used to detect such a difference. Let

$$Z_n(v; x) = \rho_{\bar{n}} \left( \hat{f}_n(v; x) - \frac{bnxc}{n} \hat{f}_n(v; 1) \right) :$$

In order to evaluate the magnitude of this CUSUM process for a fixed partial sample parameter  $x$ , we view it as an element of the function space  $L^2 = L^2(L^2[0; 1]; B(L^2[0; 1]); Q)$ , where  $Q$  is a user-specified probability measure on  $(L^2[0; 1]; B(L^2[0; 1]))$ . Using the norm on  $L^2$  to determine the size of  $Z_n(\cdot; x)$ , two natural test statistics are

$$T_n = \int_0^1 \int_H |jZ_n(v; x)|^2 dQ(v) dx ; \tag{2.4}$$

and

$$M_n = \sup_{x \in [0; 1]} \int_H |jZ_n(v; x)|^2 dQ(v) ; \tag{2.5}$$

where  $|j|$  is the complex modulus. The integrals with respect to the measure  $Q$  over  $H$  can be understood as abstract Lebesgue integrals of non-negative, measurable functions; see Definition 5.5 of Bruckner et al. (1997).

It is worth mentioning at this stage how such a measure  $Q$  might be selected, and how the integrals defining  $T_n$  and  $M_n$  are evaluated in practice. As discussed in Henze and Jiménez-Gamero (2021), a natural choice for  $Q$  is the distribution of a stochastic process with sample paths in  $L^2[0; 1]$  that can be simulated easily, and whose measure is supported by each ball in  $L^2[0; 1]$ . The distribution of standard Brownian motion satisfies both criteria, and we generally use it below. Specifically, suppose  $fW(y); y \in [0; 1]g$  is a standard Brownian motion (standard Wiener process) that is independent of  $fX_i g_{i \in \mathbb{Z}}$ . We set, for  $A \in B(L^2[0; 1])$ ,  $Q(A) = P(W \in A)$ . It follows as in Proposition 4.9 in Breiman (1968) that for each  $x \in [0; 1]$ ,

$$\int_H |jZ_n(v; x)|^2 dQ(v) = E \left[ |jZ_n(W; x)|^2 \mid X_1; \dots; X_n \right] ; \tag{2.6}$$

One approach to estimate this expectation is to approximate the inner-products  $\langle hX_j; Wi \rangle$  defining  $Z_n(W; x)$  with Riemann sums using  $D$  terms, and evaluate the expectation as a limit of  $D$ -dimensional integrals against the joint density of  $W(1=D); W(2=D); \dots; W(1)$ ; see Eq (A.1) in the Appendix for details.

Alternatively, in view of Eq (2.6) we may estimate the integrals in definitions Eqs (2.4) and (2.5) using standard Monte-Carlo integration. If  $W_1; \dots; W_B$  are **independent and identically distributed (IID)** stochastic processes with distribution  $Q$ , e.g. **IID** standard Brownian motions, independent of  $fX_i g_{i2Z}$ , then we may estimate using a large  $B$

$$T_n = \frac{1}{B} \sum_{i=1}^B \int_0^1 jZ_n(W_i; x)j^2 dx; \text{ and } M_n = \sup_{x \in [0;1]} \frac{1}{B} \sum_{i=1}^B jZ_n(W_i; x)j^2; \quad (2.7)$$

In the multivariate setting, it is standard to use a multivariate normal distribution for the measure  $Q$  and similarly apply Monte-Carlo integration; see e.g. [Hlávka et al. \(2017\)](#). Other choices of  $Q$  in the multivariate setting lead to “energy statistics” with favorable computational properties that do not require Monte-Carlo integration, as considered in [Matteson and James \(2014\)](#). It is an interesting question, that we leave as open, whether other choices of  $Q$  could be made in this setting that have similar computational benefits and can be applied to change point analysis. See also [Dette and Tang \(2024\)](#).

## 2.2.1 Asymptotic results under $H_0$

We now aim to describe the asymptotic properties of the statistics  $T_n$  and  $M_n$  under  $H_0$ . In order to do so, we assume in addition to  $H_0$  that  $fX_i g_{i2Z}$  evolves as a strictly stationary and weakly-dependent time series as characterized by a modified version of the  $L^p$ - $m$ -approximability criterion, which was introduced in [Hörmann and Kokoszka \(2010\)](#).

Assumption 2.2.1. *The FTS  $fX_i g_{i2Z}$  satisfies*

(a) *there is a measurable function  $g: S^N \rightarrow L^2[0;1]$ , where  $S$  is a measurable space, and IID innovations  $f_i g_{i2Z}$  taking values in  $S$  such that  $X_i = g(f_i; f_{i-1}; \dots)$  for  $i \in \mathbb{Z}$ ;*

(b) *for some  $p > 0$  and  $\delta > 0$ ,*

$$\sum_{m=0}^{\infty} (E[kX_i - X_{i;m}]^p)^{1/(2+\delta)} < \infty; \quad (2.8)$$

where  $X_{i;m} = g(f_i; \dots; f_{i-m+1}; f_{i-m}; f_{i-m-1}; \dots)$  with  $f_i g_{i2Z}$  being independent copies of  $f_0$ , independent of  $f_i g_{i2Z}$ .

Series satisfying Assumption 2.2.1(a) are often called “Bernoulli-shifts”, and are evidently strictly stationary. Thus Assumption 2.2.1(a) implies  $H_0$ . Assumption 2.2.1(b)

$fX_i g_{i2Z}$  is a “weak dependence” condition in the sense that it requires the process  $X_i$  to have a decaying dependence on past innovations. Typical  $L^p$ - $m$ -approximability assumes that  $E[kX_i k^\rho] < 1$ , and the sequence  $v_m = (E[kX_i - X_{i,m} k^\rho])^{1-\rho}$  is summable. Therefore  $L^{2+}$ - $m$ -approximability, which is a standard condition under which the FTS  $fX_i g_{i2Z}$  satisfies a functional central limit theorem, implies Assumption 2.2.1. The main appeal of this new formulation of condition (b) is that it only assumes  $\rho$  moments of the sequence  $kX_i k$ , with  $\rho$  being arbitrary. Since our goal is to detect general changes in the distribution, ideally we would not need to make any moment assumptions on  $kX_i k$ . This is nearly achieved by Assumption 2.2.1, but the cost is a stronger decay condition on the coefficients  $v_m$  for small  $\rho$ . For instance, if  $v_m \leq C_y m^{-\alpha}$ , for a positive  $C_y$ , then condition (b) requires that  $\alpha > (2 + \rho)/\rho$ , which is more restrictive than  $L^{2+}$ - $m$ -approximability for  $\rho < 2 + \epsilon$ . Since for most examples of interest, e.g. functional versions of autoregressive or GARCH processes,  $v_m \leq C_y m^{-\alpha}$  for a constant  $\alpha$  satisfying  $0 < \alpha < 1$ , we do not view this new condition as being overly restrictive. This assumption could also be replaced by similarly adapted mixing and near-epoch-dependence assumptions featured in [Dehling et al. \(2015\)](#), [Sharipov et al. \(2016\)](#), and [Sharipov and Wendler \(2020\)](#).

Theorem 2.2.1. *Suppose  $H_0$  holds, and that the FTS  $fX_i g_{i2Z}$  satisfies Assumption 2.2.1. If additionally*

$$\int_H k v k^\rho dQ(v) < 1; \quad (2.9)$$

*then there exists a mean zero, complex valued Gaussian process  $\tilde{X} : L^2[0;1] \times [0;1] \rightarrow \mathbb{C}$  such that*

$$T_n \stackrel{P}{\rightarrow} \int_0^1 \int_H j(v; x) \tilde{X}(v; x) dQ(v) dx; \text{ and } M_n \stackrel{P}{\rightarrow} \sup_{x \in [0;1]} \int_H j(v; x) \tilde{X}(v; x) dQ(v); \quad (2.10)$$

*Let  $G(x; x^\theta) = \min(x; x^\theta) - x x^\theta$ . The complex Gaussian process  $\tilde{X}$ , for  $(v; x) \in L^2[0;1] \times [0;1]$ , may be written as  $\tilde{X}(v; x) = \tilde{X}_1(v; x) + \mathbf{i} \tilde{X}_2(v; x)$ , where  $\tilde{X}_1$  and  $\tilde{X}_2$  are real-valued, mean zero, jointly Gaussian processes with covariances*

$$\begin{aligned} \text{Cov}(\tilde{X}_1(v; x); \tilde{X}_1(v^\theta; x^\theta)) &= \sum_{i=1}^1 \text{Cov}(\cos(hv; X_0 i); \cos(hv^\theta; X_0 i)) G(x; x^\theta) \\ &:= C_{\text{Re}}(V; V^\theta) G(x; x^\theta); \end{aligned} \quad (2.11)$$

$$\begin{aligned} \text{Cov}(\sin(hv; X_0); \sin(hv^\beta; X_\beta)) &= \sum_{i=1}^1 \text{Cov}(\sin(hv; X_{0i}); \sin(hv^\beta; X_{\beta i})) G(x; x^\beta) \\ &:= C_{Im}(v; v^\beta) G(x; x^\beta); \end{aligned} \quad (2.12)$$

and

$$\begin{aligned} \text{Cov}(\cos(hv; X_0); \sin(hv^\beta; X_\beta)) &= \sum_{i=1}^1 \text{Cov}(\cos(hv; X_{0i}); \sin(hv^\beta; X_{\beta i})) G(x; x^\beta) \\ &:= C_{Relm}(v; v^\beta) G(x; x^\beta); \end{aligned} \quad (2.13)$$

The proof of Theorem 2.2.1 is given in Appendix A.1. Theorem 2.2.1 may be viewed as an extension of the main asymptotic results of Hušková and Meintanis (2006) and Hlávka et al. (2017) to the setting of serially dependent functional data objects.

We note that the moment condition Eq (2.9) is satisfied for all  $\rho > 0$  when  $Q$  is the distribution of Brownian motion.

## 2.2.2 Consistency under $H_A$ and estimating $k$

In this subsection, we aim to develop conditions under which both  $T_n \xrightarrow{P} 1$  and  $M_n \xrightarrow{P} 1$  when  $H_A$  of (2.2) holds. Moreover, a natural estimator of  $k$  is

$$\hat{k}_n = \underset{k \in \{1, \dots, ng\}}{\text{sargmax}} \int_H \left| Z_n \left( v; \frac{k}{n} \right) \right|^2 dQ(v); \quad (2.14)$$

where  $\text{sargmax}_{k \in \{1, \dots, ng\}} f(k) = \min \{k : f(k) = \max_{1 \leq j \leq ng} f(j)\}$  is the smallest maximal argument of  $f$ , and we also investigate its consistency. Towards this, we make the following assumption:

Assumption 2.2.2. (a) With  $k$  defined in  $H_A$ , there exist measurable functions  $g_j: S^N \rightarrow L^2[0; 1]$ ,  $j = 1; 2$ , where  $S$  is a measurable space, and IID functional innovations  $f_i, g_{i2Z}$  taking values in  $S$  such that  $X_i = g_1(i; i_1; \dots; i_{i-1}; \dots)$  for  $i \leq k$ , and  $X_i = g_2(i; i_1; \dots; i_{i-1}; \dots)$ ; if  $i > k + 1$ :

(b) For each  $n \in \mathbb{N}$ , and  $i \in \{1; \dots; ng\}$ , there exist constants  $C_z > 0$ , not depending on  $n$ , and  $\gamma > 1$  so that for all  $m \geq 0$ ,  $(E[kX_i - X_{i,m}k^p])^{1=(2+\gamma)} \leq C_z m^{-\gamma}$ ; where, with  $f_i, g_{i2Z}$  being independent copies of  $f_0$ ,  $X_{i,m} = g_1(i; \dots; i_{m+1}; i_m; i_{m-1}; \dots)$  if  $i \leq k$ , and  $X_{i,m} = g_2(i; \dots; i_{m+1}; i_m; i_{m-1}; \dots)$ , if  $i > k + 1$ :

Assumption 2.2.2(b) is similar to Assumption 2.2.1(b), and in effect assumes that the series before and after the change point are each weakly dependent with similar rates of decay of the coefficients in Eq (2.8). Under this condition we have the following consistency result, whose proof is given in Appendix A.1. Let  $f_{X_0}$  and  $f_{X_A}$  denote the common characteristic functions of the observations before and after the change point  $k$ , respectively. We further define  $z : L^2[0; 1] \times [0; 1] \rightarrow \mathbb{C}$  as

$$z(v; x) = [x(1 - x) \mathbf{1}_{f_0(x) > g} + (1 - x) \mathbf{1}_{f_0(x) \leq g}] (f_{X_0}(v) - f_{X_A}(v)) :$$

Theorem 2.2.2. *If  $H_A$  of (2.2), Assumption 2.2.2, and Eq (2.9) hold, then*

$$\frac{T_n}{n} \xrightarrow{P} \int_0^1 \int_H |z(v; x)|^2 dQ(v) dx; \quad \text{and} \quad \frac{M_n}{n} \xrightarrow{P} \sup_{x \in [0; 1]} \int_H |z(v; x)|^2 dQ(v) :$$

As such, if

$$\int_H |f_{X_0}(v) - f_{X_A}(v)|^2 dQ(v) > 0; \tag{2.15}$$

then  $T_n \xrightarrow{P} 1$  and  $M_n \xrightarrow{P} 1$ . Moreover, under the same conditions

$$\hat{k}_n - k = O_P(1); \tag{2.16}$$

We note that the consistency result in Eq (2.16) is rate optimal under the AMOC alternative.

We show the following corollary in Appendix A.1.

Corollary 2.2.1. *If  $Q$  is the distribution of standard Brownian motion, condition (2.15) is always satisfied whenever  $H_A$  holds.*

The proof of this result makes use of the continuity of  $v \mapsto E \exp(\mathbf{i}hv; X_i)$ , and the fact that the measure generated by Brownian motion on  $L^2[0; 1]$  is supported on each open ball of  $L^2[0; 1]$ . As such, taking  $Q$  to be the distribution of standard Brownian motion leads to the general consistency of  $T_n$  and  $M_n$  for testing  $H_0$  versus  $H_A$ .

## 2.3 Estimating the distributions appearing in Theorem 2.2.1

Theorem 2.2.1 suggests that we may reject  $H_0$  with asymptotic level calibrated to  $\alpha$  if either

$$T_n > \hat{T}(1 - \alpha); \quad \text{or} \quad M_n > \hat{M}(1 - \alpha);$$

where  $\hat{T}(q)$  and  $\hat{M}(q)$  are estimates of the  $q$ 'th quantile of the respective limiting distributions appearing in Eq (2.10). The problem of estimating these quantiles is somewhat more difficult than for analogous change point tests for the mean or covariance structure of functional data. This is primarily due to the fact that the covariance kernels in Eqs (2.11)–(2.13), and estimates of such kernels, cannot be easily empirically diagonalized since they are defined over an infinite dimensional space. We now describe three approaches to estimate these quantiles, which we investigated further in simulation experiments that are presented in section 2.4.

### 2.3.1 Gaussian process simulation

This technique makes use of the approximation in Eq (2.7). Given realizations of  $B$  Brownian motions  $(W_1; \dots; W_B)$  used in the Monte-Carlo estimates of the integrals defining  $T_n$  and  $M_n$ , we have, according to Theorem 2.2.1, that for large  $n$ ,

$$T_n \stackrel{D}{\approx} \frac{1}{B} \sum_{i=1}^B \int_0^1 j(W_i; x) f^2 dx \approx \frac{1}{BR} \sum_{i=1}^B \sum_{r=1}^R \left[ \frac{1}{2} \left( W_i; \frac{r}{R} \right) + \frac{1}{2} \left( W_i; \frac{r}{R} \right) \right]; \quad (2.17)$$

and

$$M_n \stackrel{D}{\approx} \sup_{x \in [0,1]} \frac{1}{B} \sum_{i=1}^B j(W_i; x) f^2 dx \approx \max_{r \in \{1, \dots, R\}} \frac{1}{B} \sum_{i=1}^B \left[ \frac{1}{2} \left( W_i; \frac{r}{R} \right) + \frac{1}{2} \left( W_i; \frac{r}{R} \right) \right]; \quad (2.18)$$

where  $\mathcal{W}_1$  and  $\mathcal{W}_2$  are jointly Gaussian processes with mean zero and covariance structure described in Eqs (2.11)–(2.13). We may then approximate the distributions appearing on the right in Eqs (2.17) and (2.18) by simulating the Gaussian vector

$$\text{vec} \left[ \left( \mathcal{W}_1 \left( W_i; \frac{r}{R} \right); \mathcal{W}_2 \left( W_j; \frac{k}{R} \right) \right)_{1 \leq i, j \leq B; 1 \leq r, k \leq R} \right] \in \mathbb{R}^{2B^2R^2}; \quad (2.19)$$

The covariance matrix of this vector is a simple function of  $C_{Re}$ ,  $C_{Im}$ , and  $C_{ReIm}$  in Eqs (2.11), (2.12) and (2.13), respectively. To estimate these functions, we propose using kernel-lag window long-run covariance estimators. For example, we estimate  $C_{Re}$  with

$$\hat{C}_{Re}(V; V^\lambda) = \sum_{i=1}^{n-1} K\left(\frac{\cdot}{h}\right) \hat{\gamma}_{Re}(V; V^\lambda); \quad (2.20)$$

Here  $K : \mathbb{R} \rightarrow \mathbb{R}$  is a kernel function such that  $K(0) = 1$ , and further is a positive, symmetric, Lipschitz function with bounded support,  $h = h(n)$  is a lag-window or bandwidth parameter satisfying  $1/h + h = o(1/n)$  as  $n \rightarrow \infty$ ,

$$\hat{\gamma}_{Re}(V; V^\lambda) = \begin{cases} \frac{1}{n} \sum_{j=i+1}^n (\cos(hv; X_j - i) - \bar{Re}_n(v)) (\cos(hv^\lambda; X_j) - \bar{Re}_n(v^\lambda)) & \geq 0 \\ \frac{1}{n} \sum_{j=1}^{n-j} (\cos(hv; X_j - i) - \bar{Re}_n(v)) (\cos(hv^\lambda; X_j + i) - \bar{Re}_n(v^\lambda)) & < 0; \end{cases}$$

and

$$\bar{Re}_n(v) = \frac{1}{n} \sum_{j=1}^n \cos(hv; X_j);$$

$C_{Im}$  and  $C_{ReIm}$  may be estimated similarly. Below we use the Bartlett kernel  $K(u) = \frac{1}{2} (1 - |u|)$ , and compare values of  $h \in [1, n^{1/3}]$ ;  $2n^{1/3} \leq h \leq n$ . In the case when  $h = 1$ , we hence estimate for instance  $C_{Re}$  with  $\hat{\gamma}_{Re}$ . This would be an ideal estimator if it were known that the series under study were serially independent. Alternate kernel functions include the truncated, Parzen, Tukey-Hanning, and Quadratic Spectral kernels, though many other options exist; see e.g. [White and Domowitz \(1984\)](#), [Newey and West \(1987\)](#), and [Andrews \(1991\)](#). We also refer to [Horváth et al. \(2016\)](#) and [Rice and Shang \(2017\)](#) concerning typical bandwidth selection methods in long-run covariance estimation with [FTS](#).

We then simulate the Gaussian vector in Eq (2.19)  $J$  times, where below we set  $J = 1000$ , and subsequently compute the  $J$  simulated statistics  $(T_n^{(j)}; M_n^{(j)}); j = 1, \dots, J$ . The null-critical values of  $T_n$  and  $M_n$  are then approximated as

$$\hat{\tau}(1 - \alpha) = \inf \left\{ q : \frac{1}{J} \sum_{j=1}^J 1(T_n^{(j)} \leq q) > 1 - \alpha \right\}; \quad (2.21)$$

and

$$\hat{\mathcal{M}}(1 - \alpha) = \inf \left\{ q : \frac{1}{J} \sum_{j=1}^J \mathbf{1}(\mathcal{M}_n^{(j)} \leq q) > 1 - \alpha \right\}; \quad (2.22)$$

We note that with these settings this simulation is somewhat computationally intensive owing to the large dimension,  $2B^2R^2$ , of the Gaussian vector, whose covariance matrix must be computationally diagonalized. Choosing values of  $B$  as small as possible, whilst still maintaining the same size and power for the procedure, is of interest. To this end, Appendix A.2 presents investigations into the choice of  $B$  and whether the value of  $R$  has much influence over the procedure. In summary, there seemed to be limited gains beyond even using quite low values for  $B$ , e.g.  $B = 10$ , and little change for increasing values of  $R$  beyond  $R = 20$ . Throughout the chapter we generally take  $B = 20$  and  $R = 50$ .

### 2.3.2 Welch-style approximation ( $T_n$ only)

Using Gaussian process simulation as described above, as well as the permutation methods introduced in the following subsection, can be computationally expensive and time consuming. When it comes to estimating the null-distribution of the statistic  $T_n$ , an alternative approach that is computationally efficient is to use a ‘‘Welch’’ or ‘‘Welch-Satterthwaite’’ approximation, as introduced originally in Satterthwaite (1946) and Welch (1947); see also Chapter 4 of Zhang (2014). Per Theorem 2.2.1, the null distribution  $T_n$  is approximately that of the squared norm of a Gaussian process. Such a random variable may be expressed as a weighted sum of independent  $\chi^2$  random variables with one degree of freedom, and it has been further observed that such weighted sums are often well approximated by rescaled  $\chi^2$  random variables. This suggests the approximation for the null-distribution of  $T_n$  for large  $n$  as

$$T_n \stackrel{D}{\approx} \int_0^1 \int_H j(v; x)^2 dQ(v) dx \stackrel{D}{\approx} \chi^2_{\nu};$$

where  $\nu \in \mathbb{N}$ ;  $\nu \geq 2(0; 1)$ . The scalar parameters  $\nu$  and  $\mu$  are chosen such that the first two moments of  $\chi^2_{\nu}$  coincide with the limiting distribution, i.e. so that

$$E \chi^2_{\nu} = E \int_0^1 \int_H j(v; x)^2 dQ(v) dx; \text{ and } \text{Var}(\chi^2_{\nu}) = \text{Var} \left( \int_0^1 \int_H j(v; x)^2 dQ(v) dx \right); \quad (2.23)$$

Following Proposition 5.10.16 of [Bogachev \(1998\)](#), some calculation yields that

$$\begin{aligned} m &= E \int_0^1 \int_H j(v; x) f^2 dQ(v) dx = \int_0^1 \int_H [C_{Re}(v; v) + C_{Im}(v; v)] G(x; x) dQ(v) dx \\ &= \frac{1}{6} \int_H [C_{Re}(v; v) + C_{Im}(v; v)] dQ(v); \end{aligned}$$

and

$$\begin{aligned} \hat{\sigma}^2 &= \text{Var} \left( \int_0^1 \int_H j(v; x) f^2 dQ(v) dx \right) \\ &= \frac{2}{90} \int_H \int_H [C_{Re}^2(v; v') + 2C_{ReIm}^2(v; v') + C_{Im}^2(v; v')] dQ(v) dQ(v'); \end{aligned}$$

By replacing  $C_{Re}$ ,  $C_{Im}$ , and  $C_{ReIm}$  with consistent estimators in the above equations and evaluating the integrals using Monte-Carlo integration as in Eq (2.17), we obtain estimates of  $\hat{m}$  and  $\hat{\sigma}^2$ . Solving the equations in (2.23) for  $\hat{\sigma}$  and  $\hat{m}$  suggests setting

$$\hat{\sigma} = \frac{\hat{\sigma}^2}{2\hat{m}}; \quad \text{and} \quad \hat{m} = \frac{2\hat{m}^2}{\hat{\sigma}^2};$$

The Welch-approximation of the null critical value of  $T_n$  is then given as

$$\hat{T}(1 - \alpha) = \inf \left\{ q : P(\hat{\sigma}^{-2}(\hat{m}) > q) = 1 - \alpha \right\};$$

### 2.3.3 Permutation

Permutation methods have also been successfully implemented in empirical characteristic function based change point analysis for both scalar and vector valued data in [Hušková and Meintanis \(2006\)](#) and [Matteson and James \(2014\)](#), respectively. Their implementation is simple, although somewhat computationally time consuming. For a user specified large number  $J$ , which we set to 1000 below, the observed series  $(X_1; \dots; X_n)$  is randomly permuted  $J$  times, producing permuted samples  $(X_1^{(j)}; \dots; X_n^{(j)}); j = 1; \dots; J$ . Let  $\pi: \{1; \dots; n\} \rightarrow \{1; \dots; n\}$  denote a permutation chosen uniformly at random among all permutations, independently of  $\{X_i; i=1, \dots, n\}$ . Denoting the test statistics computed from the  $j^{\text{th}}$  permuted sample  $T_n(\pi_j)$  and  $M_n(\pi_j)$ , we set the estimated null critical values for  $T_n$  and  $M_n$  as in Eq (2.21) and Eq (2.22), respectively. When the underlying FTS is serially independent, these critical values (for large  $J$ ) furnish a test with exact size  $\alpha$ . Using this quantile also leads to a consistent test in the AMOC alternative (2.2).

Theorem 2.3.1. Suppose that Eqs (2.2) and (2.9) hold, and further that  $X_1; \dots; X_k$  are independent with distribution  $F_0$ , and  $X_{k+1}; \dots; X_n$  are independent with distribution  $F_A$ . Let  $Y$  be a random element of  $L^2[0; 1]$  with distribution  $F_0 + (1 - \alpha)F_A$ . If  $\mathbf{X} = (X_1; \dots; X_n)$ , then for all  $y \in \mathbb{R}$

$$P(T_n(\alpha) \leq y | \mathbf{X}) \stackrel{P}{=} P\left(\int_0^1 \int_H j(v; x) f^2 dQ(v) dx \leq y\right);$$

and

$$P(M_n(\alpha) \leq y | \mathbf{X}) \stackrel{P}{=} P\left(\sup_{x \in [0; 1]} \int_H j(v; x) f^2 dQ(v) dx \leq y\right)$$

as  $n \rightarrow \infty$ , where  $\mathbf{W} = \mathbf{W}_1 + i \mathbf{W}_2$  is a mean zero complex Gaussian process with covariance as in Eqs (2.11)–(2.13), where  $C_{Re}$ ,  $C_{Im}$ , and  $C_{ReIm}$  take the form

$$C_{Re}^A(v; v^\flat) = Cov(\cos(hv; Y_i); \cos(hv^\flat; Y_i)); \quad C_{Im}^A(v; v^\flat) = Cov(\sin(hv; Y_i); \sin(hv^\flat; Y_i));$$

and

$$C_{ReIm}^A(v; v^\flat) = Cov(\cos(hv; Y_i); \sin(hv^\flat; Y_i));$$

Theorem 2.3.1 implies that for serially independent data under  $H_A$ , the permutation distribution of  $T_n$  and  $M_n$  will asymptotically coincide with the limiting null distributions described in Theorem 2.2.1 as if the data were IID and drawn from the distribution  $F_0 + (1 - \alpha)F_A$ . In view of Theorem 2.2.2, rejecting  $H_0$  based on the quantiles computed via permutation then leads to a consistent test.

In general though we expect the FTS under study to be serially dependent. In that case we consider using “block-permutation”, as discussed for example in Carlstein (1986) and Lahiri (1999). For a given block length  $h$ , which is a natural number that for convenience we assume evenly divides  $n$ , we partition the FTS into  $n/h$  blocks  $\mathbf{X}_k = (X_{kh-n+1}; \dots; X_{k(h+1)-n})$ . To construct pseudo-samples of the series that attempt to preserve the serial dependence structure of the original series, we permute each of the blocks so that if  $(j)$  is the  $j$ 'th permutation of  $1; \dots; n/h$ , then we set  $(X_1^{(j)}; \dots; X_n^{(j)}) = (\mathbf{X}_{(j)(1)}; \dots; \mathbf{X}_{(j)(n/h)})$ , i.e. the blocks are permuted and then pasted together. We then proceed as above. We consider the block sizes  $h \geq 1; dn^{1-3\epsilon}; d2n^{1-3\epsilon}g$  below.

### 2.3.4 Multiple change detection

We are often also interested in evaluating for the presence of multiple change points in the sequence, so that rather than the AMOC alternative of (2.2), we consider an  $m$  change

point alternative

$$H_A: \text{ there exists } k_i = \lfloor bn_i \rfloor \text{ with } i = 1, \dots, m, \text{ and } 0 < \tau_1 < \dots < \tau_m < 1, \text{ so that} \\ F_1 = \dots = F_{k_1} \neq F_{k_1+1} = \dots = F_{k_2} \neq \dots \neq F_{k_{m+1}} = \dots = F_n: \quad (2.24)$$

A standard approach to extend tests and estimation procedures built for the [AMOC](#) alternative to detect multiple change points is to employ [binary segmentation \(BS\)](#). BS starts by applying a single change point detection procedure to the observed series. If a change point is detected, its location is estimated, and the data is split into two sub-samples based on the estimated change point. This process is repeated on each sub-sample until a stopping criteria is satisfied.

Below, to estimate multiple change points we apply [BS](#) where we declare a change point has been detected if, for a user specified level  $\alpha$ , either  $T_n > \hat{T}(1 - \alpha)$  or  $M_n > \hat{M}(1 - \alpha)$ ; where  $\hat{T}(1 - \alpha)$  and  $\hat{M}(1 - \alpha)$  are estimates of the  $1 - \alpha$  quantile defined above. In order to estimate change points on a given sub-sample, we use the estimator  $\hat{k}_n$  in [Eq \(2.14\)](#).

## 2.4 Simulation study

In this section, we present the results of several simulation experiments that aimed to evaluate the finite sample properties of the proposed characteristic function based change point detection methods, and compare them to several existing functional change point methods designed to detect change points in the mean and covariance of functional data. All simulations were carried out using the R programming language; [R Core Team \(2022\)](#). Each functional data object was simulated on an equally spaced grid of  $[0; 1]$  containing 50 points. All empirical size and power levels reported below were based on 1,000 independent simulations. We take throughout  $B = 20$  and  $R = 50$ ; see [section A.2](#) for a detailed analysis of the sensitivity of the methods to these choices. These results suggested that choosing smaller  $B$  ( $B \in \{1, \dots, 5\}$ ) generally demonstrated a modest though significant power reduction, but we noticed generally similar performance among the examples we considered for  $B \geq 10$ .

### 2.4.1 Data generation

We consider several data generating processes (DGPs) to simulate [FTS](#) with various types of change points. Each [FTS](#) is simulated after discarding a burn-in sample of length 100. The primary process that we present on is the following:

Auto-Regressive Truncated Karhunen-Loève Process (ARKL): The FTS is generated as

$$X_i(t) = \mu_i(t) + \sum_{d=1}^5 \beta_{i,d} \phi_d(t); \quad t \in [0;1]; \quad (2.25)$$

where  $(\phi_1; \dots; \phi_5)$  are the first five orthonormal cubic B-spline basis functions, and the vector  $\mu_i = (\mu_{i,1}; \dots; \mu_{i,5})^>$  follows a vector autoregression of order one, so that

$$\mu_i = \Psi \mu_{i-1} + \epsilon_i$$

The matrix  $\Psi$  is a random  $5 \times 5$  matrix generated with standard normal entries and then normalized by a constant so that  $k\Psi k_F = 1$ , where  $k k_F$  denotes the Frobenius norm. As such  $\rho$  describes the magnitude of serial dependence of the process, which we take as  $\rho = 0$  (No Dependence),  $\rho = 0.5$  (Moderate Dependence), and  $\rho = 0.8$  (Strong Dependence). Under the null hypothesis of no change point, we set the mean to be zero ( $\mu_i = 0$ ), and the error vector  $\epsilon_i$  is assumed to be a mean zero, 5-variate normal random variable with diagonal covariance matrix  $\text{diag}(1; \rho_{2}^1; \rho_{3}^1; \rho_{4}^1; \rho_{5}^1)$ .

We considered three different ways of generating data following  $H_A$  within this model:

**Mean-Change:** In general we set  $\mu_i(t) = \mu_1 f_i > k g$ . I.e. the mean changes from zero to the function  $\mu_2$  after the change point  $k$ , so that  $\rho$  describes the magnitude of the change. Typically, we set  $\mu_2$  to be a constant function; however, a non-constant  $\mu_2$  based on the sine function is further discussed later in this section.

**Covariance-Change (Cov-Change):** In this case  $\mu_i = 0$ , and for  $i \in f1; \dots; k g$ , we set  $\text{Cov}(\mu_i) = \text{diag}(1; \rho_{2}^1; \rho_{3}^1; \rho_{4}^1; \rho_{5}^1)$ , and for  $i \in fk + 1; \dots; ng$  we set  $\text{Cov}(\mu_i) = \text{diag}(\rho; \rho_{2}^1; \rho_{3}^1; \rho_{4}^1; \rho_{5}^1)$ . The magnitude of the change in the covariance is characterized by the degree to which  $\rho > 1$ .

**Distributional Change (Dist-Change):** Let  $\Sigma = \text{diag}(1; \rho_{2}^1; \rho_{3}^1; \rho_{4}^1; \rho_{5}^1)$  and  $\mu_i = 0$ . For  $i \in f1; \dots; k g$  we generate  $\mu_i \stackrel{iid}{\sim} N(0; \Sigma)$ , where  $N(0; \Sigma)$  denotes a multivariate normal distribution with mean zero and covariance matrix  $\Sigma$ , and for  $i \in fk + 1; \dots; ng$  we generate  $\mu_i = (\mu_{i,1}; \dots; \mu_{i,5})^>$  so that  $\mu_{i,j} \stackrel{iid}{\sim} (1=j^{1=4}) \chi^2(\rho; 1)$ , where  $\chi^2(\rho; 1)$  denotes a gamma distribution with scale parameter 1 and shape parameter  $\rho$  that has been standardized to have mean zero and variance one. In this case the mean function and covariance operator of the sample is homogeneous, but the distribution changes at  $k$ .  $\rho$  tending to infinity causes  $H_0$  to hold, so the magnitude of the change is larger for smaller  $\rho$ .

We also consider two further DGPs, with results on them largely contained in the appendix. In each of them we again considered  $\rho = 0$  (No Dependence),  $\rho = 0.5$  (Moderate Dependence), and  $\rho = 0.8$  (Strong Dependence).

Table 2.1:  $T_n$  size table. Empirical size for  $T_n$  test statistic using simulation, permutation, and approximation based on 1,000 simulations from a ARKL DGP with Gaussian errors. Data lengths of 100, 250, and 500 are considered under varying degrees of dependence. Several bandwidths used to compute the long-run variance estimators and permutation block sizes, both denoted  $h$ , are also considered. The nominal level is set at 0.05.

| $T_n$ TEST STATISTIC |               |       |       |                     |       |       |                   |       |       |
|----------------------|---------------|-------|-------|---------------------|-------|-------|-------------------|-------|-------|
| $n =$                | No Dependence |       |       | Moderate Dependence |       |       | Strong Dependence |       |       |
|                      | 100           | 250   | 500   | 100                 | 250   | 500   | 100               | 250   | 500   |
| Gaussian Simulation  |               |       |       |                     |       |       |                   |       |       |
| $h = 1$              | 0.052         | 0.048 | 0.050 | 0.121               | 0.123 | 0.132 | 0.386             | 0.419 | 0.398 |
| $h = n^{1=3}$        | 0.048         | 0.045 | 0.048 | 0.060               | 0.063 | 0.063 | 0.170             | 0.143 | 0.116 |
| $h = 2n^{1=3}$       | 0.040         | 0.043 | 0.042 | 0.048               | 0.049 | 0.054 | 0.098             | 0.097 | 0.065 |
| Permutation          |               |       |       |                     |       |       |                   |       |       |
| $h = 1$              | 0.047         | 0.046 | 0.046 | 0.248               | 0.271 | 0.248 | 0.597             | 0.652 | 0.663 |
| $h = n^{1=3}$        | 0.049         | 0.046 | 0.052 | 0.109               | 0.088 | 0.072 | 0.342             | 0.314 | 0.270 |
| $h = 2n^{1=3}$       | 0.048         | 0.043 | 0.050 | 0.068               | 0.060 | 0.054 | 0.191             | 0.169 | 0.131 |
| Welch Approximation  |               |       |       |                     |       |       |                   |       |       |
| $h = 1$              | 0.053         | 0.053 | 0.049 | 0.261               | 0.274 | 0.250 | 0.592             | 0.663 | 0.662 |
| $h = n^{1=3}$        | 0.049         | 0.055 | 0.052 | 0.109               | 0.099 | 0.079 | 0.354             | 0.338 | 0.284 |
| $h = 2n^{1=3}$       | 0.043         | 0.049 | 0.046 | 0.083               | 0.064 | 0.060 | 0.221             | 0.186 | 0.145 |

Auto-Regressive Ornstein–Uhlenbeck Process (AROU): We let

$$X_i(t) = \alpha_i(t) + \int_0^t \beta_i(s) X_{i-1}(s) ds + W_i(t); \quad t \in [0; 1]; \quad (2.26)$$

Here  $\alpha_i(t); i \in \mathbb{Z}; t \in [0; 1]$  are IID standard Ornstein–Uhlenbeck processes, which are Gaussian processes with mean zero and covariance  $E\alpha_i(t)\alpha_i(s) = \exp(-\lambda|t-s|)$ .

Functional autoregressive model of order one (FAR(1)): Functional autoregression is surveyed in [Bosq \(2000\)](#) and [Horváth and Kokoszka \(2012\)](#). We generate data according to the FAR(1) model

$$X_i(t) = \alpha_i(t) + \int_0^t \beta_i(s) X_{i-1}(s) ds + W_i(t); \quad t \in [0; 1];$$

where  $\beta_i(s) = 1.3390 \exp(-\lambda(s^2 + t^2))$ , and  $W_i(t); i \in \mathbb{Z}; t \in [0; 1]$  are independent and identically distributed standard Brownian motions.

Table 2.2:  $M_n$  size table Empirical size for  $M_n$  test statistic using simulation and permutation based on 1,000 simulations from a ARKL DGP with Gaussian errors. Data lengths of 100, 250, and 500 are considered under varying degrees of dependence. Several window and block sizes, both denoted  $h$ , are also considered. The nominal level is set at 0.05.

| $M_n$ TEST STATISTIC |               |       |       |                     |       |       |                   |       |       |
|----------------------|---------------|-------|-------|---------------------|-------|-------|-------------------|-------|-------|
| $n =$                | No Dependence |       |       | Moderate Dependence |       |       | Strong Dependence |       |       |
|                      | 100           | 250   | 500   | 100                 | 250   | 500   | 100               | 250   | 500   |
| Gaussian Simulation  |               |       |       |                     |       |       |                   |       |       |
| $h = 1$              | 0.023         | 0.031 | 0.032 | 0.100               | 0.103 | 0.117 | 0.355             | 0.422 | 0.421 |
| $h = n^{1=3}$        | 0.013         | 0.024 | 0.028 | 0.027               | 0.038 | 0.037 | 0.123             | 0.115 | 0.085 |
| $h = 2n^{1=3}$       | 0.008         | 0.019 | 0.018 | 0.009               | 0.021 | 0.023 | 0.038             | 0.040 | 0.037 |
| Permutation          |               |       |       |                     |       |       |                   |       |       |
| $h = 1$              | 0.039         | 0.048 | 0.043 | 0.263               | 0.305 | 0.299 | 0.605             | 0.664 | 0.693 |
| $h = n^{1=3}$        | 0.051         | 0.049 | 0.043 | 0.118               | 0.096 | 0.081 | 0.370             | 0.339 | 0.307 |
| $h = 2n^{1=3}$       | 0.048         | 0.047 | 0.045 | 0.064               | 0.066 | 0.059 | 0.209             | 0.177 | 0.152 |

## 2.4.2 Null simulations

For a no-change simulation when  $n = 100$  ( $n = 500$ ),  $J = 1000$ , and  $h = n^{1=3}$ , the time to compute  $T_n$  and the critical value was approximately 20.63 (22.16) seconds for the simulation approach, 0.51 (0.94) seconds for the Welch approximation, and 0.74 (2.23) seconds for the permutation approach on a standard laptop. Although computational improvements may be possible, this was sufficient for our needs.

Tables 2.1 and 2.2 show the empirical size with nominal level  $\alpha = 0.05$  when data were generated according to the ARKL DGP with values of  $n \in \{100, 250, 500\}$ . Several strengths of dependence are also explored. To attempt to mitigate the effects of dependence, each approach can be slightly modified. When simulating the Gaussian process or using a Welch-style approximation, the long-run covariance estimation can be done with a larger kernel bandwidth. With the permutation approach, block permutation can be used in place of the traditional permutation. We denote both the kernel bandwidth and the length of the block with  $h$ . We investigate  $h = 1; n^{1=3}; 2n^{1=3}$  in each setting. Beyond using these standard choices,  $h$  could also be selected using a data-driven approach; see Horváth et al. (2016) and Rice and Shang (2017).

In each table, all three methods for estimating critical values introduced in section 2.3 exhibited close to nominal size when the simulated process did not exhibit serial de-

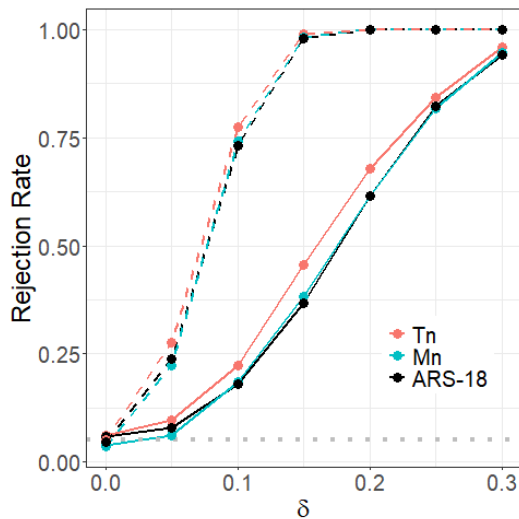
pendence, even with  $n = 100$ . The simulation approach appeared to be more robust to dependence than the permutation method, and the Welch approximation was the least robust. When the block lengths were sufficiently large each method exhibited approximately nominal size for moderately serially dependent data. Tables for nominal size  $\alpha = 0.01$  are given in Appendix A.3, and a similar pattern was observed. The statistic  $T_n$  generally had closer to nominal size for the examples we considered, whereas using the statistic  $M_n$  tended to be more undersized. AROU and FAR(1) generated data are discussed in Appendix A.3, for which the results were similar.

### 2.4.3 Single change point

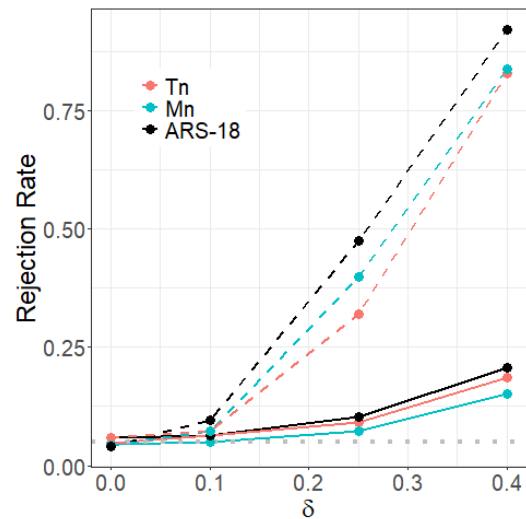
Figure 2.1 investigates the performance of the methods under the mean-change alternative. We compared the proposed measures to the method of Aue et al. (2018), which we designated ARS-18. The data are independently generated by each DGP with a mid-sample mean change, i.e.  $k = n/2$ . The Gaussian simulation approach with the bandwidth parameter  $h = 1$  is presented; however, unreported simulations demonstrated that the Welch approximation and permutation approaches, as well as increasing  $h$ , returned similar results. In each figure, sample sizes 100 and 500 are respectively indicated by solid and dashed lines.

Throughout Figure 2.1, the methods have good power and typically perform similarly. We observed interestingly that the proposed test based on  $T_n$  outperformed ARS-18 under the ARKL DGP. We think this can be explained by how the function defining the change in the mean, in this case a constant shift, aligns with the primary modes of variability of Brownian motion defining the measure  $Q$ . This is supported by considering Figures 2.1b and 2.1c where the characteristic functional-based approaches perform well, but not exceeding that of ARS-18. Unreported simulations demonstrate that integrating against the measure defined by an OU process rather than Brownian motion increased the power when the data were generated via AROU. Hence additional knowledge can increase power of the characteristic functional-based approach. Nonetheless, simulations suggest that using a general measure such as Brownian motion is sufficient in many cases.

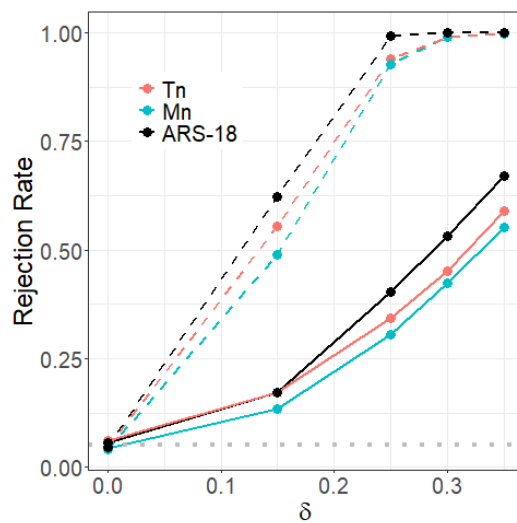
To further investigate the effect of the choice of  $Q$ , we considered a non-constant change that did not tend to align with the leading principal components of Brownian motion; namely, a mean change of the form  $\mu_i(t) = c \sin(2\pi K t) 1_{t > k}$  for  $K = 1; 2; 3; 4$  and  $c = 1; 1/2; 1/4$ . Figure 2.2a presents the results of the simulations when  $n = 100$ . As  $K$  increases, the mean change becomes increasingly orthogonal to a Brownian motion. This appears to explain why power loss is observed for the proposed methods, while the ARS-18 approach maintains power for each value of  $K$ . This highlights the importance and



(a) ARKL DGP



(b) AROU DGP



(c) FAR(1) DGP

Figure 2.1: Data generating process power curves. Power curves as a function of the change magnitude based on mid-sample mean changes on data generated via different DGPs.  $T_n$  simulation (red) and  $M_n$  simulation (blue) are compared, along with ARS-18 (black) for two different data length settings— $n = 100$  (solid) and  $n = 500$  (dashed). 1;000 simulations are computed at each point, with  $h = 1$ . The characteristic approaches show good power in all cases.

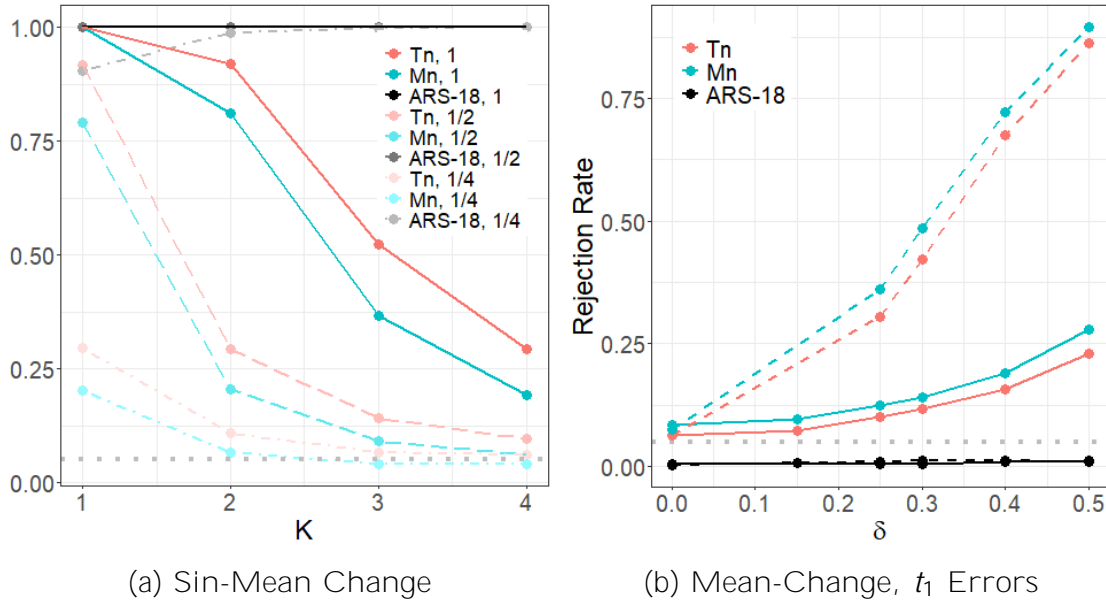


Figure 2.2: Mean change power curves. Power curves based on mid-sample mean changes on data generated via ARKL DGP.  $T_n$  simulation (red) and  $M_n$  simulation (blue) are compared, along with ARS-18 (black) for length 100 data. 1;000 simulations are computed at each point, with independence assumed ( $h = 1$ ). (a) considers the mean change of the form  $y_i(t) = c \sin(2\pi Kt) 1_{fi} > n=2g$ . Several choices for  $c$  are considered—1, 1=2, and 1=4. The characteristic approaches show loss of power while ARS-18 maintains good power for increasing values of  $K$ . (b) considers a mean change, but with errors with distribution  $t_1$  rather than normality. In this case the characteristic approaches show much higher power than ARS-18, which has no power.

influence of the choice of  $Q$ . Similar issues also arise in empirical characteristic function and energy-distance based methods for scalar and vector valued data.

Figure 2.2b examines a constant mean change under errors which have a  $t_1$  distribution. In this case, ARS-18 shows no more than trivial power, while the characteristic functional-based approaches still exhibit reasonable power.

Figure 2.3 investigates the performance of the characteristic functional-based methods under covariance and distribution-based change points. In the case of an eigenvalue change, we also compared to the method of Horváth et al. (2022) which we designated HRZ-22.

Under an eigenvalue change of Figure 2.3a, both  $T_n$  and  $M_n$  exhibit power, even for relatively small changes in  $\lambda$ . However, HRZ-22 shows higher power. We expect some of

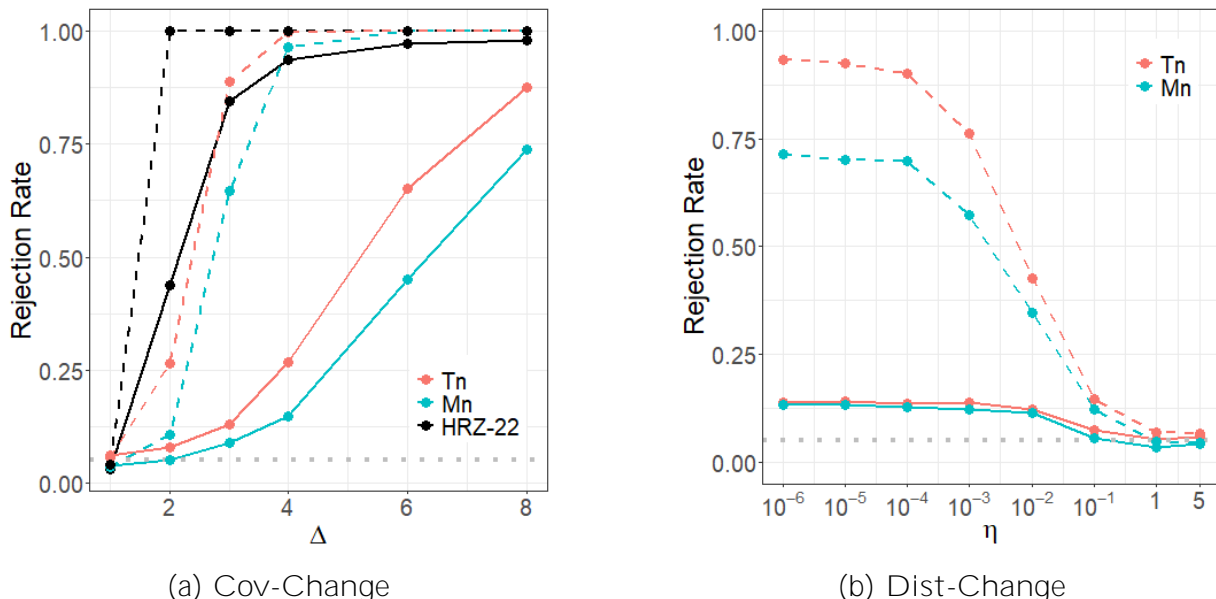


Figure 2.3: Higher order change power curves. Power curves based on mid-sample changes on data generated via different the ARKL DGP.  $T_n$  simulation (red) and  $M_n$  simulation (blue) are compared, along with HRZ-22 (black) for two different data length settings—100 (solid) and 500 (dashed). 1;000 simulations are computed at each point, with independence assumed ( $h = 1$ ). The characteristic approaches show good power in all cases.

this difference is due to the generalized nature of  $T_n$  and  $M_n$  and note that increasing the sample size closes the separation between the approaches.

Figure 2.3b investigates the detection of distributional changes in the data. To our knowledge, there is no comparable functional change point detection method in this situation. Interestingly, even for relatively small changes in distribution, both approaches have some power, which improves for decreasing values of the shape parameter of the distribution, .

Figure 2.4 shows the power of change point detection from ARKL DGP for a mild mean change ( $i(t) = 0.151fi \quad ng$ ) at different change point locations within the sample  $2 f0.5;0.75;0.8;0.9;0.95g$ . The  $T_n$  approach performed the best, with  $M_n$  performing similarly to ARS-18. We also considered a modified ARS-18 method, denoted mARS-18, which is more akin to  $T_n$  in that it is based on an integrated CUSUM process rather than a maximally selected CUSUM process. Given its similar performance to ARS-18,

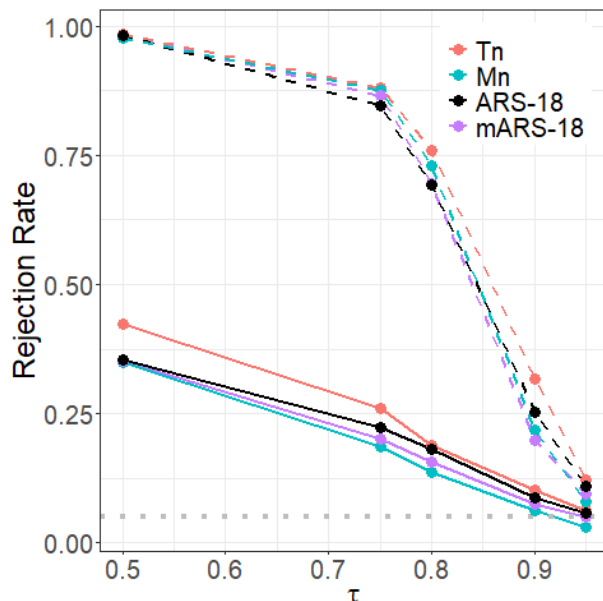


Figure 2.4: Location power curves. Rejection rate for a minor mean change (0.15) from ARKL generated data of length 100 (solid) and 500 (dashed) in 1,000 simulations, with independence assumed ( $h = 1$ ). The  $T_n$  (red),  $M_n$  (blue), ARS-18 (black) and mARS-18 (purple) statistics are compared.  $T_n$  outperforms the other methods which are generally comparable.

we attribute the apparent discrepancy between ARS-18 and  $T_n$  to the nature of the mean change and how it relates to the measure  $Q$ .

## 2.5 Data applications

In this section, we present the results of two applications using the proposed characteristic functional-based change point detection method. We considered Spanish electricity price data and high-frequency S&P500 stock returns data.

Due to its superior performance in the simulation experiments, we present the results when we used the change point detector  $T_n$  as well as Gaussian simulation to estimate the null-quantiles. We set the bandwidth parameter  $h = n^{1/3}$ , although the reported results changed little for the other bandwidths that we considered. Multiple change point detection was carried out using BS as described in section 2.3.4.

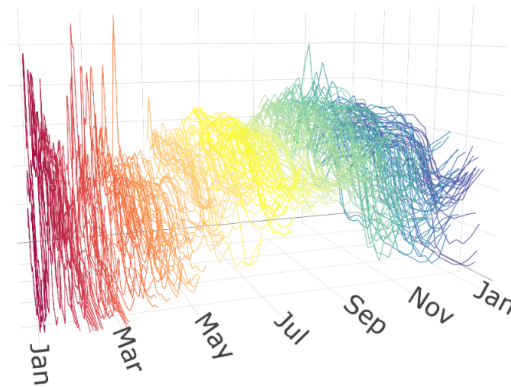


Figure 2.5: Spanish electricity. Daily, hourly observed Spanish electricity spot prices in 2014.

### 2.5.1 Spanish electricity data

The data we considered was the hourly electricity spot prices in Spain in the year 2014 as provided by the Spanish electricity market operator ([www.omie.es](http://www.omie.es)); see [Gonzalez et al. \(2018\)](#) and [Mestre et al. \(2021\)](#). The data are plotted as daily [FTS](#) using linear interpolation in [Figure 2.5](#). A popular approach to forecast such data is to use techniques in functional time series analysis; see e.g. [Jan et al. \(2022\)](#) and [Martínez-Álvarez et al. \(2019\)](#).

The goal of this analysis was to identify change points in the distribution of the model residuals computed from a standard [FTS](#) forecasting model fit to the data. Knowledge of change points in the distribution of model residuals is useful in informing the construction of prediction intervals, as well as assessing the limitations of such models.

The forecasting model that we consider is sometimes termed the Hyndman-Ullah model, abbreviated “HU”, and was initially put forward in [Hyndman and Ullah \(2007\)](#). The  $HU(\mathcal{J})$  model is a potentially nonlinear and non-stationary [FTS](#) model. It supposes that the [FTS](#) under consideration may be well-approximated by its projection onto  $\mathcal{J}$  ortho-normal basis

functions

$$X_i(t) = \hat{m}(t) + \sum_{j=1}^J \hat{\alpha}_{ij} \hat{v}_j(t); \quad t \in [0; 1];$$

where the  $\hat{m}(\cdot)$  is the sample mean function,  $\hat{\alpha}_{ij}$  is referred to as the principal component score and the  $\hat{v}_j(\cdot)$  are the eigenfunctions of the sample covariance operator.

There are three steps to fit and forecast with HU( $J$ ) models:

1. Each  $X_i$  is approximated as  $X_i(t) \approx \hat{m}(t) + \sum_{j=1}^J \hat{\alpha}_{ij} \hat{v}_j(t)$ .  $J$  may be selected according to a number of criteria. Here we use total variation explained; see Chapter 8 of [Ramsay and Silverman \(2005\)](#).
2. To each of the FPC series,  $\hat{v}_j; j = 1; \dots; J$ , we fit a scalar time series model. Notably the models can be distinct for different score series. Here, we use Holt-Winters exponential smoothing with a weekly seasonality fit using the `ets()` function in the `forecast` package in R; see [Hyndman and Khandakar \(2008\)](#) for details.
3. Forecasts or fitted values are constructed as  $\hat{X}_i(t) = \hat{m}(t) + \sum_{j=1}^J \hat{\alpha}_{ij} \hat{v}_j(t)$ , and residual curves are computed as  $r_i(t) = X_i(t) - \hat{X}_i(t)$ .

We set  $J = 7$ , where 7 components were selected to explain at least 99% the data variability. We also investigated  $J = 3$  components, which explain 95% of data variability, and the results changed little.

The residuals curves  $r_i(t); i = 1; \dots; n; t \in [0; 1]$  are shown in Figure 2.6a. Applying BS with the level set at  $\alpha = 0.05$  resulted in three detected change points on the dates March 8 (p-value  $p = 0.002$ ), May 24 ( $p = 0.002$ ), and October 1 ( $p = 0.008$ ). The segmentation is indicated by different colors in Figure 2.6a, and also shown on the original data in Figure 2.6b. The change detected on March 8 appears to separate the highly volatile winter month residuals with the less volatile spring months. The residuals appear to become even less volatile leading into the summer months. The final change identified on October 1 appears to demarcate when the more volatile winter prices start again.

## 2.5.2 S&P 500 data

Stock market assets are traded at such a high volume throughout the day that it is natural to treat their asset prices as nearly continuously observed functional data. In this section,

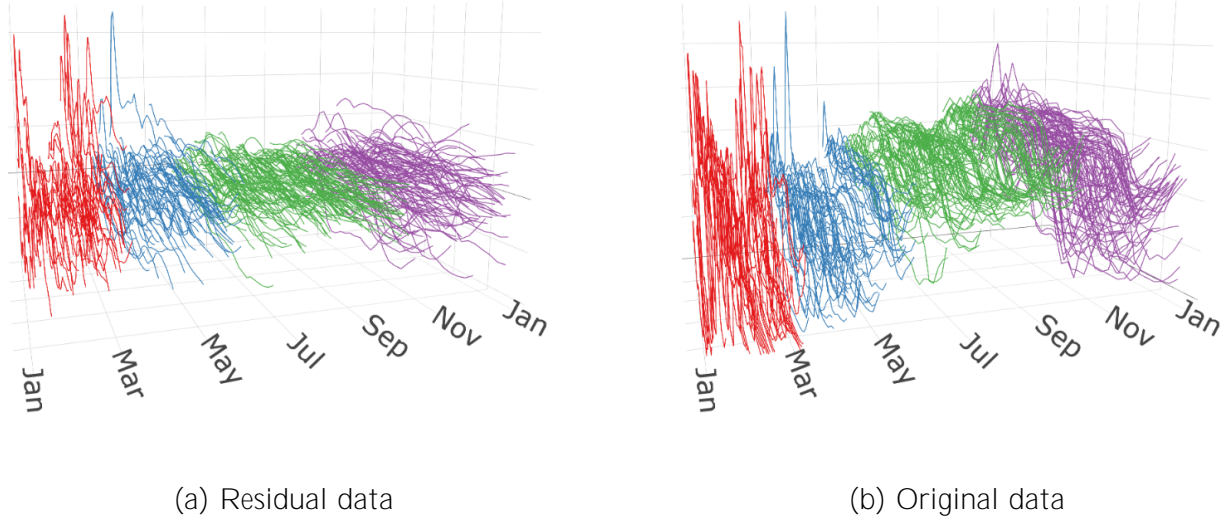


Figure 2.6: Spanish electricity changes. Changes detected on the residuals of the HU(7) ETS forecast for the 2014 Spanish electricity data, placed on the residual and original electricity spot price data. Colors indicate different segments of stationarity, showing three change points detected using the  $T_n$  simulation approach.

we consider five years of the SPDR S&P 500 ETF, obtained at one-minute resolution. This is a fund that tracks the S&P 500 market index. We transformed these data into the overnight cumulative intraday log return curves as follows.

Definition 2. Let  $P_j(t)$ ,  $j = 0; \dots; n$ , be the price of a financial asset at time  $t$  on day  $j$ . As in Rice et al. (2023), the overnight cumulative intraday returns (OCIDRs) are defined as

$$R_j(t) = 100 (\log P_j(t) - \log P_{j-1}(1)); \quad j = 1; \dots; n \quad t \in [0; 1]:$$

We consider the OCIDR curves between 1 January 2019 and 31 December 2023 generated from one minute resolution observations. This period notably overlaps with the COVID-19 global pandemic as well as the onset of the Russo-Ukrainian war. Figure 2.7 shows the OCIDR price curves during this period.

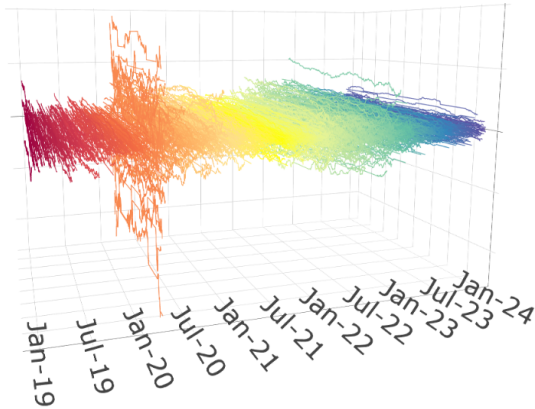


Figure 2.7: SPDR S&P 500. SPDR S&P 500 ETF OCIDRs based on one-minute resolution data from 2019 to 2023.

We take as the goal of this analysis to evaluate for change points the distribution of these curves. Using [BS](#) at level  $\alpha = 0.05$  as detailed in section [2.3.4](#), five change points were detected. Each approximately homogeneous sub-segment is given a different color and shown in [Figure 2.8](#). The first change detected was on 13 February 2020 ( $p < 0.001$ ), approximately coinciding with an apparent shift in the variability of the data, as well as the early stages of COVID-19 pandemic. The next change was detected on 4 May 2020 ( $p < 0.001$ ), at which point COVID-19 and public health responses to it were more widespread. This period had apparently higher variability than the initial or final segments of the data, but less so than the early-pandemic period. Another change was found on 17 November 2020 ( $p = 0.015$ ), and coincides with the beginning of a relatively low volatility period. The next change was detected over a year later, on 3 January 2022 ( $p < 0.001$ ). This demarcates a period of higher variability, and coincides with the lead up to the Russo-Ukrainian war that formally began with the invasion on 24 February 2022. The last change detected was on 10 January 2023 ( $p < 0.001$ ), and the final segment exhibited more typical variability levels.

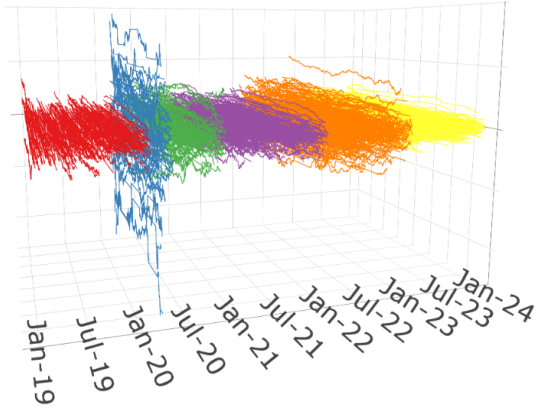


Figure 2.8: SPDR S&P 500 changes. S&P 500 OCIDR curves based on one-minute resolution data from 2019 to 2023. Changing colors segments demarcated by five change points detected using the  $T_n$  with [binary segmentation](#).

## 2.6 Conclusion

This work presents fully functional methodology for detection and consistent estimation of distributional change points in [functional data](#). Several methods for thresholding the proposed test statistics were presented. The assumptions placed for the asymptotic results have been shown to be mild, permitting wide practical use. Simulations confirm the proposed characteristic-based tests are well sized and exhibit good power, even when compared to change detection tuned to specific changes. Applications in electricity and financial data are given as examples where this approach may outperform existing techniques.

As we saw in the development of the test statistics in Eqs (2.4) and (2.5), what is needed in order apply these results for any given separable  $H$ , not necessarily  $L^2[0; 1]$ , is the ability to define suitable measures over  $B(H)$  that both define useful (powerful) integrated empirical characteristic functional metrics, and we can computationally integrate against in practice. When for instance  $H = L^2(\mathbb{R})$  and  $Q$  is a probability measure over  $B(H)$  satisfying Eq (2.9), Theorem (2.2.1) remains valid, but how to choose  $Q$  is more difficult. We leave this issue open here.

# Chapter 3

## Graph-Based Change Point Analysis for Functional Data

### 3.1 Introduction

Multivariate change point detection methods highlight that [cumulative sum \(CUSUM\)](#)-based statistics are readily computable but often rely on parametric assumptions or asymptotic properties to detect changes and may perform poorly under certain high-dimensional scenarios; see [Galeano and Peña \(2007\)](#), [Lavielle and Teyssi re \(2006\)](#), [Aue et al. \(2009\)](#), [Zhang et al. \(2010\)](#), [Wang et al. \(2018\)](#), [Kao et al. \(2018\)](#), [Enikeeva and Harchaoui \(2019\)](#), [Chenouri et al. \(2020\)](#), and [Ramsay and Chenouri \(2023\)](#). To mitigate this, [Chen and Zhang \(2015\)](#) presented a graph-based approach. This approach only requires a definition of distances between observations. It has shown efficacy detecting changes in scalar and multivariate settings; see [Chen and Zhang \(2013\)](#), [Shi et al. \(2017\)](#), [Chen and Friedman \(2017\)](#), [Chu and Chen \(2019\)](#), [Chen \(2019a\)](#), [Chen \(2019b\)](#), [Zhang and Chen \(2021\)](#), [Nie and Nicolae \(2021\)](#), and [Zhou and Chen \(2022\)](#). A recent review by [Chu and Chen \(2022\)](#) highlights key developments for graph-based change point estimation.

Fully functional methods to detect higher order changes for [functional time series \(FTS\)](#) are largely missing from the literature. Recent work by [Horv th et al. \(2025\)](#) consider an approach based on empirical characteristic functionals to detect distributional changes. The graph-based approach proposed by [Chen and Zhang \(2015\)](#) is more general than typically presented on multivariate applications and can be directly applied to [functional data](#). However, the lack of functional literature means its performance, and tuning parameter choices, have not been investigated to our knowledge.

In this chapter, we fill this gap by investigating the efficacy of the graph-based framework in the context of retrospective functional data change point estimation. An extensive simulation study is conducted to evaluate the size and power of graph-based methods under various scenarios, including the low sample settings. The effect of tuning parameter choices, robustness under parametric and non-parametric assumptions, and performance of graph-based methods are evaluated. Practical recommendations for default values of the tuning parameters are given along with the performance of graph-based methods under ideal and non-ideal selection of the tuning parameters. Performance of the graph-based methods for detecting mean, covariance, and distributional changes is evaluated and compared to existing functional methods. [Binary segmentation \(BS\)](#) is used to extend the tests to the multiple change scenario. Several examples of [functional time series](#) are studied.

The remainder of this chapter is organized as follows. The four graph-based change point detection statistics are presented in section [3.2](#), along with their theoretical properties. Approaches for computing thresholds for the statistics are also given. Section [3.3](#) investigates the proposed test statistics via simulations in retrospective scenarios. The simulations examine performance under various settings such as variety of observed data length, dependence structure, change point types, and location of the change(s). Performance of the graph-based statistics under tuning parameter choices are considered. The graph-based methods are compared to existing fully functional change point detection methods. Section [3.4](#) applies graph-based change point detection on multi-year pedestrian counts, high-frequency stock returns, and continuous electricity prices. Finally, section [3.5](#) provides a discussion on graph-based methods and our contributions. All code and data used in this chapter can be found at [gi thub. com/j rvanderdoes/fungraphs](https://github.com/jrvanderdoes/fungraphs) in the R package fungraphs.

## 3.2 Functional graph-based change point analysis

We consider [FTS](#) of the form  $fX_i(t) : i = 1; \dots; n; t \in [0; 1]g$ . These observations are often assumed to take value in the Hilbert space  $L^2[0; 1]$  of real-valued, measurable, and square-integrable functions on the unit interval which is equipped with an inner product. More generally, we can consider data in  $L^p[0; 1]$  for  $p = 1; 2; \dots$ . For  $X(t); Y(t) \in L^p[0; 1]$ , let the corresponding norm be denoted

$$\|X - Y\|_p = \left( \int_0^1 |X(t) - Y(t)|^p dt \right)^{1/p} : \quad (3.1)$$

Below we focus on  $L^1$  and  $L^2$  norms, referred to as simply  $L^1$  and  $L^2$ .

Graph-based change point methods work by treating the data as part of a graph. A graph  $G$  can be described as a set of vertices and a set of edges that connect the vertices. Graph-based change point detection treats the observations as graph vertices, connects the observations via edges, and investigates the resultant edge-defined clusters. Edges are defined using a distance metric and a method for selecting edges based on the distances. Intuitively, a change is detected when there is a cluster of strongly connected observations, or vertices with a relatively large number of connecting edges, which are less connected to observations (vertices) outside the cluster. Formally, changes are detected using a two-sample test statistic based on edge-count.

### 3.2.1 Single change analysis

In the framework of the [at most one change \(AMOC\)](#), detecting a change can be framed as a hypothesis testing problem. Let the probability law governing the process  $X_i$  be denoted by  $F_i$  for  $i = 1; \dots; n$ . We consider the null hypothesis

$$H_0 : F_1 = \dots = F_n \tag{3.2}$$

against the alternative hypothesis

$$H_A : \text{There exists an integer } k = 1; \dots; n \text{ such that } F_1 = \dots = F_k \neq F_{k+1} = \dots = F_n \tag{3.3}$$

Section [3.2.2](#) discusses how one may adapt the proposed test to a more general alternative which allows for multiple changes in the sample of [FTS](#).

Graph-based change point detection is based on transforming observations into a graph and analyzing potential clusters. Additional details on the graph construction are contained in section [3.2.3](#). Clusters in the graph are determined by an edge-count metric that examines the number of edges connecting observations. Changes in the data are detected by comparing the edge-counts to that which would be expected due to randomness. Previous works in the context of multivariate data have identified four potential statistics, namely: original, weighted, max-type, and generalized edge-count statistics ([Shi et al., 2017](#); [Chen and Friedman, 2017](#); [Chu and Chen, 2019](#)). Each statistic splits the observed data into 2 groups at observation location  $k = 1; 2; \dots; n$ . Let  $i$  and  $j$  denote observations,  $i; j = 1; \dots; n$  and  $i \neq j$ . Denote a potential edge between  $i$  and  $j$  as  $(i; j)$ . The test

statistics are based on three quantities that measure the edge counts,

$$\begin{aligned}
R_0(k) &= \sum_{(i,j) \in 2G} \mathbb{I}(g_i(k) \notin g_j(k)); \\
R_1(k) &= \sum_{(i,j) \in 2G} \mathbb{I}(g_i(k) = g_j(k) = 0); \quad \text{and} \\
R_2(k) &= \sum_{(i,j) \in 2G} \mathbb{I}(g_i(k) = g_j(k) = 1);
\end{aligned}$$

where  $g_i(k) = \mathbb{I}(i > k)$ , i.e. the indicator function that a given observation  $i$  is grouped into the second group. Because  $k$  divides the observations into two groups, these quantities can measure the number of edges between the two groups ( $R_0(k)$ ) or within the groups ( $R_1(k)$  and  $R_2(k)$ ). If the observations before and after  $k$  come from different distributions, the observations in each group should be closer to each other than observations in the other group. Relatively small  $R_0(k)$  or large  $R_1(k)$  and  $R_2(k)$  provide evidence to reject the null in Eq (3.2) and conclude that there is a structural change in the FTS at time  $k$ . Unfortunately due to the curse of dimensionality, observations in the same group may not appear close together. Measuring the closeness and remoteness of observations can be valuable in mitigating power and size degradation in high dimensional settings. In FTS, this refers to time series that are observed at a high resolution. The proposed graph-based statistics are defined:

$$Z_0(k) = \frac{R_0(k) - \mathbb{E}(R_0(k))}{\sqrt{\text{Var}(R_0(k))}}; \quad (3.4)$$

$$Z_w(k) = \frac{R_w(k) - \mathbb{E}(R_w(k))}{\sqrt{\text{Var}(R_w(k))}}; \quad (3.5)$$

$$S(k) = \begin{pmatrix} R_1(k) & \mathbb{E}(R_1(k)) \\ R_2(k) & \mathbb{E}(R_2(k)) \end{pmatrix}^{\succ} \quad R^{-1} \begin{pmatrix} R_1(k) & \mathbb{E}(R_1(k)) \\ R_2(k) & \mathbb{E}(R_2(k)) \end{pmatrix}; \quad \text{and} \quad (3.6)$$

$$M(k) = \max(Z_w(k); jZ_{\text{di}}(k)); \quad (3.7)$$

where  $R = \text{Var} \begin{pmatrix} R_1(t) \\ R_2(t) \end{pmatrix}$  and

$$\begin{aligned} R_w(k) &= \frac{n-k-1}{n-2} R_1(k) + \frac{k-1}{n-2} R_2(k); \\ R_d(k) &= R_1(k) \quad R_2(k); \\ Z_{\text{di}}(k) &= \frac{R_d(k) - E(R_d(k))}{\sqrt{\text{Var}(R_d(k))}}; \quad \text{and} \\ R_d(k) &= R_1(k) \quad R_2(k); \end{aligned}$$

We respectively refer to the statistics as the original (3.4), weighted (3.5), generalized (3.6), and max-type (3.7) edge-count statistics.

The original and weighted edge-count statistics are designed to detect mean changes. As discussed for the case of scalar and multivariate data in [Chen and Chu \(2023\)](#), the original statistic tends to perform well when the change point is located near the middle of the data. However, its performance diminishes when the number of observations on either side of the change are uneven, i.e. the change point is near either boundary. This effect is due to inaccurate variance estimation for segments that are smaller, i.e. closer to the boundary. The weighted statistic is computed based on the weighted average of  $R_1(k)$  and  $R_2(k)$  to accommodate the effect of the sample size in each segment. This leads to improved performance near the end points while maintaining the performance for mid-sample changes.

In the multivariate settings, the generalized and max-type edge-count statistics are recommended for general changes. The generalized and max-type test statistics typically exhibit similar performance to each other ([Chen and Zhang, 2015](#); [Chen and Friedman, 2017](#); [Chu and Chen, 2019](#); [Chen and Chu, 2023](#)). These two statistics are designed to be more robust to high-dimensional data by using both the closeness and remoteness of observations to define the statistics. [Chu and Chen \(2019\)](#) show that the generalized edge-count statistic can be decomposed into the uncorrelated quantities  $Z_w(k)$  and  $Z_{\text{di}}(k)$ , where the first is the weighted statistic that is particularly sensitive to location changes and the second is sensitive to scale changes. The max-type edge-count statistic takes the maximum of the uncorrelated quantities, i.e. the max-type statistic is defined as the maximum of  $Z_w(k)$  and  $jZ_{\text{di}}(k)j$ .

A change point is proposed at the location where the test statistic takes its maximum

value, i.e.,

$$\hat{k}_0 = \operatorname{argmax}_{k \in \{f_1, \dots, ng\}} Z_0(k); \quad (3.8)$$

$$\hat{k}_w = \operatorname{argmax}_{k \in \{f_1, \dots, ng\}} Z_w(k); \quad (3.9)$$

$$\hat{k}_S = \operatorname{argmax}_{k \in \{f_1, \dots, ng\}} S(k); \quad \text{and} \quad (3.10)$$

$$\hat{k}_M = \operatorname{argmax}_{k \in \{f_1, \dots, ng\}} M(k); \quad (3.11)$$

The smallest value of  $k$  is taken in the case of ties. In practice,  $k$  is restricted to the interval  $n_0 \leq k \leq n_1$  where  $1 < n_0 < n_1 < n$  to prevent change point estimation based on an inadequate sample size. [Chen et al. \(2020\)](#) suggest using  $n_0 = 0.05n$  and  $n_1 = 0.95n$ . We take this approach, but encourage further investigations into this choice.

The proposed change point is deemed significant if the test statistic from Eqs (3.4)–(3.7) is at least as big as some properly calibrated threshold. One method to determine the threshold is by using permutation. When the data are assumed to be exchangeable and under the null hypothesis in Eq (3.2), test statistics computed from permuted data should be similar to the statistic computed from the original data. In practice, random shuffling is used to minimize computation time.

Let  $T_n$  denote a test statistic computed from Eqs (3.8)–(3.11) on the original data. For a user specified number  $M$ , which is often less than the total possible number of permutations, the observed data  $(X_1, \dots, X_n)$  is randomly permuted  $M$  times, producing permuted samples  $(X_1^{(m)}, \dots, X_n^{(m)}); m \in \{1, \dots, M\}$ . Denoting the test statistics computed from the  $m^{\text{th}}$  permuted sample  $T_n^{(m)}$ , the threshold for the test statistic at significance  $\alpha \in (0, 1)$  is defined as

$$t_{n;(\alpha)} = \inf \left\{ q : \frac{1}{M} \sum_{m=1}^M \mathbf{1}(T_n^{(m)} \leq q) > 1 - \alpha \right\}; \quad (3.12)$$

When the underlying data are serially independent, this critical value (for large  $M$ ) will furnish a test with exact size  $\alpha$ .

When the data are not exchangeable, the dependency structure can be modeled using a [FTS](#) model, e.g. a functional [autoregressive moving average \(ARMA\)](#) model ([Aue et al., 2015](#)) or the projection approach proposed by [Hyndman and Ullah \(2007\)](#). The residuals of well-fit models are approximately exchangeable, and may be used to generate pseudo-samples. Alternatively, [Chen \(2019a\)](#) proposes using circular block permutation for multivariate data to threshold the test statistics; see also block permutation discussions

in [Carlstein \(1986\)](#) and [Lahiri \(1999\)](#). Selection of the bandwidth length in such methods has been a topic of discussion, e.g. see [Horváth et al. \(2016\)](#) and [Rice and Shang \(2017\)](#) for data-driven bandwidth selection.

Permutation tests and random shuffling can become computationally expensive as  $n$  becomes large. Hence, analytic formulas to approximate the permutation  $p$ -values have been proposed. [Chen and Zhang \(2015\)](#) and [Chu and Chen \(2019\)](#) propose approximations which are distribution-free and do not depend on the underlying graph  $G$  nor the data used to construct the graph. The approximations are based on the fact that the limiting distributions of the test statistics converge to respective Gaussian processes, with known covariance functions. However in this chapter, we focus on random shuffling and leave details of approximation  $p$ -values to other works; see [Chen and Zhang \(2015\)](#), [Chu and Chen \(2019\)](#), and [Chen \(2019a\)](#).

A proposed change point is deemed significant at a  $1 - \alpha$  level if  $T_n > t_{n;(1-\alpha)}$  where  $T_n$  is any of the test statistics defined in Eqs (3.8)–(3.11) and  $t_{n;(1-\alpha)}$  is the corresponding threshold defined in Eq (3.12).

### 3.2.2 Multiple change analysis

For many applications, the data appear to exhibit more than one change point. Evaluating data for the presence of multiple change points can again be phrased as a hypothesis testing problem. This problem compares the null hypothesis defined in Eq (3.2) against the  $m$  change alternative hypothesis

$$H_A : F_1 = \dots = F_{k_1} \neq F_{k_1+1} = \dots = F_{k_2} \neq \dots \neq F_{k_{m+1}} = \dots = F_n; \quad (3.13)$$

which states there are  $m$  changes,  $1 \leq k_1 < \dots < k_m \leq n$ .

We use [binary segmentation \(BS\)](#) to estimate multiple change points. Recall that [BS](#) recursively applies a single-change point detection method. If a change is deemed significant, the data are segmented at that point, and the single-change method is reapplied to the resulting sub-segments. For reference, the [BS](#) algorithm is outlined in [Algorithm 1](#).

`BINSEG( $\cdot$ ;  $u$ ;  $n$ )` returns estimates of the number of changes points  $\hat{m}$  and their locations  $\hat{k}_1, \dots, \hat{k}_m$ . The consistency of [BS](#) for multiple change point detection has been shown under various conditions, e.g. see [Scott and Knott \(1974\)](#), [Vostrikova \(1981\)](#), [Venkatraman \(1992\)](#), [Zhang and Chen \(2021\)](#), and [Rice and Zhang \(2022\)](#).

---

Algorithm 1 Binary segmentation algorithm.  $\text{BINSEG}(\cdot; u; n)$

---

```

`   start index;  $u$    end index;  $n$    threshold
if  $u = 1$ , then STOP, else compute  $\hat{k} := \operatorname{argmax}_{k < u} |T_{;u}(k)|$ ;  $T = |T_{;u}^{\hat{k}}|$ 
if  $T > t_{n,(1-\alpha)}$  then
    add  $\hat{k}$  to the set of estimated change points
    run  $\text{BINSEG}(\cdot; \hat{k}; t_{n,(1-\alpha)})$  and  $\text{BINSEG}(\hat{k}; u; t_{n,(1-\alpha)})$ 
    else STOP
end if

```

---

### 3.2.3 Graph construction

Graph-based change point detection is based on transforming observations into a graph and detecting changes based on clusters of that graph. With the test statistics established, we now discuss how to construct the graphs. Although distributional properties of the test statistics have the same asymptotic properties under very mild restrictions on the graph, the graph construction affects the power of the test statistics. [Chen and Friedman \(2017\)](#) discuss the history and assumptions for several popular graphs in the multivariate setting.

Functional observations  $X_i(t); X_j(t) \geq L^p[0; 1]$  can be related using the appropriate  $p$ -norm, given in Eq (3.1). Based on the pairwise distances, edges of the graph can be selected in various ways. We consider three common trees for the construction of the graph. Examples of each tree for bi-variate normal random variables are given in Figure 3.1 for reference.

Minimal spanning trees (MST) connect all vertices (observations) while minimizing the total distance across the edges ([Friedman and Rafsky, 1979](#)). Under MST, a vertex may have multiple edges and all vertices are connected. An example MST graph is shown in Figure 3.1a.

Minimal distance pairing (MDP) pairs vertices so that the total distance between all pairs is minimal. If the number of vertices is odd a pseudo-observation with distance 0 to all vertices is included ([Rosenbaum, 2005](#)). MDPs connect vertices such that each vertex has only a single edge and only two vertices are connected to each other. An example MDP graph is shown in Figure 3.1b.

Nearest neighbour linking (NNL) finds the nearest  $N$  neighbors to each vertex,  $N \geq 1; \dots; ng$ . In NNLs, a vertex has at least  $N$  edges and all vertices do not need to be connected together. In this work, we focus on  $N = 1$ . An example NNL graph is shown in Figure 3.1c.

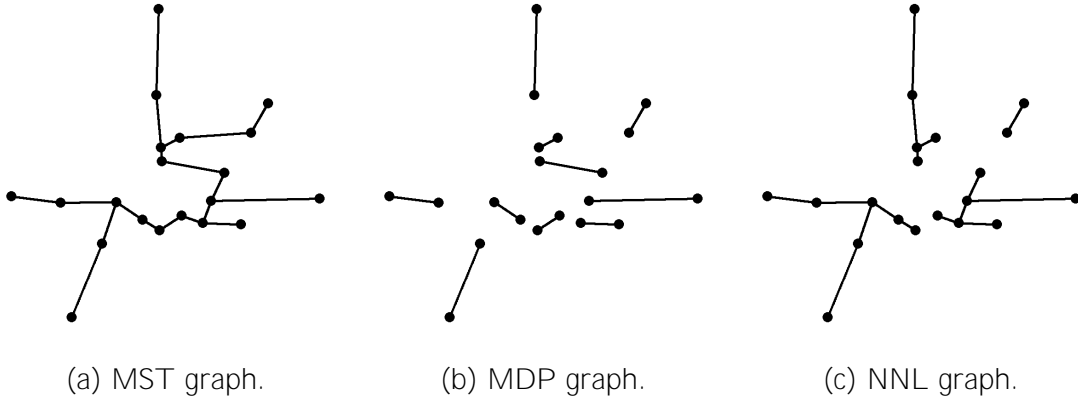


Figure 3.1: Example graph constructions. Graphs constructed on 20 observations which were randomly placed according to a multivariate normal distribution. The graphs are constructed using the following trees: (a) minimal spanning tree (MST), (b) minimal distance pairing (MDP), and (c) nearest neighbour linking (NNL). No orthogonal trees were used.

Although not considered in this work, domain-based or weighted trees may also be valuable, e.g. see [Nie and Nicolae \(2021\)](#) and [Zhou and Chen \(2022\)](#). Computational considerations may necessitate approximations for graph estimation ([Liu and Chen, 2022](#); [Chen and Chu, 2023](#)).

Intuitively, clusters in the graph may be more clearly defined when graphs are denser, i.e. there are additional edges in the graph. One way to increase the density of a graph is by the union of less dense graphs. A  $K$ -tree is the union of the first to  $K$ 'th tree. The  $i$ 'th tree is the tree created under the same procedure, e.g. MST, as the previous 1 to  $i - 1$  trees subject to the constraint that this tree does not contain any edge that is in the previous trees. Thus, each tree is orthogonal to the other  $K - 1$  trees. See Figure 3.2 for an example of a NNL-2.

The effectiveness of graph-based change point detection when using a  $K$ -tree is influenced by the choice of  $K$ . Intuitively, if  $K$  is too large or too small then no clusters will be formed regardless of the underlying distribution, and hence there will be no power to detect changes. Section 3.3 investigates the choice of  $K$ .

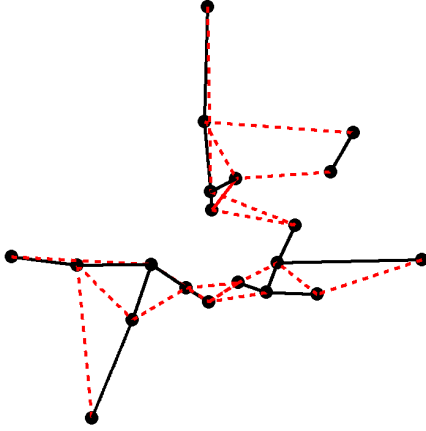


Figure 3.2: Example  $K$ -graph construction. A NNL-2 graph constructed on 20 observations which were randomly placed according to a multivariate normal distribution. The black lines indicate the first NNL tree; see Figure 3.1c. The red lines indicate the second NNL tree, which is orthogonal to the first tree.

### 3.3 Simulations

The practical implications of the proposed tests statistics and related graph-construction parameters are explored in this section. Under various scenarios, several values for the tuning parameters are studied. The graph-based methods are also compared to existing functional methods. When considering changes in the mean, the graph-based methods are compared to the functional mean-change method proposed by [Aue et al. \(2018\)](#). This is denoted ARS-18. When considering changes in the covariance kernel, the graph-based methods are compared to the functional method proposed by [Horváth et al. \(2022\)](#). This is denoted HRZ-22. When considering distributional changes, the graph-based methods are compared to the functional method proposed by [Horváth et al. \(2025\)](#). The simulation-based thresholds for the empirical characteristic functional approach is used and the approach is denoted HRV-25.

Graph-based change point detection methods have several tuning parameters, such as: the tree type, number of orthogonal trees, distance metric, and the test statistic. This section investigates the effect of these choices on the effectiveness of graph-based change point detection. We consider MST, MDP, and NNL trees. All trees were constructed via greedy algorithms. We considered  $K = 1;3;5;7;15$  for the number of orthogonal trees.

We focused on  $L^1$  and  $L^2$  distances. All 4 graph-based test statistics are considered, and we denote each by their first letter in subsequent tables, i.e. original (O), weighted (W), max-type (M), and generalized (G).

Simulation studies were conducted looking at the size and power of the methods under various scenarios. In the single change scenarios, changes in the mean, covariance, and distribution were considered. Dependent data were also investigated. In the multiple change scenario, combinations of these changes were considered. Reported simulations had data which were simulated for lengths varying from  $n = 15$  to  $n = 250$ .

The simulated [functional data](#) were generated using an auto-regressive [Karhunen-Loéve](#) expansion; see [Aue et al. \(2018\)](#). Denoted ARKL, the data were generated such that

$$X_i(t) = \varphi_i(t) + \sum_{d=1}^4 \varphi_{i,d}(t); \quad t \in [0;1]; \quad (3.14)$$

where  $(\varphi_1; \dots; \varphi_4)$  are typically the first four orthonormal cubic B-spline basis functions, and the vector  $\varphi_i = (\varphi_{i,1}; \dots; \varphi_{i,4})^>$  follows a vector autoregression of order one, so that

$$\varphi_i = \Psi \varphi_{i-1} + \varepsilon_i$$

The matrix  $\Psi$  is a random  $4 \times 4$  matrix generated with standard normal entries and then normalized by a constant so that  $k\Psi k_F = 1$ , where  $k k_F$  denotes the Frobenius norm. As such  $\rho$  describes the magnitude of serial dependence of the process, which we take as  $\rho = 0$  by default to generate independent data. Under the null hypothesis of no change point, we set the mean to be zero ( $\varphi_i = 0$ ), and the error vector  $\varepsilon_i$  is assumed to be a mean zero, 4 variate normal random variable with diagonal covariance matrix  $\text{diag}(3;2;1;0.5)$ .

Mean changes are created in ARKL by setting  $\varphi_i(t) = 1$  for  $t > k$ , i.e. the mean changes from zero to the function  $1$  after the change point  $k$ .

Covariance changes in ARKL are created by setting  $\rho = 0$  and for  $i \in [1; \dots; k]$  letting  $\text{Cov}(\varphi_i) = \text{diag}(3;2;1;0.5)$ , and for  $i \in [k+1; \dots; n]$ ,  $\text{Cov}(\varphi_i) = \text{diag}(3-\delta; 2-\delta; \delta; 0.5)$ . The magnitude of the change in the covariance is characterized by the degree to which  $\delta > 1$ .

Distributional changes in ARKL are created by first letting  $\text{Cov}(\varphi_i) = \text{diag}(3;2;1;0.5)$  and  $\rho = 0$ . Then for  $i \in [1; \dots; k]$  we generate  $\varphi_i \stackrel{iid}{\sim} N(0; \Sigma)$ , where  $N(0; \Sigma)$  denotes a multivariate normal distribution with mean zero and covariance matrix  $\Sigma$ , and for  $i \in [k+1; \dots; n]$  we generate  $\varphi_i = (\varphi_{i,1}; \dots; \varphi_{i,5})^>$  so that  $\varphi_{i,j} \stackrel{iid}{\sim} \Gamma(\alpha_j; 1)$  ( $1=j=4$ ), where  $\Gamma(\alpha; 1)$  denotes a gamma distribution with scale parameter 1 and shape parameter  $\alpha$  that has been

Table 3.1: Size table under  $L^2$  distance. Table for the empirical sizes of the graph-based change point detection methods. Combinations of test statistics, tree types, and orthogonal tree counts tuning parameters were examined. Each graph-based test statistic was computed using the  $L^2$  distance. Several data lengths were considered,  $n = 15; 25; 50; 100; 200$ . In general, the graph-based change point detection methods were well-sized.

| Statistic = | 15    |       |       |       | 25    |       |       |       | 50    |       |       |       |
|-------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|             | O     | W     | M     | G     | O     | W     | M     | G     | O     | W     | M     | G     |
| MDP-1       | 0.060 | 0.000 | 0.000 | 0.000 | 0.040 | 0.031 | 0.031 | 0.033 | 0.016 | 0.016 | 0.013 | 0.013 |
| MDP-3       | 0.067 | 0.068 | 0.068 | 0.076 | 0.046 | 0.037 | 0.039 | 0.039 | 0.060 | 0.060 | 0.054 | 0.055 |
| MDP-5       | 0.070 | 0.069 | 0.069 | 0.073 | 0.044 | 0.047 | 0.047 | 0.049 | 0.043 | 0.044 | 0.048 | 0.047 |
| MDP-7       | 0.070 | 0.078 | 0.078 | 0.079 | 0.057 | 0.051 | 0.051 | 0.047 | 0.052 | 0.054 | 0.055 | 0.056 |
| MDP-15      | -     | -     | -     | -     | -     | -     | -     | -     | 0.055 | 0.055 | 0.057 | 0.057 |
| NNL-1       | 0.064 | 0.044 | 0.044 | 0.048 | 0.052 | 0.051 | 0.051 | 0.054 | 0.061 | 0.059 | 0.059 | 0.052 |
| NNL-3       | 0.069 | 0.059 | 0.059 | 0.057 | 0.054 | 0.056 | 0.056 | 0.056 | 0.055 | 0.059 | 0.059 | 0.058 |
| NNL-5       | 0.056 | 0.066 | 0.066 | 0.064 | 0.056 | 0.054 | 0.054 | 0.056 | 0.052 | 0.067 | 0.067 | 0.067 |
| NNL-7       | 0.057 | 0.059 | 0.063 | 0.066 | 0.051 | 0.053 | 0.053 | 0.052 | 0.048 | 0.065 | 0.065 | 0.060 |
| NNL-15      | -     | -     | -     | -     | -     | -     | -     | -     | 0.055 | 0.064 | 0.064 | 0.057 |
| MST-1       | 0.052 | 0.043 | 0.048 | 0.050 | 0.050 | 0.044 | 0.045 | 0.042 | 0.062 | 0.058 | 0.057 | 0.063 |
| MST-3       | 0.074 | 0.079 | 0.073 | 0.075 | 0.054 | 0.049 | 0.047 | 0.042 | 0.055 | 0.059 | 0.060 | 0.056 |
| MST-5       | 0.063 | 0.065 | 0.066 | 0.071 | 0.056 | 0.055 | 0.052 | 0.053 | 0.052 | 0.056 | 0.058 | 0.056 |
| MST-7       | 0.050 | 0.054 | 0.063 | 0.064 | 0.048 | 0.048 | 0.044 | 0.041 | 0.041 | 0.059 | 0.063 | 0.057 |
| MST-15      | -     | -     | -     | -     | -     | -     | -     | -     | 0.054 | 0.062 | 0.060 | 0.057 |

| Statistic = | 100   |       |       |       | 200   |       |       |       |
|-------------|-------|-------|-------|-------|-------|-------|-------|-------|
|             | O     | W     | M     | G     | O     | W     | M     | G     |
| MDP-1       | 0.063 | 0.063 | 0.061 | 0.065 | 0.061 | 0.061 | 0.052 | 0.060 |
| MDP-3       | 0.046 | 0.048 | 0.045 | 0.049 | 0.073 | 0.073 | 0.067 | 0.067 |
| MDP-5       | 0.064 | 0.064 | 0.061 | 0.065 | 0.054 | 0.054 | 0.055 | 0.052 |
| MDP-7       | 0.059 | 0.060 | 0.057 | 0.053 | 0.058 | 0.058 | 0.052 | 0.055 |
| MDP-15      | 0.052 | 0.052 | 0.045 | 0.045 | 0.058 | 0.058 | 0.056 | 0.055 |
| NNL-1       | 0.048 | 0.067 | 0.067 | 0.064 | 0.060 | 0.054 | 0.053 | 0.060 |
| NNL-3       | 0.047 | 0.054 | 0.053 | 0.052 | 0.077 | 0.059 | 0.061 | 0.061 |
| NNL-5       | 0.061 | 0.064 | 0.065 | 0.062 | 0.063 | 0.061 | 0.059 | 0.060 |
| NNL-7       | 0.055 | 0.059 | 0.059 | 0.058 | 0.066 | 0.055 | 0.055 | 0.051 |
| NNL-15      | 0.045 | 0.054 | 0.054 | 0.053 | 0.062 | 0.061 | 0.061 | 0.064 |
| MST-1       | 0.047 | 0.044 | 0.047 | 0.054 | 0.064 | 0.060 | 0.056 | 0.055 |
| MST-3       | 0.049 | 0.045 | 0.044 | 0.052 | 0.078 | 0.059 | 0.061 | 0.057 |
| MST-5       | 0.055 | 0.058 | 0.058 | 0.051 | 0.073 | 0.058 | 0.059 | 0.053 |
| MST-7       | 0.065 | 0.058 | 0.059 | 0.061 | 0.053 | 0.053 | 0.054 | 0.050 |
| MST-15      | 0.051 | 0.055 | 0.055 | 0.054 | 0.058 | 0.059 | 0.060 | 0.064 |

standardized to have mean 0 and variance 1. In this case the mean function and covariance operator of the sample is homogeneous, but the distribution changes at  $k$ . Note that tending to infinity causes  $H_0$  to hold, so the magnitude of the change is larger for smaller values of  $k$ . We also consider an centered exponential distribution created under the same framework.

For brevity, when the data were generated via ARKL only the parameters that were different from the previously stated defaults are given. Additionally, 1,000 independent simulations per scenario were conducted unless otherwise stated. In order to detect changes, the random shuffling method was used to compute the  $p$ -values,  $M = 1,000$ . Reported

Table 3.2: Size table under varying distance metrics. Table for the empirical sizes of the graph-based change point detection statistics with varying tree types and orthogonal tree counts using  $L^2$  and  $L^1$  distance metrics. Data were length  $n = 50$ . Both  $L^2$  and  $L^1$  were typically well-sized, with the exception of MDP-1 which was very conservative.

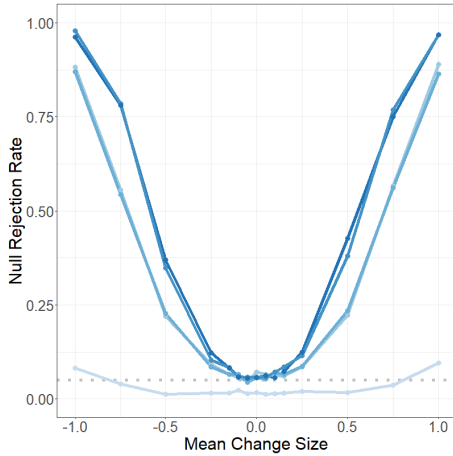
| Distance =  | Original Data |       |       |       |       |       |       |       |
|-------------|---------------|-------|-------|-------|-------|-------|-------|-------|
|             | $L^1$         |       |       |       | $L^2$ |       |       |       |
| Statistic = | O             | W     | M     | G     | O     | W     | M     | G     |
| MDP-1       | 0.018         | 0.018 | 0.018 | 0.018 | 0.010 | 0.010 | 0.010 | 0.010 |
| MDP-3       | 0.063         | 0.063 | 0.057 | 0.061 | 0.051 | 0.051 | 0.046 | 0.048 |
| MDP-5       | 0.054         | 0.054 | 0.057 | 0.057 | 0.057 | 0.057 | 0.056 | 0.056 |
| MDP-7       | 0.056         | 0.056 | 0.057 | 0.056 | 0.053 | 0.055 | 0.061 | 0.059 |
| MDP-15      | 0.052         | 0.052 | 0.055 | 0.055 | 0.056 | 0.057 | 0.054 | 0.054 |
| NNL-1       | 0.056         | 0.059 | 0.060 | 0.062 | 0.050 | 0.052 | 0.052 | 0.049 |
| NNL3        | 0.052         | 0.051 | 0.051 | 0.05  | 0.059 | 0.063 | 0.064 | 0.069 |
| NNL5        | 0.061         | 0.052 | 0.051 | 0.046 | 0.058 | 0.064 | 0.064 | 0.067 |
| NNL7        | 0.061         | 0.052 | 0.052 | 0.058 | 0.058 | 0.057 | 0.057 | 0.060 |
| NNL15       | 0.061         | 0.051 | 0.051 | 0.052 | 0.053 | 0.057 | 0.057 | 0.060 |
| MST-1       | 0.058         | 0.064 | 0.064 | 0.064 | 0.046 | 0.059 | 0.056 | 0.064 |
| MST-3       | 0.056         | 0.054 | 0.053 | 0.046 | 0.06  | 0.068 | 0.067 | 0.061 |
| MST-5       | 0.065         | 0.056 | 0.055 | 0.060 | 0.056 | 0.062 | 0.064 | 0.062 |
| MST-7       | 0.057         | 0.055 | 0.058 | 0.059 | 0.065 | 0.053 | 0.053 | 0.051 |
| MST-15      | 0.052         | 0.046 | 0.049 | 0.054 | 0.051 | 0.047 | 0.048 | 0.056 |

$p$ -values were compared to the nominal significance level of  $\alpha = 0.05$ .

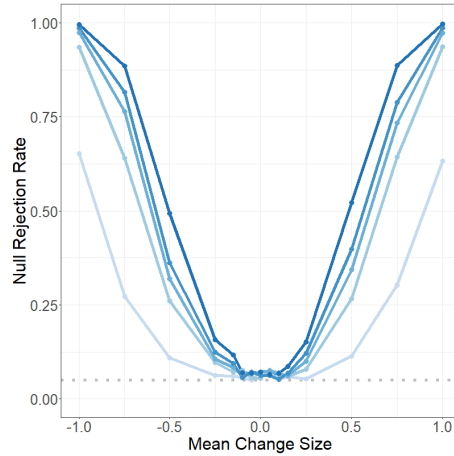
### 3.3.1 No change simulations

Under the no-change null hypothesis, Table 3.1 investigates the empirical size of the graph-based tests using the  $L^2$  distance metric. Table 3.1 shows that for all investigated sample sizes— $n = 15; 25; 50; 100; 200$ —the graph-based tests were reasonably well-sized. The various edge-count metrics and tree types performed similarly. An increasing number of orthogonal trees resulted in similar or better sizes, provided the sample was sufficiently long. Since change point detection is often considered more difficult for smaller samples, we note that when  $n = 50$  the graph-based methods are well-sized. Throughout this section, we use  $n = 50$  for simulations.

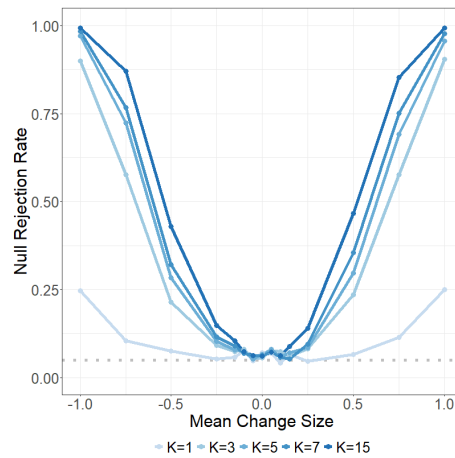
Table 3.2 examines the size of the test statistic under  $L^1$  and  $L^2$  distance metrics. Both  $L^1$  and  $L^2$  perform well over the various scenarios. Unreported values of  $L^p$  generally performed well, with the exception that  $L^1$  was generally poorly-sized. We report the remaining simulations in this section using the  $L^2$  distance, although similar results were seen for  $L^1$ .



(a) MDP graph



(b) MST graph



(c) NNL graph

Figure 3.3: Mean power curves for normal errors. Power curves as a function of the change magnitude based on mid-sample mean changes. The change magnitude varied from  $-1$  to  $1$  with samples of length  $n = 50$ . The number of orthogonal trees,  $K$ , are colored by increasing blue intensity. The dotted horizontal gray line is the nominal significance level. Power curves were constructed using the generalized test statistic as in Eq (3.6). Power curves by graph types are displayed in each figure: (a) MDP, (b) MST, and (c) NNL. Power tended to increase with a larger number of orthogonal trees. MST and NNL were marginally more powerful than MDP.

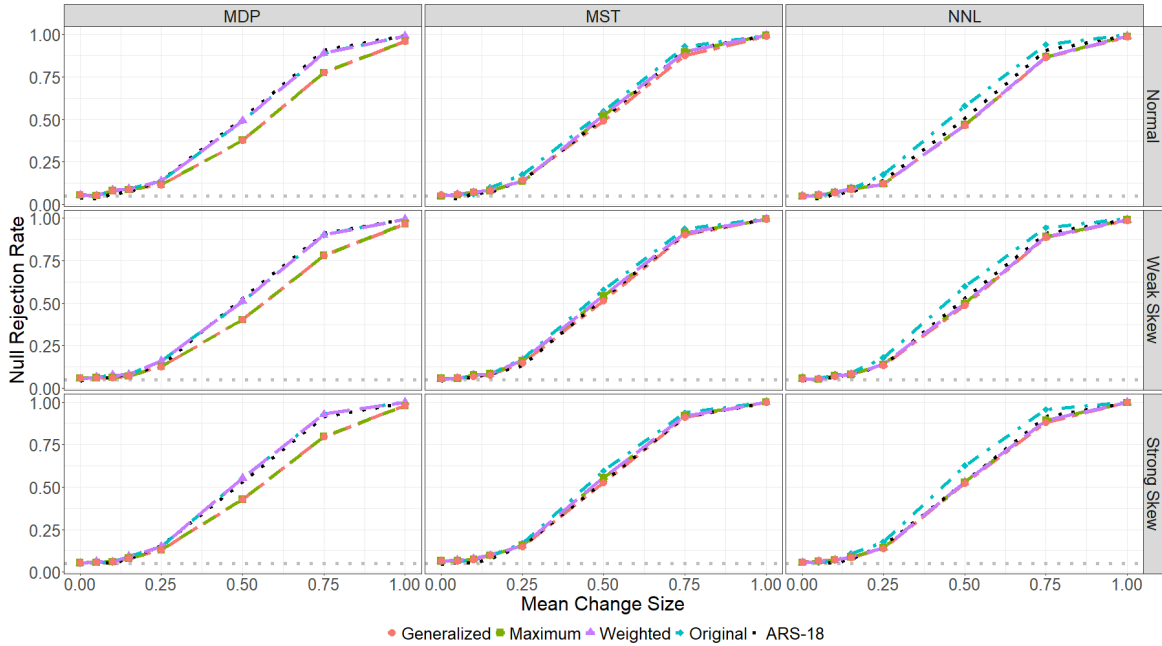


Figure 3.4: Mean power curves for skewed errors. Power curves as a function of the change magnitude based on mid-sample mean changes with normal and skewed-normal error distributions. The change magnitude varied from 0 to 1 with samples of length  $n = 50$ . The top plot used data normal errors while the middle and bottom plots used increasingly skewed-normal errors. The methods are generally comparable. The weighted and original methods performed better than the generalized or max-type statistic, especially for the MDP graph. The original test statistic exhibited power comparable to ARS-18.

### 3.3.2 Single change simulations

Under the AMOC alternative, we considered the effect of the number of orthogonal trees. Figure 3.3 examines the effect of the number of orthogonal trees,  $K$ , for mean change point detection. The data were simulated using ARKL as in Eq (3.14) for length  $n = 50$  samples. The change magnitudes  $\delta = 0; 0.05; 0.1; 0.15; 0.25; 0.5; 0.75; 1$  were considered for the mid-sample change. The errors were normally distributed. The results when using the generalized test statistic are shown in Eq (3.6); unreported simulations showed similar patterns for the other graph-based test statistics.

For each graph construction, increasing the number of orthogonal trees tended to increase the power of the test. However, the increased power comes at the cost of requiring additional data and computational time to fit the graph and assuming that the changes

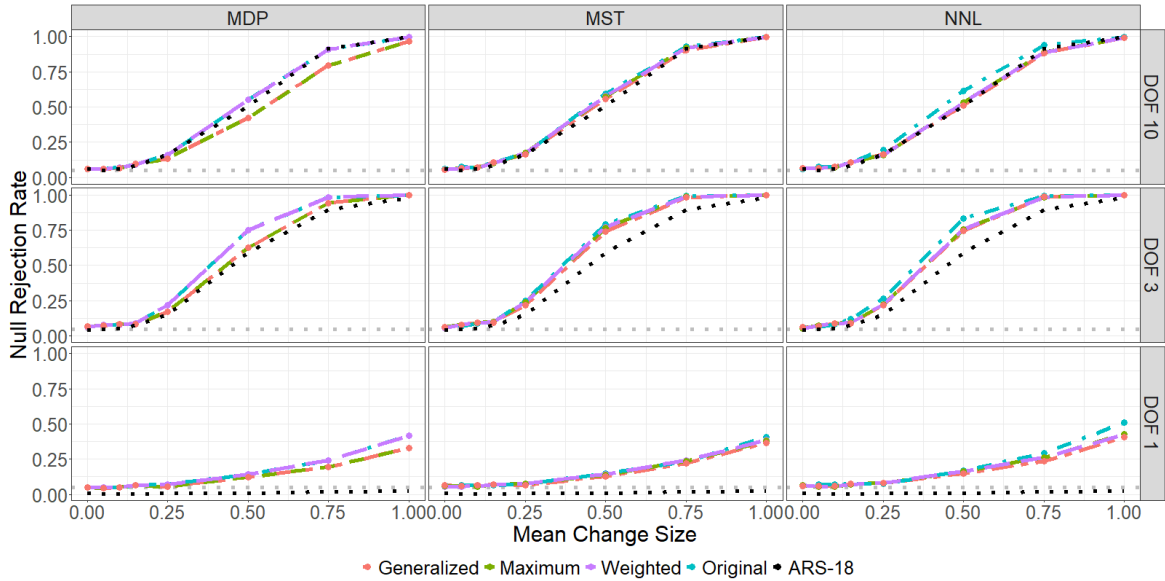


Figure 3.5: Mean power curves for  $t$ -distribution errors. Power curves as a function of the change magnitude for a mid-sample mean change with errors distributed according to a  $t$ -distribution. The change magnitude varied from 0 to 1 with samples of length  $n = 50$ . The dotted horizontal gray line is the nominal significance level. The top figure was generated from data with  $t_{10}$  errors, the middle with  $t_3$  errors, and the bottom with  $t_1$  errors; see also Figure 3.4 for power under Gaussian errors. As the normality assumption was increasingly violated, the graph-based methods increasingly outperformed ARS-18 in terms of size and power. The graph types performed similarly. The test statistics were also similar, although original and weighted typically performed better than generalized and max-type.

are sufficiently separated. Hence, while increasing  $K$  can lead to increased power, it may overlook changes. Thus, the choice of  $K$  is strongly data dependent.

In this section, we consider only 15-trees due to the increased power and its additional robustness to non-normal errors. Nonetheless, typically all choices are reasonable, with the notable exception of 1-trees often having low power. Yet even the power loss of 1-trees can be mitigated using max-type or generalized test statistics. Therefore, the conclusions in this section generally apply to any number of orthogonal trees. When the distance between changes is unknown, a smaller value for  $K$  is recommended. Practically it may be valuable to begin with a larger  $K$  and shrink  $K$  when no changes are detected to ensure no edge changes are present.

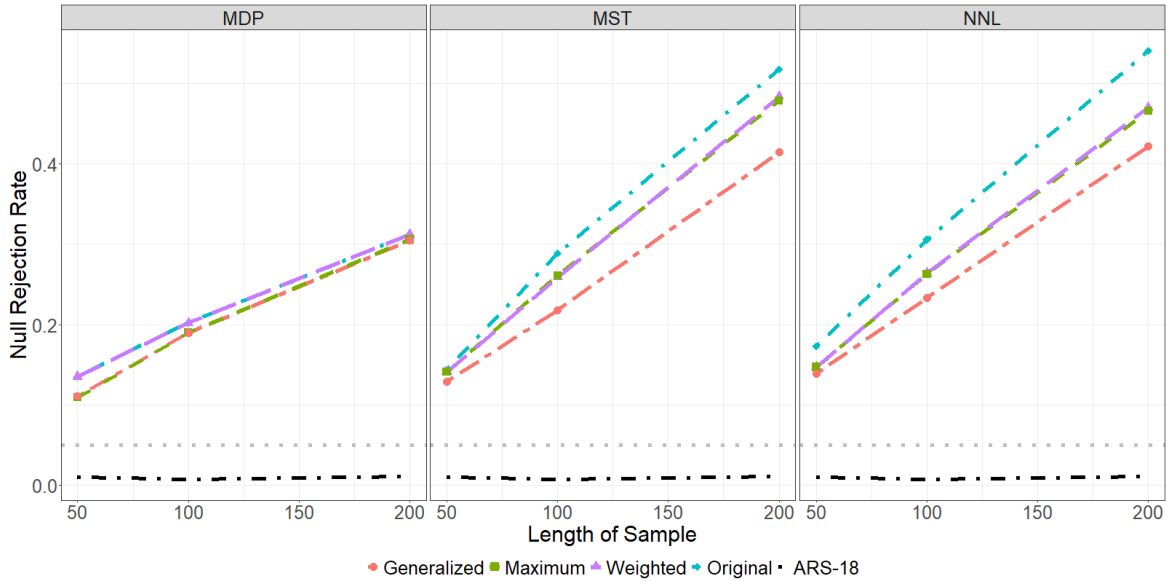


Figure 3.6: Power curves for  $t_1$  errors with varying data length. Power curves as a function of data length for a mid-sample mean change with  $t_1$  errors. The change magnitude was set to 0.5 for data lengths of  $n = 50; 100; 200$ . The dotted horizontal gray line is the nominal significance level. Improvement over increasing data length  $n$  is observed for all graph-based approaches, but not for ARS-18. MST and NNL performed better than MDP. The original statistic performed better than the weighted and max-type. The weighted and max-type statistics performed better than the generalized statistic.

Power curves investigating the single mean change alternative under normally-skewed errors are given in Figure 3.4 for  $n = 50$ . In each simulation, the first half of the sample had mean 0 which abruptly changed at the midpoint to the second half mean,  $\mu = 0; 0.05; 0.1; 0.15; 0.25; 0.5; 0.75; 1$ . The graph-based methods exhibited similar power to each other and ARS-18. As expected, the weighted and original methods, which are tuned for mean changes, performed better than the other graph-based methods in all cases, although the differences were typically small.

Figure 3.5 compares the power of graph-based approaches and ARS-18 for mean changes under varying levels of non-normality. Mean changes were considered such that  $\mu = 0; 0.05; 0.1; 0.15; 0.25; 0.5; 0.75; 1$ . The non-normality came in the form of heavy tails with the errors distributions being  $t_{10}$ ,  $t_3$ , and  $t_1$ . Figure 3.5 shows that the graph-based methods were more robust and outperformed ARS-18 when the error distributions had heavier tails. Again the weighted and original methods, which are tuned to the mean

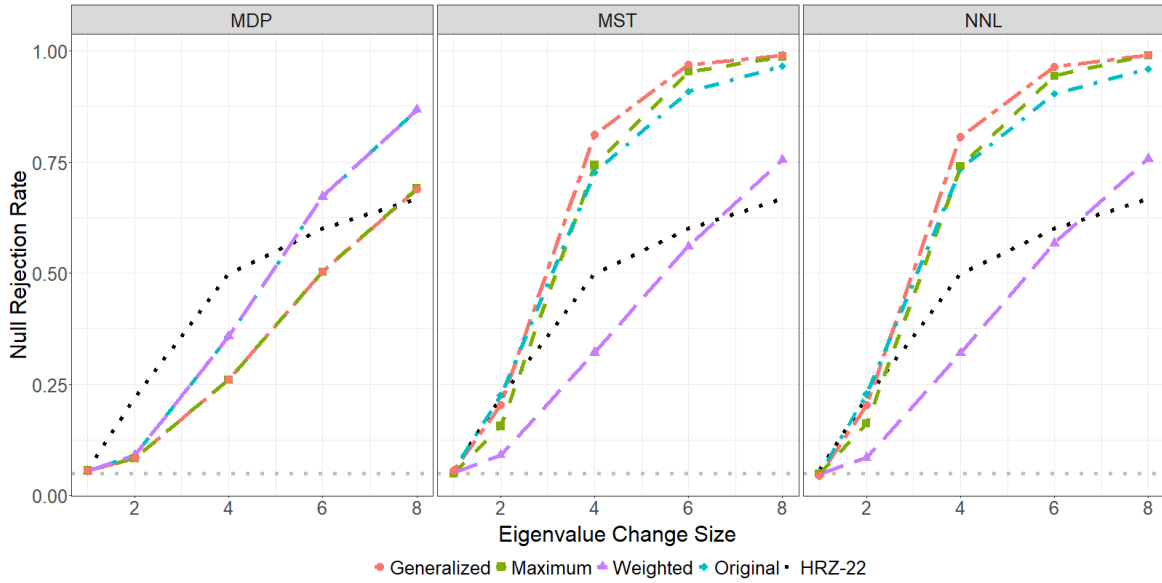


Figure 3.7: Eigenvalue change power curves. Power curves as a function of the change magnitude based on mid-sample eigenvalue changes. The change magnitude was varied from 1 to 8 for sample length  $n = 50$ . The dotted horizontal gray line is the nominal significance level. For MST and NNL graphs, the graph-based statistics quickly surpassed HRZ-22 in terms of power. MDP graphs performed worse than HRZ-22 for smaller eigenvalue changes and better for larger eigenvalue changes. Interestingly, the weighted statistic was not strongly influenced by the graph type, where the other graph-based test statistics exhibited increased power for MST and NNL graphs.

change, performed better than the generalized and max-type graph-based methods.

A clear loss in power was observed as the errors became increasingly heavy-tailed; see Figure 3.5. The effects of data length,  $n$ , on this power loss are considered in Figure 3.6. Figure 3.6 shows the power of the graph-based and ARS-18 methods for  $n = 50; 100; 200$  when the errors were distributed according to a  $t_1$  distribution. The data were generated using the ARKL as in Eq (3.14) with a small, centered mean change of  $\delta = 0.5$ . The ARS-18 did not benefit from increased sample sizes while the graph-based approaches increased in power. Interestingly, MDP does not benefit as heavily from increasing  $n$  when compared to MST and NNL. The original statistic performed better than the weighted and max-type statistics. The weighted and max-type statistics performed better than the generalized statistic.

A valuable feature of the graph-based approaches is their ability to capture changes

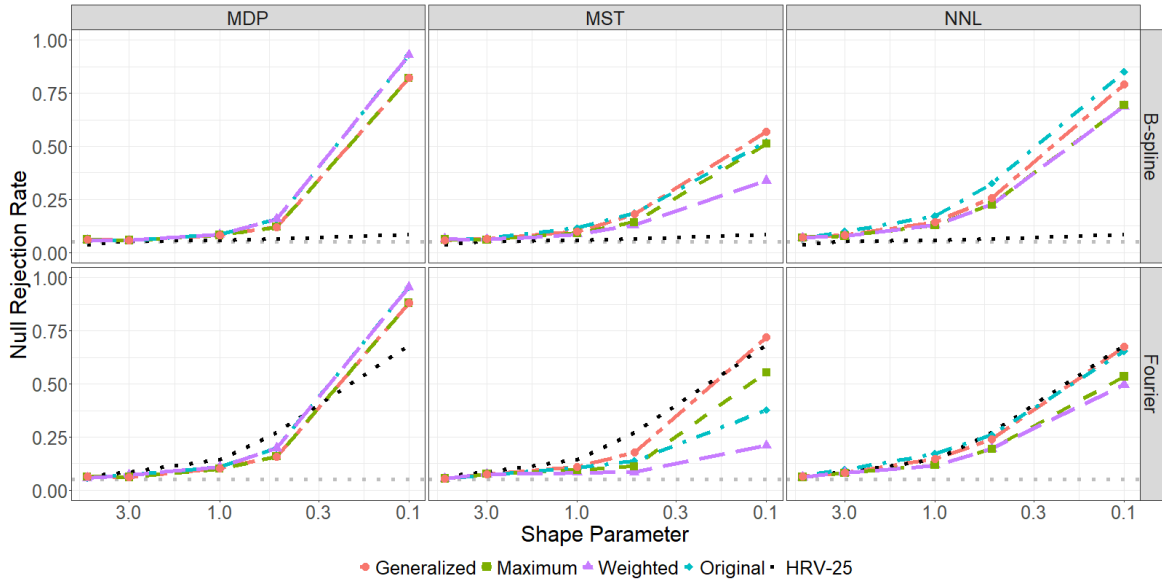


Figure 3.8: Distributional change power curves. Power curves as a function of the change magnitude based on mid-sample distributional changes. The change magnitude was varied from 5 to 0:1 for sample length  $n = 50$ . The dotted horizontal gray line is the nominal significance level. Both the graph-based and HRV-25 methods improved with increased distributional differences. The graph-based methods performed better for extreme cases and for the B-spline basis, but HRV-25 performed better for moderate changes with the Fourier basis.

of various nature, e.g. changes in the data mean or changes in the error distributions. Detecting changes in the eigenvalues of the models innovation covariance is shown in Figure 3.7. For  $n = 50$ , data were generated per ARKL as in Eq (3.14) with normal errors. The eigenvalue changes were considered for  $\lambda = 1; 2; 4; 6; 8$ . HRZ-22 outperformed the graph-based approaches for small changes, but was quickly surpassed as the change magnitude became more extreme.

Figure 3.8 examines the power of the methods when detecting distributional changes midway through data of length  $n = 50$ . The first half of the data had errors with a normal distribution, and the second half had errors with a standardized Gamma distribution. The shape parameter values considered were  $\lambda = 5; 3; 1; 0.5; 0.1$ . The graph-based and HRV-25 methods both exhibited power in detection of distributional changes. The power increased for more significant changes. The graph-based methods performed better than HRV-25 for large differences and when using the B-spline basis. HRV-25 performed better for

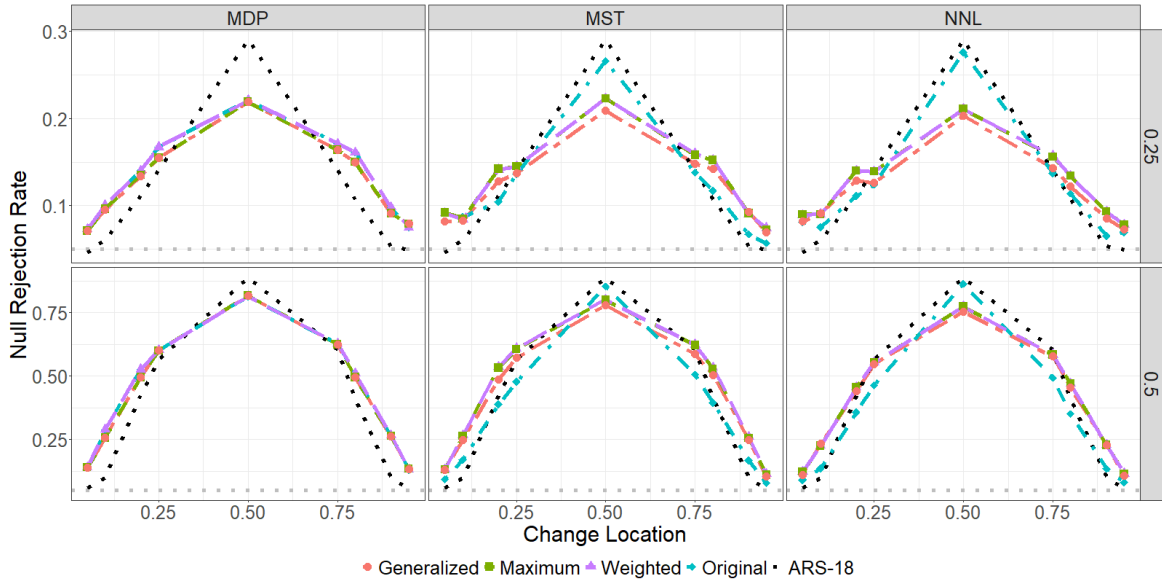


Figure 3.9: Mean change location power curves. Power curves as a function of the change location based on mean changes. Change magnitudes of  $\delta = 0.25; 0.5$  for length  $n = 100$  data was considered. The change location  $k$  was varied from 5 to 95. The dotted horizontal gray line is the nominal significance level. The graph-based approaches exhibited more robustness to changes at the edges of the data than ARS-18. ARS-18 exhibited more power for mid-sample changes, particularly for mild changes. The weighted statistic performed best among graph-based methods, particularly with MST and NNL graph types.

moderate changes with the Fourier basis. Consideration of the data structure is crucial when selecting the optimal approach for detection of distributional changes.

Often real data have changes skewed to one end of the sample. Thus, we also considered the performance of the models when the location of the change was modified. For  $n = 100$ , change point locations of  $k = 5; 10; 20; 25; 50; 75; 80; 90; 95$  were examined. At each change location, mild ( $\delta = 0.25$ ) and moderate ( $\delta = 0.5$ ) mean changes were considered. Figure 3.9 shows the graph-based methods were more robust than ARS-18 to changes at the edge of the sample. Yet, ARS-18 exhibited better detection under mid-sample mean changes, specifically when the change was mild.

The graph-based change point detection approach proposed in section 3.2 is based on serial independence of the data. When the data are serially dependent, models can often be used to capture the dependence, and then change point analysis conducted on the independent, or near-independent, residuals; see section 3.4.3. Yet sometimes dependency exists

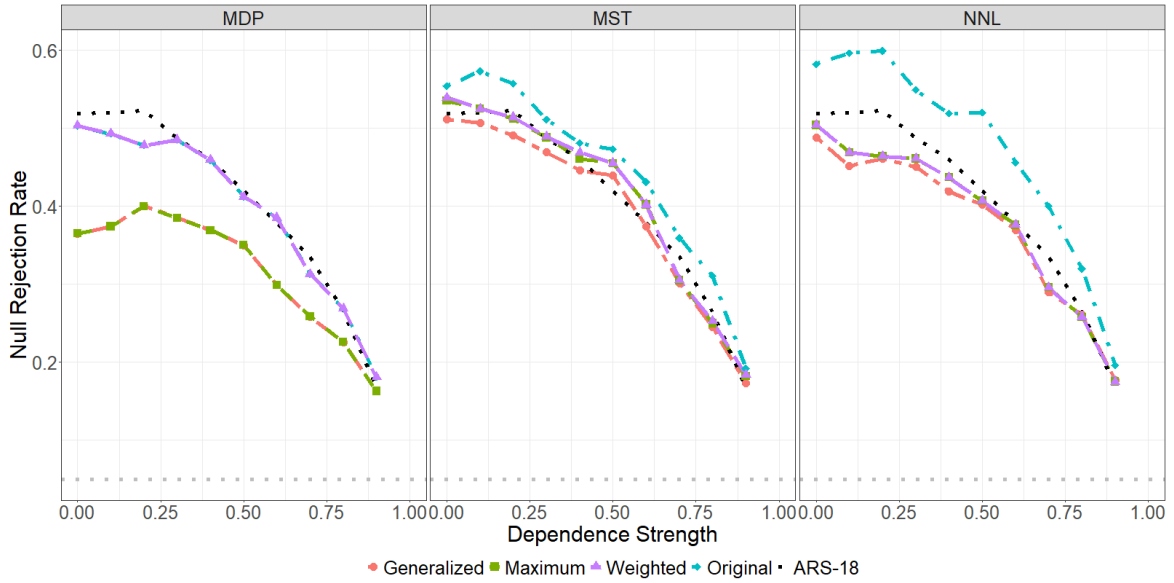


Figure 3.10: Size-adjusted mean change power curves under dependency. Power curves as a function of dependency magnitude based on mid-sample mean changes. Change magnitude = 0.5 for  $n = 50$  length data was considered. The dependency magnitude was varied from 0 to 0.9. The size-adjusted rejection rate is reported. Loss of power due to dependency is typically comparable between graph-based and ARS-18 methods, while MST and NNL performed better than MDP.

undetected in the data or remains in model residuals. Hence it is valuable to consider the robustness of the graph-based methods under dependency. Figure 3.10 examines the ability of the method to detect a mean change in data with dependence,  $\rho = 0; 0.1; \dots; 0.9$ . As expected, power decreased as the strength of the dependency increased when accounting for inflated size. The graph-based and ARS-18 methods exhibited similar power degradation over the increased dependence. The choice of graph played a significant role in the power, with MST and NNL more robust than MDP. We note that while we assumed independence, ARS-18 can be modified such that estimation is more robust under dependence.

Under a single change point, the graph-based approaches were well sized and exhibited good power under a variety of scenarios. The performance of graph-based change point detection methods were generally similar or better than existing functional change point detection methods. Robustness to tuning parameter selections and assumption violations was observed.

### 3.3.3 Multiple change simulations

While single change simulations are valuable to understand performance of a method due to a specific change, often multiple change simulations better reflect reality. We applied [binary segmentation \(BS\)](#) to detect multiple changes for the graph-based and existing functional change point detection methods. Using only [BS](#) highlighted the differences between approaches; however, other multiple change point extensions may improve the power for all approaches. In this section  $K = 5$  for orthogonal trees to mimic real-life scenarios when the structure of the data is largely unknown. Similar results were observed when  $K = 15$ . The max-type test statistic in [Eq \(3.11\)](#) is used for the graph-based methods due to its robustness in the single change scenarios. In each setting, only 100 simulations were run as this was sufficient to observe the detection trends.

Detected change locations for the graph-based and ARS-18 methods were recorded in the mean changed interval, or epidemic, scenario. Data of length  $n = 200$  were simulated with the mean changing from 0 to 0.75 to 0. The changes occurred at 60 and 170. 100 simulations were run. The true mean function and histograms of the detected locations are given in [Figure 3.11](#). The graph-based and ARS-18 methods typically detected changes in the neighborhood of the true changes. However, the graph-based methods had more early sample false detections and ARS-18 had more mid-sample false detections.

[Figure 3.12](#) considers 100 simulations of  $n = 250$  length samples containing 3 changes of different types. The graph-based and HRV-25 methods exhibited power on the first mean change. However, the graph-based methods exhibited more power for the other changes. Per [Figure 3.8](#), we suppose the power of the HRV-25 method would improve using a different basis in ARKL, e.g. a Fourier basis. The robustness of parameter choices is of note for the graph-based methods.

## 3.4 Applications

This section considers three real-world examples, related to pedestrian counts, stock returns, and electricity prices. For brevity, and due to its strong performance in [section 3.3](#), we only report the results from the max-type test statistic defined in [Eq \(3.11\)](#) based on MST-15 using  $L^2$  distance. Nominal significance of  $\alpha = 0.05$  is used. Multiple changes were detected using [binary segmentation](#). An unreported sensitivity analysis on the parameter choices showed robustness to these choices.

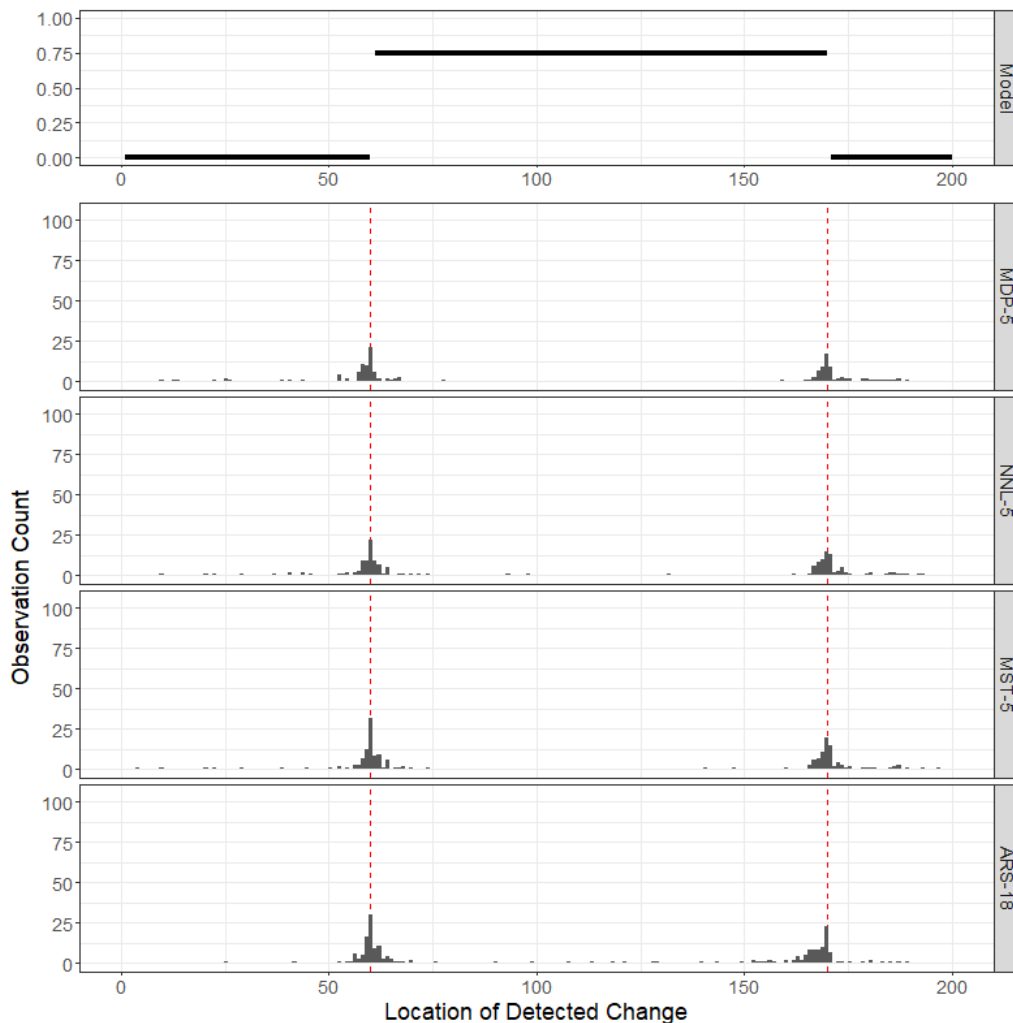


Figure 3.11: Multiple mean change location estimates. Estimates of mean change point locations based on data with 2 changes. For data of length  $n = 200$ , the magnitude change changed from 0 to 0.75 to 0 at locations  $k_1 = 60$  and  $k_2 = 170$ . The true mean model is given in the top plot. Histograms for the estimated change point locations from 100 simulations are given in the lower plots. Vertical red lines indicate the locations of the true changes. The graph-based methods and ARS-18 show robustness to the changed interval, or epidemic change, scenario. The graph-based methods had more early sample false detections while ARS-18 had more mid-sample false detections. The graph-based methods were generally comparable.

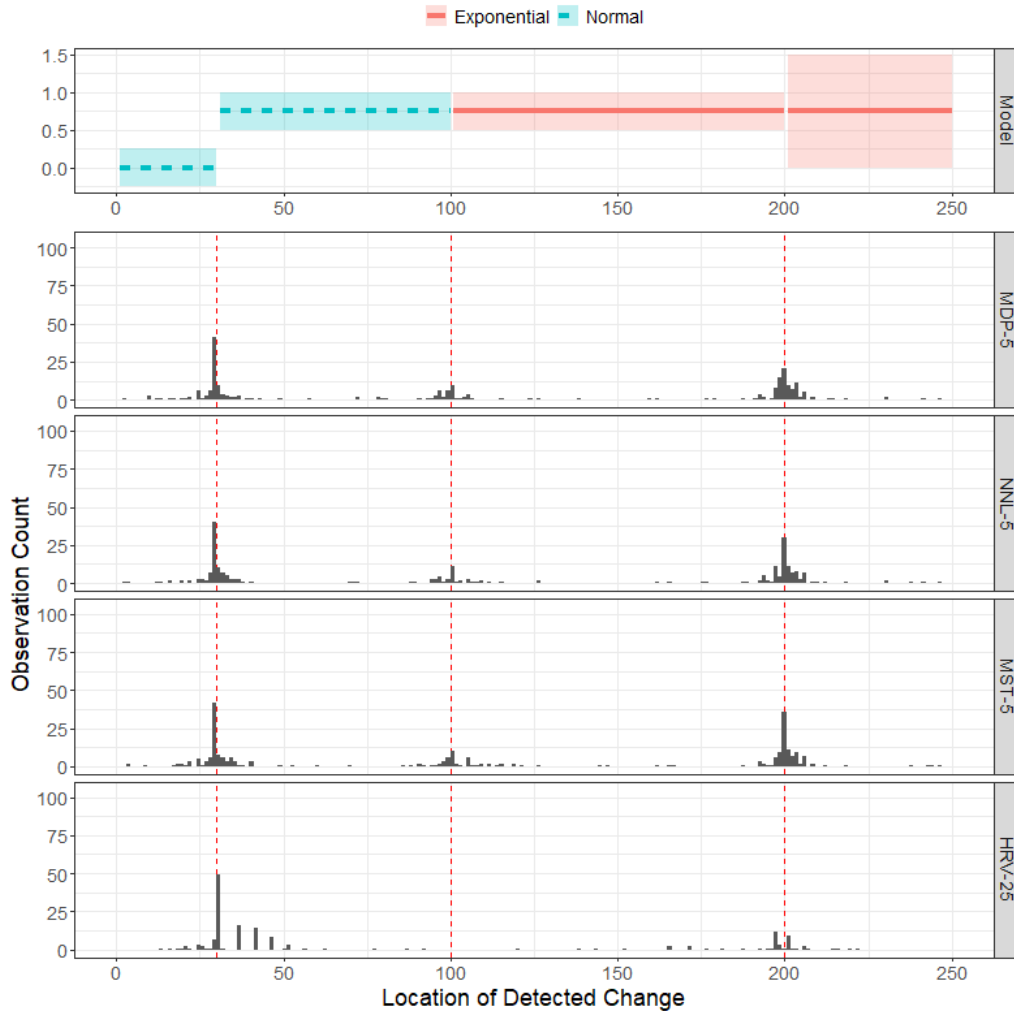


Figure 3.12: Multiple distributional change location estimates. Estimates of distributional change point locations based on data with 3 changes. For data of length  $n = 250$ , the data changed at locations  $k_1 = 30$ ,  $k_2 = 100$ , and  $k_3 = 200$ . A mean change  $\mu = 0.75$  occurred at  $k_1$ . A distributional change for the errors  $\epsilon$ 's changed from normal to exponential, both standardized, at  $k_2$ . A covariance change  $\Sigma = 3$  occurred at  $k_3$ . A representation of the true model is given in the top plot. Histograms for the estimated change point locations from 100 simulations are given in the lower plots. Vertical red lines indicate the locations of the true changes. The graph-based methods and HRV-25 commonly detected the first change, with HRV-25 perhaps performing slightly better. However, the graph-based methods detected the distributional and variance changes far more commonly.

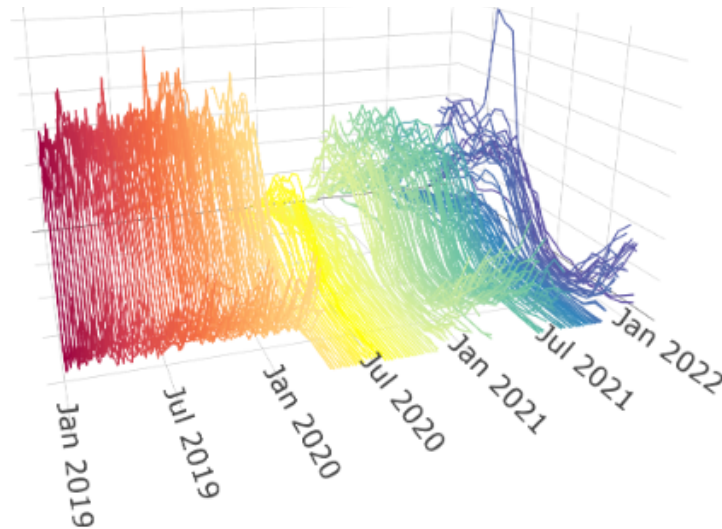


Figure 3.13: Melbourne pedestrian counts. Hourly pedestrian counts on a street in Melbourne Australia from 2019 to 2021.

### 3.4.1 Australian pedestrian counts

The city of Melbourne Australia uses a system of automated sensors to monitor information about pedestrian activity throughout the city. These observations are used to inform development and promote city vibrancy (City of Melbourne, 2024). We are interested in potential changes in pedestrian count data collected at one of the monitored crossroads from 1 January 2019 to 31 December 2021. The data were recorded every hour and are shown in Figure 3.13.

Figure 3.14 shows approximately homogeneous regions demarcated by the 7 detected changes with colored segments with differing line types. The detected changes seem to match the apparent distributional changes even before considering actual events. The first change is detected on 22 December 2019 ( $p < 0.01$ ), which shows a change of max daily pedestrian counts, perhaps due to the seasonal change. The next change is detected on 21 March 2020 ( $p < 0.01$ ), near the onset of COVID-19 and related lockdowns in Melbourne. Several changes follow throughout the year, mirroring the easing and strengthening of restrictions throughout the city. The change on 2 August 2020 ( $p < 0.01$ ), is near another lockdown with the onset of additional variants, followed by a change on 25 October 2020 ( $p < 0.01$ ). This October change is near the end of another round of lockdowns, at which point Melbourne had been under lockdown longer than any other city. The change

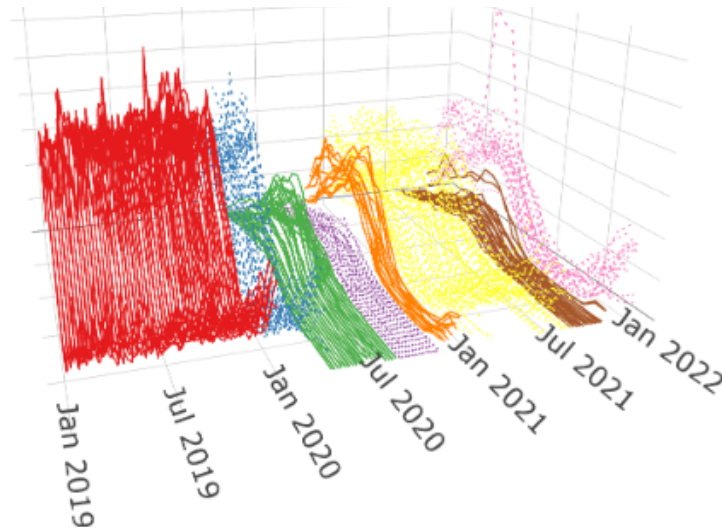


Figure 3.14: Segmented Melbourne pedestrian counts. Segmented hourly pedestrian counts on a street in Melbourne Australia (2019 - 2021). Colors and line types indicate approximately homogeneous regions defined by change points detected using MST-15 and the max-type test statistic in Eq (3.11) with  $L^2$  distance.

detected on 6 December 2020 ( $p = 0.02$ ) was at time where people were excited to enjoy the summer weather. A change is detected on 7 August 2021 ( $p = 0.01$ ), near another few short lockdowns in the Australian winter. The final change is detected on 24 October 2021 ( $p < 0.01$ ) as people again walk the streets in the summer.

### 3.4.2 Twitter stock returns

Stock prices are commonly thought of as [functional time series](#), due to the high frequency of trades and constant availability of prices. In this section, the now de-listed Twitter<sup>1</sup> stock data are considered. The data were recorded for every minute of trading hours from January 2019 to the end of December 2020. However, instead of directly studying the price data (which likely contains trends), intraday cumulative returns (CIDRs) are used ([Gabrys et al., 2010](#)).

If  $P_i(t_j)$ ,  $i = 1; \dots; n$ ,  $j = 1; \dots; m$  is the price of a financial asset at time  $t_j$  on day  $n$ ,

<sup>1</sup>X, after acquisition by Elon Musk on 28 October 2022

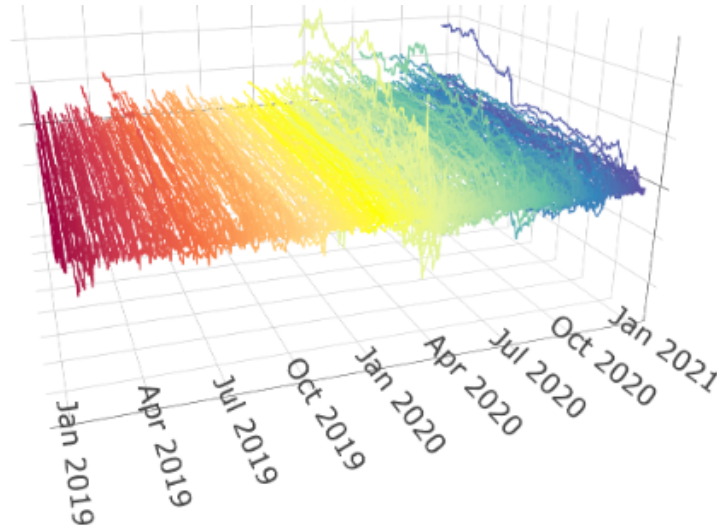


Figure 3.15: Twitter CIDRs. Twitter daily CIDRs (2019 - 2020), recorded at a one-minute resolution.

then the CIDR curves are defined as

$$r_i(t_j) = 100 [\log P_i(t_j) - \log P_i(t_1)]; \quad j = 1, \dots, m; \quad i = 1, \dots, n;$$

The CIDRs standardize the returns and ensure the starting value of each day is zero. Changes in mean are widely removed, and higher order changes are of primary interest in the CIDRs. The Twitter CIDRs are shown in Figure 3.15.

Approximately homogeneous regions in the daily CIDR Twitter data are colored and given different line types in Figure 3.16. The detected changes align well with known stock market events. The first change is on 31 October 2019 ( $p = 0.01$ ), which starts a period of relative stability and corresponds to a stock market surge and the lowering of the US federal funds rate. The next change on 4 February 2020 ( $p < 0.01$ ) aligns with stock market crash and early instability due to the onset of coronavirus 2019 (COVID-19) in the US. Likewise the change on 4 May 2020 ( $p < 0.01$ ) relates to the early lockdowns and continued instability of COVID-19. Both periods are defined by volatility reflecting investor sentiment and market conditions related to lock-down and work-from-home policies. The final change is on 17 December 2020 ( $p < 0.01$ ), around the time that the federal funds rate reached its lowest.

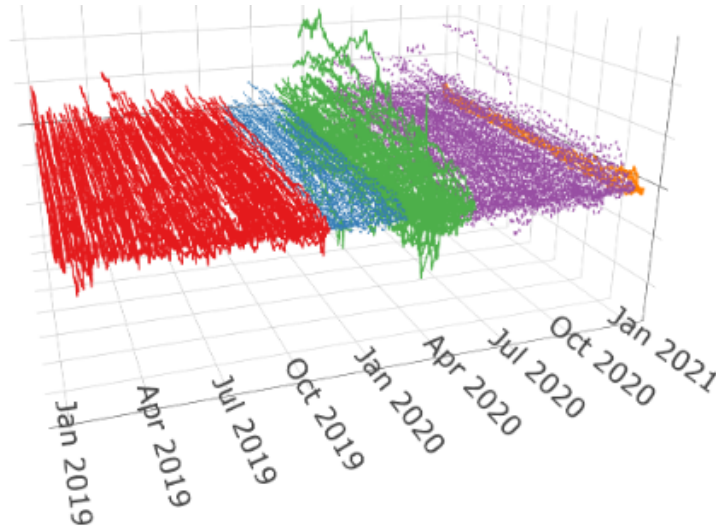


Figure 3.16: Segmented Twitter CIDs. Segmented daily Twitter CIDs (2019 - 2020), recorded at one-minute resolution. Approximately homogeneous regions are indicated by colors and line types. The regions are demarcated by estimated changes from the max-type graph-based test statistic defined in Eq (3.11) using a MST-15 graph and  $L^2$  distance.

### 3.4.3 Spanish electricity prices

Many FTS exhibit strong dependence between observations. Gonzalez et al. (2018), Mestre et al. (2021), and Horváth et al. (2025) consider hourly electricity prices for Spain in 2014. Such data exhibits natural trends; see Figure 3.17a. Hyndman and Ullah (2007) propose forecasting FTS through modeling components of functional principal component analysis. We term this approach  $HU(J)$ . The approach states that if data can be well-approximated by its projection onto  $J$  orthonormal basis functions then

$$X_i(t) \approx \bar{X}(t) + \sum_{j=1}^J a_{ij} \hat{f}_j(t); \quad t \in [0;1];$$

where the  $\bar{X}(\cdot)$  is the sample mean function,  $a_{ij}$  is the principal component score, and  $\hat{f}_j(\cdot)$  is the eigenfunction of the sample covariance operator.

The  $HU(J)$  model first requires approximation of the data, where  $J$  is selected according to some criteria. We used total variance explained (TVE); see Ramsay and Silverman (2005). We explored several values for TVE and observed little difference. Here we use

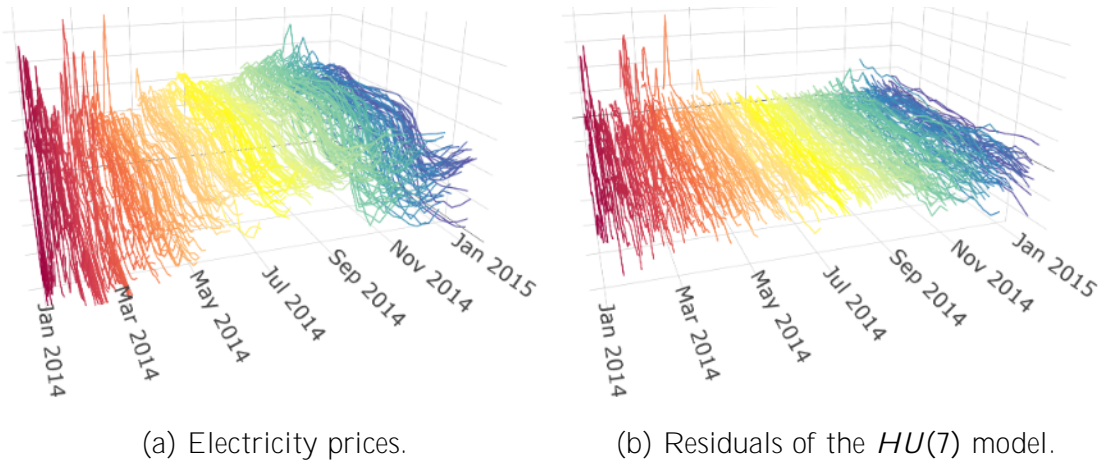


Figure 3.17: Electricity and its model residuals. Hourly electricity prices in 2014 for Spain and the  $HU(7)$  electricity model residuals. (a) The hourly electricity prices. (b) The residuals of the electricity prices when modeled using the  $HU(7)$  model with a Holt-Winters exponential smoothing model and weekly seasonality.

$J = 7$ , which explained 99% of the TVE, similar to that of Horváth et al. (2025). On each score series  $x_{ij}; i = 1; \dots; n$  for  $j = 1; \dots; J$ , a scalar time series model is fit. We used Holt-Winters exponential smoothing with a weekly seasonality fit as implemented in the forecast package in R; see Hyndman and Khandakar (2008) for details. The  $HU(J)$  model constructs the functional forecasts and fitted values as  $\hat{X}_i(t) = \hat{\mu}(t) + \sum_{j=1}^J \hat{\alpha}_{ij} \hat{f}_j(t)$ . Residual curves for the data can be computed as  $r_i(t) = X_i(t) - \hat{X}_i(t)$ . Following this, the residuals curves for the electricity prices are shown in Figure 3.17b.

Three change were detected on the  $HU(7)$  model residuals. These changes generally align with the seasons and are detected on the dates February 24 ( $p = 0.01$ ), May 4 ( $p < 0.01$ ), and September 30 ( $p < 0.01$ ). The segmentation of the residuals is indicated by different colors and line types in Figure 3.18. The residuals exhibit significantly more volatility in the winter models than in the summer months. These changes are similar to those detected in previous works, e.g. Horváth et al. (2025).

### 3.5 Discussion

Graph-based change point detection is an active area of research, and a valuable addition to the toolbox of functional time series analysis. Graph-based methods exhibited strong

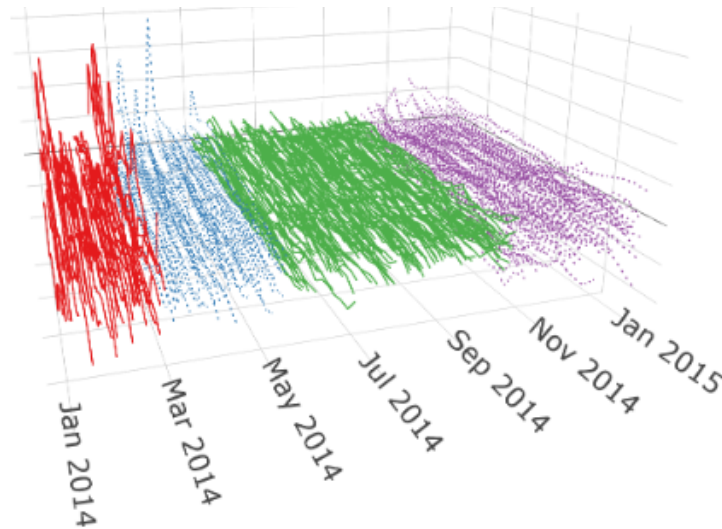


Figure 3.18: Segmented electricity residuals. Segmentation for the residuals of the  $HU(7)$  electricity model based on max-type graph-based change point detection using MST-15 and  $L^2$  distance. Regions of approximate homogeneity are indicated by line type and line color.

versatility and may be of valuable in many fields. In simulations, graph-based change point detection often provided significant improvement over existing functional methods. Even in low sample sizes, graph-based methods demonstrated sensitivity to various types of changes. They are also non-parametric and robust to violations of assumptions. Graph-based change point detection performed well for a number of tuning parameters, i.e. tree type, number of orthogonal trees, test statistics, and distance metric.

In this chapter, we offered several reasonable “default” tuning parameter choices and considerations when selecting these parameters. Even when the tuning parameters are not chosen optimally, the graph-based methods often exhibited power in a variety of scenarios. The graph-based methods returned reasonable segmentations when applied to multi-year pedestrian counts, high-resolution stock returns, and continuous electricity prices.

We hope graph-based methods will be increasingly used for functional time series and encourage further investigation. Graph-based methods are often computational fast and may find value in fields where data sizes are large. We also encourage examination into approximations for graphs and  $p$ -values to further improve on the computational efficiency of the methods. Investigations into dependent data and confidence intervals for change points detected via graph-based methods would also be valuable contributions.

# Chapter 4

## fChange: Functional Change Point Analysis in R

### 4.1 Introduction

To facilitate [functional data analysis \(FDA\)](#), specialized software continues to emerge in major programming languages, e.g. see FDApy ([Golovkine, 2021](#)) and scikit-fda ([Ramos-Carreño et al., 2024](#)) packages in Python ([Van Rossum and Drake Jr, 1995](#)), FDA ([Hein, 2016](#)) and ElasticFDA ([Tucker, 2018](#)) in Julia ([Bezanson et al., 2017](#)), or BFDA ([Yang and Ren, 2019](#)) in MATLAB ([The MathWorks Inc., 2024](#)), among others. As interest in FDA grows, general frameworks and packages designed to integrate FDA with established packages are becoming commonplace. One of the most well-known FDA packages is fda ([Ramsay et al., 2022](#)) in R and MATLAB. This general-purpose package relies on basis expansion to convert discrete observations into functional observations. Analysis is then typically done on the coefficients and combined with the basis functions to construct various models and perform inferences. The package fda.usc ([Febrero-Bande and Oviedo de la Fuente, 2012](#)) represents [functional data \(FD\)](#) in a discrete form and supplies several regression, classification, and non-parametric approaches. The recent packages funData ([Happ-Kurz, 2020](#)) and tidyfun ([Scheipl et al., 2024](#)) extend the use of FD to the popular family of packages in tidyverse ([Wickham et al., 2019](#)), including ggplot2 ([Wickham, 2016](#)). Other well-used R packages include those that focus on linear models, e.g. funGp ([Betancourt et al., 2024](#)) and refund ([Goldsmith et al., 2024](#)).

Despite the prodigious growth in software, general-purpose FD packages often overlook sequentially collected FD. When FD are collected sequentially, they are referred to as

functional time series (FTS). [Bosq \(2000\)](#) provides an introduction to FTS, while a more recent review can be found in Chapter 8 of [Kokoszka and Reimherr \(2017\)](#).

Many packages devoted to FTS are found in R. The package `ftsa` ([Hyndman and Shang, 2024](#)) provides a framework for FTS and functions such as functional [autoregressive integrated moving average](#). Visualization is a key component of `rainbow` ([Shang and Hyndman, 2022](#)). White noise tests for FTS are implemented in `wwntests` ([Kim and Petoukhov, 2023](#)) and `FTSgof` ([Kim et al., 2024](#)).

FTS applications often have periods of stability that are interrupted by some structural change. Investigation into such changes is included in the field of change point analysis. We refer readers to Chapter 8 of [Horváth and Rice \(2024\)](#) for a review of change point analysis with FTS.

Packages on change point detection for FTS are largely nonexistent. Currently, one must project FD into finite dimensions and detect changes using existing scalar or multivariate approaches. [Functional principal component analysis \(FPCA\)](#) is widely-used to project FD into finite dimensions. Implementations for FPCA include `fpca` ([Peng and Paul, 2011](#)) or `fdapace` ([Zhou et al., 2024](#)) (R), `PACE` ([Yao et al., 2015](#)) (MATLAB), `fdasrvf` ([Tucker, 2024](#)) (R, MATLAB, Python), or `ElasticFDA` ([Tucker, 2018](#)) (Julia). Once in finite dimensions, many packages exist for change point detection, e.g. see `change-point` for R ([Killick and Eckley, 2014](#)), `ruptures` for Python ([Truong et al., 2018](#)), or `Changepoints` for Julia ([Fairbrother et al., 2022](#)).

We present `fChange`, a library for FTS and change point detection in R. The package is available on CRAN and GitHub. Rather than fitting and storing the data with functions and coefficients, the package directly uses the discretized observations of functional objects that would be routinely collected. In addition to easy importation of data, the package includes example FTS data and allows simulation of FTS through various representations. Data processing and investigation methods are included, including specialized visualizations. Extensive implementations to characterize the data sets' stationarity, dependence, autocorrelation, and similarity to white noise are included ([Mestre et al., 2021](#); [Kim et al., 2023](#); [Kim and Petoukhov, 2023](#)). For change point detection, `fChange` includes projection- ([Berkes et al., 2009](#); [Aue et al., 2009](#); [Hušková and Meintanis, 2006](#)), tuned- ([Sharipov et al., 2016](#); [Bucchia and Wendler, 2017](#); [Aue et al., 2018, 2020](#); [Sharipov and Wendler, 2020](#); [Horváth et al., 2022](#); [Jiang et al., 2023](#); [Wegner and Wendler, 2024](#)), and general- ([Horváth et al., 2025](#)) change point detection methods. Extension from single to multiple change point detection are included via [binary segmentation](#) and [elbow plot](#) methods ([Scott and Knott, 1974](#); [Vostrikova, 1981](#); [Venkatraman, 1992](#); [Rice and Zhang, 2022](#)).

For each function, and for the package as a whole, comprehensive documentation and examples are provided. Further documentation can be found on the package’s website [jrvanderdoes.github.io/fChange](https://jrvanderdoes.github.io/fChange). The fChange functions are explored in subsequent sections using continuous electricity prices, cancer mortality, and long-term US interest rate examples. Section 4.2 introduces fChange’s representation of FTS data and useful functions innate to this representation. The functionality of methods in fChange are discussed in section 4.3, which is broken up into several subsections for clarity. Data simulation and input is discussed in section 4.3.1. A variety of visualizations are presented in section 4.3.2. Data characterization and change point analysis is respectively explored in sections 4.3.3 and 4.3.4. Section 4.4 concludes.

## 4.2 Data representation

In fChange, data are stored as discretized functions of the form  $X_i(t)$  for  $i = 1; \dots; n$  and  $t \in \mathbb{R}^+$ . For ease, we refer to  $X(t) = (X_1(t); \dots; X_n(t))$  as the data,  $t$  as the functional parameter, and some element  $X_i(t)$  for  $i \in \{1; \dots; n\}$  as an observation. While stored discretely, the data are treated as functions throughout the package.

### 4.2.1 The dfts class

All functions in fChange use data organized into dfts (discretely observed functional time series) class objects. This standard interface simplifies function calls and ensures that all necessary attributes are included. The class contains the following attributes:

- name: A string for the name of the functional time series.
- data: A matrix of size  $r \times n$  where each column is the functional observation  $i$ , for  $i = 1; \dots; n$  and the rows indicate the time points  $t \in [a; b]$ .
- labels: A vector of the length  $n$  for the observation names, e.g. numbers or dates.
- fparam: A vector of length  $r$  for the discrete points  $t$ ,  $t \in [a; b]$  where the observations are evaluated.

The dfts() function ensures the data are complete and well-formatted, including using a format that minimizes run times in fChange functions. All attributes besides data can

be user-specified or filled by default. Any 2-dimensional data that can be cast as a matrix or data.frame can be used as the data to create a dfts object. Further, conversions from popular [functional data](#) classes such as fda's fd, ftsc's fts, ftsc's fts, fda.usc's fdata, and funData's funData objects are automatic.

When only the parameter data is given, name is inferred from the data variable name, labels are inferred from the column names, or equivalent, of the input data, and fparam is computed as  $r$  evenly spaced points on the  $[0; 1]$  interval where  $r$  is the resolution—typically the number of rows in the data. In general, each observation does not need to be observed at evenly spaced fparam points. However, it is assumed that all observations are observed the same fparam points, which may include missing values for some observations. If this assumption is violated, the data can be pre-processed for use as a dfts object. Additional details are discussed in section [4.3.1](#).

## 4.2.2 Functions of the dfts class

Many mathematical functions are extended for dfts objects to simplify data manipulation. For example, mathematical operations like addition and multiplication are directly applied to the data attribute. More complex functions such as mean(), cumsum(), center(), sd() and var() are also included. The function mean() returns the pointwise mean at each point of the functional parameter fparam. The function center() removes the mean, or median, of the data pointwise. The variance function, var(), can return a pointwise or operator variance.

The function summary() provides an overview of a data set. It returns a conglomerate figure and descriptive information. An example overview is provided in [Figure 4.1](#). The descriptive information includes the length, resolution, and the p-values from the KPSS and stationarity tests on the data. The plots and information can also account for specified change points.

The electricity spot price in Spain is included in fChange as electricity. The electricity prices are measured at hourly resolution throughout 2014. The following code and [Figure 4.1](#) is used to summarize the electricity data.

```
R> summary(electricity)
```

| Segment | Observations | kpss  | stationarity | Resolution |
|---------|--------------|-------|--------------|------------|
| 1-365   | 365          | 0.000 | 0.000        | 24         |

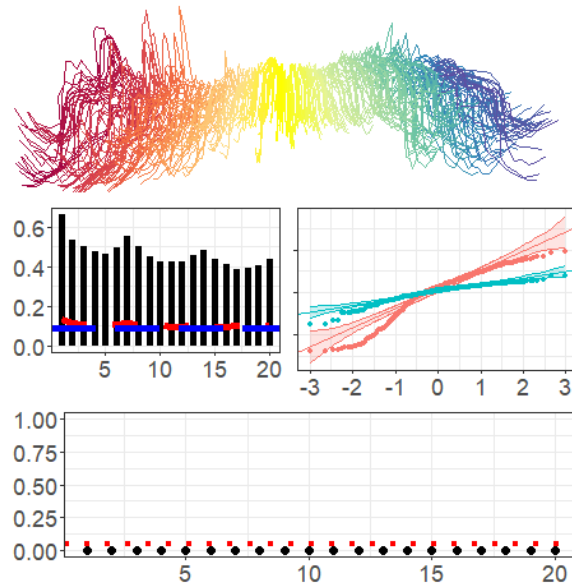


Figure 4.1: Electricity summary. Example view of electricity from the function summary(). The top plot is a spaghetti plot of the data. The middle left plot is an ACF plot with 20 lags, where the blue line is the threshold under the strong white noise assumption and the red line is the threshold under the weak white noise assumption. The middle right is a QQ plot for the first 2 **functional principal components**. The bottom plot displays the  $p$ -values for white noise tests up to 20 lags. The red dotted line is the nominal size,  $\alpha = 0.05$ .

The summary figure has several subplots. At the top, a rainbow spaghetti plot is given for general reference. The middle left plot displays the functional autocorrelation function of the data (Mestre et al., 2021). 20 lags are shown by default. The blue line indicates the threshold under strong white noise and the red under weak white noise. The middle right plot displays a QQ-plot displaying the distributions for the leading principal component scores. By default 2 components are used. The bottom plot computes a joint white noise test up to some maximum number of lags and displays the corresponding  $p$ -values (Kim et al., 2023). The dotted red line is the nominal size. 20 lags are considered by default with nominal size  $\alpha = 0.05$ . The summary plot is modified when change points are indicated. When changes are specified, each segment can be investigated or the demeaned data can be considered; see parameters changes and demean. More detailed visualizations for data are discussed in section 4.3.2.

## 4.3 Functionality of fChange

The fChange package provides extensive tools for preparing and analyzing [FTS](#), with a focus on change point analysis. In this section, several key functions are investigated with illustrative examples.

### 4.3.1 Data

Data can be imported into fChange by using `dfts()` on any `matrix`, `data.frame`, or other 2-dimension object which inherits from these classes as well as from other popular representations; see section 4.2. Native fChange data sets include `electricity`, `cancer`, and `rates` which are discussed below. An explicit call to the `dfts()` is written as:

```
R> cancer <- dfts(cancer)
```

The `electricity` data comprises the 2014 hourly spot prices of electricity in Spain as reported by the market operator OMIE. The `rates` data are the US daily interest rate curves from 1990 to 2023, with rates unevenly spaced from 1 month to 360 months retrieved from the US Department of the Treasury. The `cancer` data are the US female breast cancer mortality percentages (out of total deaths) from 1950 to 2021 reported by the World Health Organization.

Data generating processes (DGPs) are also included in fChange through the family of functions `generate_*`, e.g. `generate_brownian_motion()` or the `generate_data()` wrapper function. The implementations in fChange extend beyond that observed in other packages such as `sde` ([Iacus, 2022](#)) by allowing for evenly and unevenly spaced observation times. General information on three common DGPs are discussed below. Further information for these, as well as other DGPs, are left to the package documentation.

Auto-Regressive Truncated Karhunen-Loève Process (ARKL): Akin to the work by [Aue et al. \(2018\)](#), the ARKL process is defined as

$$X_i(t) = \alpha_i(t) + \sum_{j=1}^d \beta_{ij} \phi_j(t); \quad t \in [0;1]; \quad (4.1)$$

where  $(\phi_1, \dots, \phi_d)$  are the first  $d$  orthonormal basis functions, and  $\beta_{ij} = (\beta_{i1}, \dots, \beta_{id})^T$  is a vector that follows a vector autoregression of order one, so that

$$\beta_{ij} = \Psi \beta_{i,j-1} + \epsilon_{ij}$$

The matrix  $\Psi$  is a random  $d \times d$  matrix with [independent and identically distributed \(IID\)](#) standard normal entries that is normalized by a constant so that  $k\Psi k_F = 1$ , where  $k k_F$  denotes the Frobenius norm. Hence  $\Psi$  describes the magnitude of serial dependence in the process. Changes to the mean ( $\mu$ ), eigenvalues ( $\text{var}(\lambda)$ ), or distribution ( $f(\lambda)$ ) can be directly implemented.

In fChange, ARKL can be generated from `generate_karhunen_loeve()` or the general function `generate_data()` using the ‘kl’ process details. The robustness of this method allows many possible changes that can be specified by input parameters. For brevity and clarity, we direct the reader to the package examples and documentation.

**Auto-Regressive Ornstein–Uhlenbeck Process (AROU):** The AROU process is defined as

$$X_i(t) = \mu_i(t) + \rho X_{i-1}(t) + O_i(t); \quad t \in [0;1]; \quad (4.2)$$

where  $fO_i; i \in \mathbb{Z}; t \in [0;1]g$  are [IID](#) standard Ornstein–Uhlenbeck processes, which are Gaussian processes with mean zero and covariance  $E O_i(t) O_j(s) = \exp(-\lambda |t-s|)$ . An AROU process can be generated from `generate_ornstein_uhlenbeck()` or the general function `generate_data()` using the ‘ou’ process details, e.g.

```
R> generate_ornstein_uhlenbeck(N = 100,
> resolution = c(0, 0.5, 0.75, 1))
```

**Functional autoregressive model of order one (FAR(1)):** The FAR(1) process, as described in [Bosq \(2000\)](#) and [Horváth and Kokoszka \(2012\)](#), is defined as

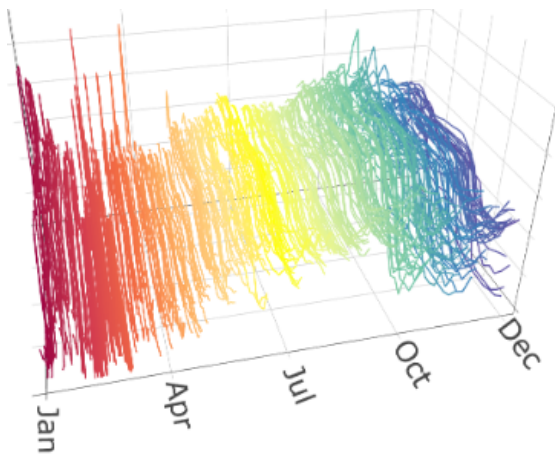
$$X_i(t) = \mu_i(t) + \int_0^1 \gamma(s;t) X_{i-1}(s) ds + W_i(t); \quad t \in [0;1];$$

where  $\gamma(s;t) = 1.3390 \exp(-\lambda(s^2 + t^2))$ , and  $fW_i(t); i \in \mathbb{Z}; t \in [0;1]g$  are [IID](#) standard Brownian motions. The FAR(1) data generating process can be simulated using `generate_far1()` or `generate_data()` with the ‘far1’ process details.

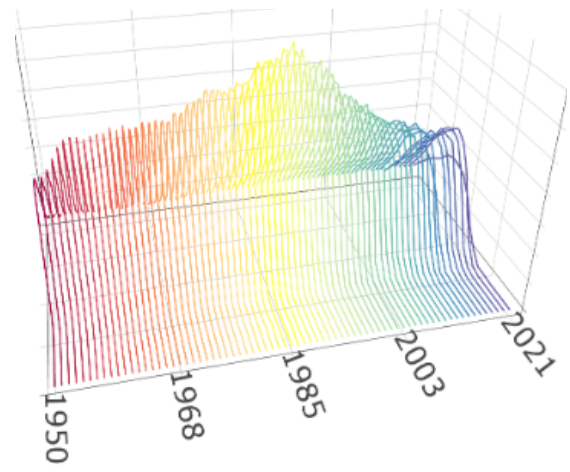
```
R> generate_brownian_motion(N = 100)
```

### 4.3.2 Visualizations

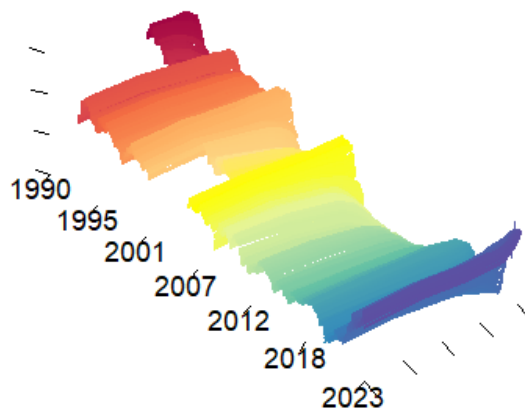
Often a more focused view of the data than given by `summary()` is desired. Using `plot()` on `dfts` objects can create a number of customizable visualizations. Several plots are



(a) Electricity prices



(b) Breast cancer mortality



(c) US interest rates

Figure 4.2: Spaghetti plots. Rainbow spaghetti plots for electricity, cancer, and rates.

presented in this chapter. The value of the `changes` parameter which contains change point locations is shown in section 4.3.4.

The default view of `FTS` in `fChange` is a *rainbow spaghetti plot*. This plot has lines in a 3-dimensional environment which represent the observations. An example was given in Figure 4.1 by the `summary()` function. By default, the plot is interactive in `plot()` when `type` is set as `'spaghetti'`. However, with many observations it can be computationally burdensome and hence setting `type` to `'fast'` is recommended as the number of observations

$n$  or the resolution  $r$  becomes large.

```
R> plot(electr icity)
R> plot(cancer)
R> plot(rates, type = 'fast')
```

The resultant figures are given in Figure 4.2. By default, the plot is titled with the `dfts` parameter name and the axes are labeled according to `labels` and `fparam`. Additional parameters for these plots which clarify the data under change points are explored in section 4.3.4.

The `plot()` function can also be used to create 2-dimensional plots. A rainbow plot stacks the observations in 2-dimensions, akin to `rainbow` (Shang and Hyndman, 2022). Similarly, a banded plot mitigates significant overlap in rainbow plots by plotting bands rather than every observation. These plots are computed by respectively setting `type` to 'rainbow' or 'banded'.

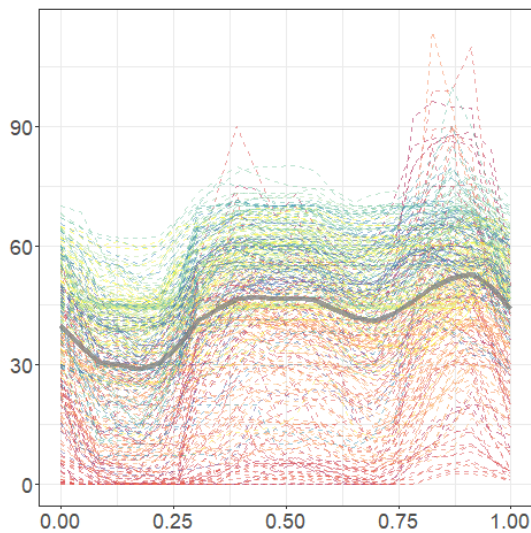
```
R> plot(electr icity, type = 'rai nbow')
R> plot(cancer, type = 'rai nbow')
```

Banded plots are particularly useful under change points and hence only rainbow plots are given in Figure 4.3.

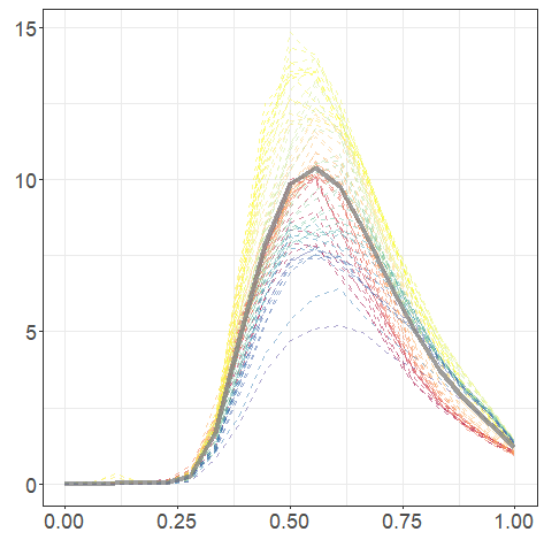
### 4.3.3 Data processing and diagnostic tests

Since many `FTS` contain at least some missingness, several imputation functions are included with `fChange` in the function `impute()`. For example, we use linear imputation on `rates` for this work; however a sensitivity analysis showed that different imputation methods did not significantly influence the subsequent analyses. For data with a more complex missing structure, the flexibility of `fChange` permits data to be imputed using other packages and then easily imported. Discussion on included methods can be found in the documentation and more sophisticated approaches for `FD` imputation are left to other works, e.g. Rao and Reimherr (2021).

Evaluating `FTS` for stationarity and the presence of serial correlation are valuable for analysis. There is extensive literature on white noise and stationarity testing; e.g. see Gabrys and Kokoszka (2007), Horváth and Kokoszka (2012), Zhang (2016), Kokoszka et al. (2017), and Kim et al. (2023). Implementations can be found throughout `fChange` such



(a) Electricity prices



(b) Breast cancer mortality

Figure 4.3: 2-dimensional rainbow plots. 2-dimensional rainbow plots for electricity and cancer. The thin, dotted, and colored lines are the observations. The thick gray lines are the pointwise means of the data.

as `stationarity_test()`, `portmanteau_tests()`, `kpss_test()`, and `acf()` or `pacf()` functions.

Test statistics for stationarity, white noise, and even change point hypothesis tests often rely on the functional partial sum process,

$$S_{n;X}(x; t) = \frac{1}{n} \sum_{i=1}^{bnxc} X_i(t) \quad (4.3)$$

where  $x \in [0; 1]$ . This can be used to define the **cumulative sum (CUSUM)** process

$$Z_{n;X}(x; t) = S_{n;X}(x; t) - xS_{n;X}(1; t) \quad (4.4)$$

Let  $\lambda_1, \lambda_2, \dots$  be the eigenvalues of the long-run covariance function such that

$$\lambda_i \int_0^1 d(t; s) \lambda_i(s) ds; \quad i = 1; 2; \dots; \quad (4.5)$$

where

$$d(t; s) = \lim_{n \rightarrow \infty} n \text{Cov}(X(t); X(s));$$

for the sample mean

$$X(t) = \frac{1}{N} \sum_{i=1}^N X_i(t); \quad (4.6)$$

See chapter 8 of [Horváth and Rice \(2024\)](#). Estimated eigenelements are defined as

$$\hat{\lambda}_i \hat{v}_i(t) = \int \hat{d}(t; s) \hat{v}_i(s) ds; \quad \text{for } i = 1; 2; \dots; \quad (4.7)$$

for estimated kernel  $\hat{d}(t; s)$ . Estimation of the long-run covariance and its spectrum is often done using a kernel bandwidth long-run covariance estimator, i.e.

$$\hat{d}(t; s) = \sum_{h=1}^{\infty} K\left(\frac{h}{q}\right) \hat{\gamma}_h(t; t^0)$$

for bandwidth parameter  $q$ , kernel  $K$ , and

$$\hat{\gamma}_h(t; t^0) = \frac{1}{n} \sum_{i=1}^n [X_i(t) - X(t)][X_{i+jh_j}(t^0) - X(t^0)];$$

where  $X(t)$  is the mean estimate in Eq (4.6). See [Aue et al. \(2018\)](#) and [Rice and Shang \(2017\)](#) for additional details. In fChange, the long-run covariance estimator can be run with `long_run_covariance()`. Estimation requires the use of some kernel function and bandwidth. The default kernel used throughout the package is the Bartlett kernel,

$$K(u) = \begin{cases} 1 - |u| & |u| \leq 1 \\ 0 & \text{otherwise} \end{cases};$$

Other common kernels are also implemented, such as: truncated, Parzen, Tukey-Hanning, quadratic spectral, Daniell, and flat top kernels. Additionally, a user-defined kernel can be supplied to the functions. We leave kernel definitions and further discussions on the choice of kernel to works such as [Politis \(2003\)](#) and [Horváth and Rice \(2024\)](#). The choice of bandwidth is often of primary interest when using kernels as any reasonable kernel can achieve close to optimal results ([Epanechnikov, 1969](#)). The bandwidth can be chosen manually or adaptively; see function `adaptive_bandwidth()` ([Rice and Shang, 2017](#)). By default the bandwidth is set to  $2n^{1/5}$ .

### 4.3.3.1 Projection

It is common practice to reduce the dimension of a **FTS**. Such dimension reduced models are often constructed by projecting the functional observations into finite dimensions and then applying a scalar or multivariate model on the resulting finite dimensional data. Perhaps the most common approach to project **FTS** into finite dimensions is **FPCA**; e.g. see [Berkes et al. \(2009\)](#) and [Aue et al. \(2009\)](#).

**FTS** can be represented as

$$\mathbf{X}'_i(t) = \int c(t; s) \mathbf{X}'_i(s) ds; \quad \text{for } 1 \leq i < I; \quad (4.8)$$

where  $\lambda_1; \lambda_2; \dots$  are eigenvalues and  $\phi_1; \phi_2; \dots$  are the eigenfunctions of the covariance operator of the data, i.e.

$$c(s; t) = \text{Cov}(X(t); X(s))$$

Estimated eigenelements are defined

$$\hat{\mathbf{X}}_i(t) = \int \hat{c}(t; s) \phi_i(s) ds; \quad \text{for } i = 1; \dots; n; \quad (4.9)$$

for

$$\hat{c}(t; s) = \frac{1}{n} \sum_{i=1}^n [X_i(t) - \bar{X}(t)][X_i(s) - \bar{X}(s)];$$

and the estimated mean defined in Eq (4.6).

Intuitively, **FPCA** iteratively projects infinite-dimensional data, such as **FTS**, onto directions that contain the most variability within the data. That is,

$$\hat{\mathbf{X}}_{i^*} = \int [X_i(t) - \bar{X}(t)] \phi_i(t) dt; \quad i = 1; \dots; n; \text{ and } i^* = 1; \dots; K \quad (4.10)$$

onto some dimension  $K$ . We refer to  $\hat{\mathbf{X}}_{i^*}$  as the (functional) principal component scores. Since the directions maximize the variability explained, the leading **functional principal components (FPCs)**,  $\phi_i(t)$ , can capture a significant portion of the “information” from the data. In practice, the first  $K$  **FPCs** explain the majority of the variability in the function and can be used to reduce the dimensionality of the data.

A common approach to select  $K$  is [total variance explained \(TVE\)](#). For  $\text{TVE } v \in [0; 1]$ , the number of components recommended is

$$K = \min \left\{ K : \frac{\sum_{j=1}^K \hat{\lambda}_j}{\sum_{j=1}^7 \hat{\lambda}_j} > v \right\} \quad (4.11)$$

where  $\hat{\lambda}$ 's are the empirical eigenvalues in Eq (4.9).

The function `pca()` returns estimates for the properly scaled eigenvalues, eigenvectors, and eigenscores of the `dfts` object. The parameter `TVE` specifies the maximum permitted loss in this projection.

```
R> pca(cancer, TVE = 0.95)
```

#### 4.3.3.2 Stationarity test

The stationarity test proposed by [Horváth et al. \(2014\)](#) considers the null hypothesis

$$H_0 : X_i(t) = \mu_i(t) + \epsilon_i(t) \quad (4.12)$$

versus the alternative

$$H_0 : X_i(t) \notin \mu_i(t) + \epsilon_i(t) : \quad (4.13)$$

More specific alternatives are also considered in [Horváth et al. \(2014\)](#).

It is common to test such hypotheses by using an integrated-type or supremum-type test statistic. The integrated-type test statistic is defined as

$$T_n = \iint Z_{n;X}^2(x; t) dt dx;$$

while the supremum-type test statistic is defined as

$$M_n = \iint \left( Z_{n;X}(x; t) - \int Z_{n;X}(y; t) dy \right)^2 dt dx;$$

When the data are stationary and suitably weakly dependent, it can be shown that asymptotically

$$T_n \xrightarrow{P} \sum_{i=1}^7 \int B_i^2(x) dx \quad (4.14)$$

and

$$M_n \overset{P}{\rightarrow} \sum_{i=1}^d \int \left( B_i(x) - \int B_i(y) dy \right)^2 dx; \quad (4.15)$$

where  $\overset{P}{\rightarrow}$  denotes convergence in distribution as defined in Billingsley (1968),  $B_1; B_2; \dots$  are independent standard Brownian bridges, and  $\lambda_1; \lambda_2; \dots$  are the eigenvalues defined in Eq (4.5). These limits can be approximated by Monte Carlo simulation taking the summation to  $d$ , when  $d$  is suitable large and estimating the eigenvalues  $\lambda_i$  as defined in Eq (4.7). Alternatively, (block) resampling can be used to threshold the statistic.

```
R> stationarity_test(electricity)$pvalue
R> stationarity_test(cancer, critical = 'resample',
> blocksize = 3)$pvalue
R> stationarity_test(rates, statistic = 'Mn')$pvalue
```

The stationarity test using the critical values defined from the asymptotic distribution in Eqs (4.14) and (4.15) was applied to the three data sets. Each was shown to exhibit significant non-stationarity with the  $p$ -values from the above code respectively  $p < 0.001$ ,  $p = 0.002$ , and  $p = 0.022$  (output suppressed).

For each hypothesis test in fChange, both the integrated-type and supremum-type test statistics are implemented. For brevity below, we only present the integrated-type test statistics, hereafter also referred to simply as “test statistics”. Unreported investigations into the supremum-type test statistics returned generally comparable results.

#### 4.3.3.3 KPSS test

A functional Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test, which expands the test from scalar data as originally presented by Kwiatkowski et al. (1992), is another way to evaluate the degree of stationarity in some FTS. Kokoszka and Young (2016) consider the trend stationary null hypothesis

$$H_0 : X_i(t) = \mu(t) + \lambda_i(t) + \epsilon_i(t) \quad (4.16)$$

compared to the random walk alternative hypothesis

$$H_0 : X_i(t) = \mu(t) + \lambda_i(t) + \sum_{j=1}^i u_j(t) + \epsilon_i(t) : \quad (4.17)$$

See also [Chen and Pun \(2019\)](#).

The null hypothesis defined in Eq (4.16) can be compared to Eq (4.17) by the test statistic

$$R_n = \iint Z_{n;X}^2(x; t) dt dx$$

where

$$Z_{n;X}(x; t) = S_n(x; t) - \frac{bnxc}{n} S_n(1; t) - \frac{1}{2} n^{3/2} \hat{\Lambda}(t) \left( \frac{bnxc}{n} \left( \frac{bnxc}{n} - 1 \right) \right)$$

with  $S_n(x; t)$  as defined in Eq (4.3),

$$\hat{\Lambda}(t) = \frac{1}{S_n} \sum_{i=1}^n \left( i - \frac{n+1}{2} \right) X_i(t); \quad \text{and} \quad S_n = \sum_{i=1}^n \left( i - \frac{n+1}{2} \right)^2 :$$

Under stationarity, it can be shown that the test statistic asymptotically behaves such that

$$R_n \xrightarrow{P} \sum_{i=1}^{\infty} \lambda_i \int V_i^2(x) dx$$

where  $\lambda_1; \lambda_2; \dots$  are the eigenvalues from Eq (4.5),  $V_1; V_2; \dots$  are IID second-level Brownian bridges,

$$V(x) = W(x) + (2x - 3x^2)W(1) + (6x - 6x^2) \int_0^1 W(y) dy;$$

and  $fW(x); 0 \leq x \leq 1$  are standard Brownian motions. Block resampling can also be used to threshold the test statistic.

```
R> kpss_test(electri ci ty)$pval ue
R> kpss_test(cancer, cri ti cal = 'resampl e', h=3)$pval ue
R> kpss_test(rates)$pval ue
```

On the three data sets, the stationarity test is highly significant. The same is not true for the KPSS test, which is highly significant for el ectri ci ty and cancer (both  $p < 0.001$ ) but insignificant for rates.

#### 4.3.3.4 Mitigate stationarity violations

To mitigate stationarity violations, two common approaches are differencing and detrending. When an FTS  $X_i(\cdot)$  is differenced, a new FTS  $Y(\cdot)$  is defined as the  $d$ th difference of the data  $X(\cdot)$ , i.e.

$$Y_i(t) := (1 - B)^d X_i(t)$$

where  $B$  is the backshift operator. Differencing electricity leads to approximately stationary data, per the stationary and KPSS tests.

```
R> electricity_diff <- diff(electricity, lag = 1, differences = 1)
```

Detrending attempts to capture, and then remove, the dependency structure of a time series with the goal that the residuals of detrended time series are approximately stationary. Hyndman and Ullah (2007) present an approach to detrend FTS by projecting the data into finite dimensions, fitting scalar models on each component, and recombining the results to estimate a functional trend. We denote this method  $HU(J)$ , where  $J$  indicates the number of components used. It is common to define and select projection components by functional principal components analysis; see section 4.3.3.1,

The  $HU(J)$  model is implemented in `projection_model()`, where  $J$  can be selected via TVE. The function returns information such as the fit, residuals, and forecasts. The cancer data was modeled using  $HU(5)$ , where 5 explained at least 99% TVE. Each component was modeled using an error-trend-seasonality (ETS) model as implemented in Hyndman and Khandakar (2008). According to the stationarity and KPSS tests, the residuals of the model is approximately mean stationary. Figure 4.5b shows the model residuals.

```
R> cancerresiduals <- projection_model(cancer, model = 'ets',  
>   TVE = 0.99)$residuals
```

Other methods of detrending, such as using a linear model, can be computed in existing packages and imported into fChange.

#### 4.3.3.5 White noise tests

A white noise FTS is a stationary FTS with zero autocorrelation. Determination of white noise can inform on the goodness of fit or utility of some model. The package fChange

contains several statistics to detect if a FTS is white noise, including those based on ‘single-lag’, ‘multi-lag’, ‘spectral’, ‘independence’, and ‘imhof’ methodologies. For brevity, the details of these tests are left other works, e.g. [Gabrys and Kokoszka \(2007\)](#), [Horváth et al. \(2013\)](#), [Characiejus and Rice \(2020\)](#), and [Kim et al. \(2023\)](#).

The white noise tests are implemented in the `portmanteau_tests()`. Applying a variety of the white noise tests on `electricity_diff` and `rates` reject the plausibility of white noise with  $p < 0.001$ .

```
R> portmanteau_tests(electricity_diff, test = 'variety')
```

The tests reveals that `cancerresiduals` are plausibly white noise ( $p > 0.1$ ); see [Figure 4.5b](#).

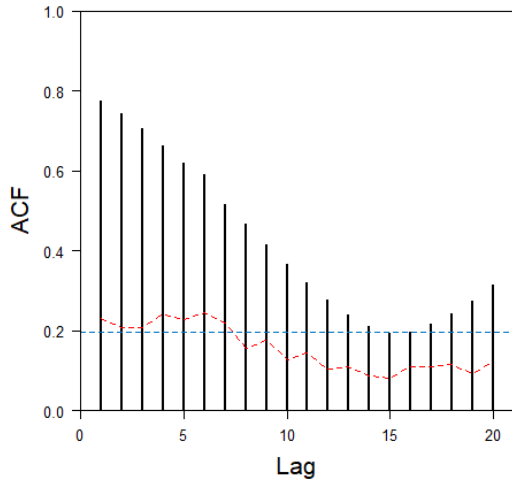
Functional autocorrelation (ACF) and functional partial autocorrelation (PACF) functions can also be used to assess the degree of whiteness in data ([Kokoszka et al., 2017](#)). We leave the details of this analysis to other works beyond noting that implementations for the functional autocorrelation and the functional partial autocorrelation functions are available using `acf()` and `pacf()`. Both tests produce thresholds under the assumptions of strong and weak white noise.

The ACF of `cancer` and `cancerresiduals` are computed below, with the resultant figures given in [Figure 4.4](#). The ACF plots suggests that there is not strong evidence to reject the white noise assumption for the  $HU(5)$  model residuals.

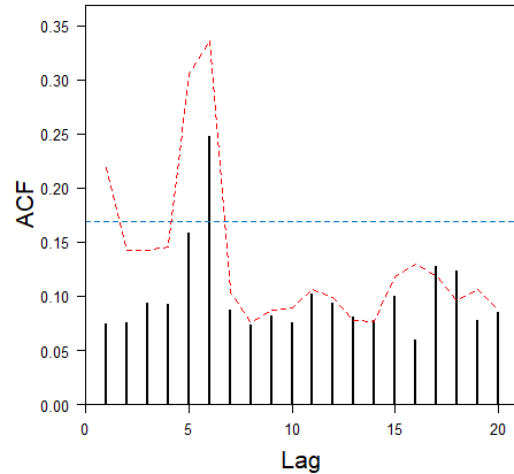
```
R> acf(cancer)
R> acf(cancerresiduals)
```

#### 4.3.4 Change point detection

Literature on functional change point detection has grown prodigiously in recent years, and `fChange` offers an expansive array of change point detection methods. Both fully functional and projection-based methods are included. The change point detection methods in `fChange` can broadly be grouped into three categories: mean, second-order, and general. Many methods are based on the [CUSUM](#) process from [Eq \(4.4\)](#). For each method, both integrated-type and supremum-type test statistics defined; however, we only report integrated-type statistics in this document for brevity. The critical values for these statistics can generally be computed using simulation, resampling, or Welch-form approximations.



(a) ACF of cancer data



(b) ACF of cancer model residuals

Figure 4.4: Autocorrelation functions. Functional autocorrelation functions computed from cancer and cancerresiduals. The threshold under strong white noise is given in blue and the threshold under weak white noise is given in red. (a) The cancer data exhibits strong serial correlation. A large number of violations are observed when compared to the assumption of strong and weak white noise. (b) The cancerresiduals exhibit no strong serial correlation. Only a lag 6 spike is observed, which is below the weak white noise threshold. The spike may be due to chance, or an indication that the weak white noise assumption is more appropriate.

For ease, the user interface for all change methods is through the function `fchange()`. In `fchange()`, the parameter `method` specifies the testing method, `statistic` specifies the test statistic (integrated-type as ‘Tn’ or supremum-type as ‘Mn’), and `critical` specifies the approach for determining the critical value (‘simulation’, ‘resample’, or ‘welch’). We note that the ‘resample’ option can be used to compute the critical values via bootstrapping or permutation based on the the parameter `replace`. Other parameters are discussed when appropriate.

#### 4.3.4.1 Mean change

Functional mean change methods aim to test the null hypothesis

$$H_0 : E[X_1(\cdot)] = \dots = E[X_n(\cdot)] \quad (4.18)$$

where equality is understood in the  $L^2$ -sense. Detection methods aim to identify violation of Eq (4.18), for example the [at most one change \(AMOC\)](#) alternative is

$$H_A : E[X_1(\cdot)] = \dots = E[X_k(\cdot)] \neq E[X_{k+1}(\cdot)] = \dots = E[X_n(\cdot)] \quad (4.19)$$

for some change point  $k = \lfloor bnxc \rfloor$ ,  $x \in [0; 1]$ . There are several methods included in fChange to detect such changes.

**Projection:** The projection-based mean change method detects changes through the [CUSUM](#) of the principal components defined in Eq (4.10) ([Berkes et al., 2009](#); [Aue et al., 2009](#)).

The integrated test statistic is defined as

$$T_n^{(PM)} = \int_0^1 \frac{1}{n} \sum_{i=1}^d \hat{\Lambda}_i \left( \sum_{i=1}^{\lfloor bnxc \rfloor} \hat{\Lambda}_i \cdot x \sum_{i=1}^n \hat{\Lambda}_i \cdot \right)^2 dx;$$

with the estimates coming from Eqs (4.9) and (4.10). Under the null hypothesis in Eq (4.18),

$$T_n^{(PM)} \xrightarrow{P} \int_0^1 \sum_{i=1}^d B^2(x) dx;$$

The projection mean-based change point detection can be run by setting the parameter method to ‘projmean’.

```
R> fchange(cancer, method = 'projmean')[c('location', 'pvalue')]
```

The cancer data has a mean change detected in 1991 ( $p < 0.001$ ).

**Fully Functional:** The fully functional mean change method does not require projection of the data, but is rather based on the  $L^2$ -norm ([Aue et al., 2018](#)). The integrated test statistic is defined as

$$T_n^{(M)} = \int_0^1 \int \int Z_{n;x}(x; t) \int Z_{n;x}(x; t) dx$$

based on Eq (4.4), which under Eq (4.18) satisfies

$$T_n^{(M)} \stackrel{P}{\rightarrow} \int_0^1 \sum_{i=1}^7 B_i^2(x) dx :$$

This can be run when method is set to ‘mean’.

```
R> fchange(rates, method = 'mean')[c('location', 'pvalue')]
```

The rates data has a mean change detected on 31 December 2007 ( $p < 0.001$ ).

Robust: The robust mean change method for Eq (4.18) generalizes the Wilcoxon two-sample test, and is less sensitive to outliers (Wegner and Wendler, 2024). This approach uses 2-sample  $U$ -statistics

$$U_n(x) = \sum_{i=1}^{bnxc} \sum_{j=bnxc+1}^n h(X_i; X_j) :$$

to construct the test statistics

$$T_{n,h}^{(RM)} = \frac{1}{n^3} \int_0^1 \iint U_n(x) \iint U_n(y) dx dy :$$

When  $h(x; y) = x - y$ , the traditional (maximum) CUSUM-statistic can be computed, and when  $h(x; y) = (x - y) \iint x - y \iint$  the Wilcoxon-type statistic emerges, which is more robust, because  $h$  is bounded. Under the null-hypothesis of stationarity, we have the weak convergence of the test statistics

$$T_{n,h} \stackrel{P}{\rightarrow} \sum_{i=1}^7 h_i \int B_i^2(x) dx ;$$

where  $h_{:1} > h_{:2} > \dots$  are the ordered eigenvalues of the long-run-covariance operator of the (unobserved) transformed time series  $(h_1(X_i))_{i \in \mathbb{N}}$ . For  $\mathcal{X}$  an independent copy of  $X$ , then  $h_1(x) = E[h(x; \mathcal{X})]$ . Functional mean change point detection based on  $h(x; y) = (x - y) \iint x - y \iint$  (Wilcoxon-type) is run by setting the parameter method to ‘robustmean’ in `fchange()`. The critical values can be computed from the long-run-variance estimator based on the estimated transformed time series  $(\hat{h}_1(X_i))_{i=1, \dots, n}$  with  $\hat{h}_1(X_i) = \frac{1}{n-1} \sum_{j \neq i} h(X_i; X_j)$  when `critical` is set to ‘simulation’. When `critical` is set to ‘resample’, a dependent wild bootstrap can be used.

```
R> fchange(electricity, method = 'robustmean')[c('location', 'pvalue')]
```

The electricity data has a mean change detected on May 23rd ( $p < 0.001$ ).

#### 4.3.4.2 Second-order change

Eigenvalues can define both mean changes and higher order changes in data. Eigenvalue tests consider the partial sample eigenvalues  $\hat{\lambda}_d^{(j)} = (\hat{\lambda}_1^{(j)}; \dots; \hat{\lambda}_d^{(j)})'$  which are estimated as in Eq (4.8), but with  $c(\cdot; \cdot)$  replaced by

$$c_i(t; s) = \text{Cov}[X_i(t); X_i(s)] : \quad (4.20)$$

A hypothesis test can be constructed with the null hypothesis

$$H_0 : \hat{\lambda}_d^{(1)} = \dots = \hat{\lambda}_d^{(n)} \quad (4.21)$$

compared to

$$H_A : \hat{\lambda}_d^{(1)} = \dots = \hat{\lambda}_d^{(k)} \neq \hat{\lambda}_d^{(k+1)} = \dots = \hat{\lambda}_d^{(n)} \quad (4.22)$$

for change point  $k = \lfloor bnx_c \rfloor$ ,  $x_c \in [0; 1]$ . The number of eigenvalues  $d$  can be specified or selected using the TVE; see Eq (4.11).

Eigenvalue: Aue et al. (2020) investigate both joint and individual eigenvalue changes by testing Eq (4.21) against Eq (4.22). Consider the test statistic for the joint eigenvalue test

$$T_n^{(JE)} = \int_0^1 \hat{\lambda}_d^{>}(x) \hat{\lambda}_d^{-1}(x) dx$$

where  $\hat{\lambda}_d^{>}(x) = \hat{\rho}_{\bar{n}} \left( \hat{\lambda}_d(x) - \frac{bnx_c}{n} \hat{\lambda}_d(1) \right)$  and  $\hat{\lambda}_d(x)$  is the vector of the first  $d$  empirical eigenvalues coinciding with estimating the covariance based on  $X_1; \dots; X_{\lfloor bnx_c \rfloor}$  as in Eq (4.9). The test statistic for individual eigenvalues can be defined as

$$T_{n;j}^{(IE)} = \int_0^1 \frac{\hat{\rho}_{\bar{n}}}{\hat{\lambda}_j} \left| \hat{\lambda}_j(x) - \frac{bnx_c}{n} \hat{\lambda}_j(1) \right| dx; \quad j = 1; \dots; d;$$

Here  $\hat{\lambda}_j^2 = \hat{\lambda}_d(j; j)$  and  $\hat{\lambda}_d$  is an estimate for the long-run covariance of the empirical eigenvalue process  $\lambda_n$ .

It can be shown that under the null hypothesis in Eq (4.21), the asymptotic behavior of the test statistics is such that

$$T_n^{(JE)} \xrightarrow{P} \int_0^1 \sum_{j=1}^d B_j^2(x) dx$$

and

$$T_{nj}^{(IE)} \stackrel{P}{\rightarrow} \int_0^1 jB_j(x)jdx;$$

where  $B_1; B_2; \dots$  are independent Brownian bridges. To run this in `fchange()`, the parameter `method` can be set to ‘eigenjoint’ or ‘eigensingle’. The eigenvalue or maximum eigenvalue of interest is set with `eigen_number`.

```
R> fchange(cancer, method = 'eigenjoint',
> eigen_number = 3)[c('location', 'pvalue')]
R> fchange(cancer, method = 'eigensingle',
> eigen_number = 3)[c('location', 'pvalue')]
```

For the cancer data, analysis of the first 2 eigenvalues—jointly and individually—revealed little evidence for a change point. The 3rd eigenvalue is also individually insignificant, but the joint eigenvalue test detects a change in 1976,  $p = 0.031$ .

Trace: Eigenvalues can also be used to detect changes in the trace of the covariance kernel (Aue et al., 2020). To test Eq (4.21) against Eq (4.22), we consider the test statistic

$$T_n^{(T)} = \frac{1}{\hat{\Lambda}_n} \int_0^1 jZ_{n;X}(x; t)jdx;$$

where  $\hat{\Lambda}_n$  is the estimated long-run variance (LRV),

$$\hat{\Lambda}_n^2 = \sum_{i=1}^T \text{Cov}(jjX_0 - X_{jj}^2; jjX_i - X_{jj}^2);$$

$X(t)$  is defined in Eq (4.6), and  $Z_{n;X}(x; t)$  is a CUSUM process as defined in Eq (4.4) but applied to  $jjX_i - X_{jj}^2$ .

When the null hypothesis Eq (4.21) holds, then

$$T_n^{(T)} \stackrel{P}{\rightarrow} \int_0^1 jB(x)jdx:$$

for Brownian bridge  $B$ . This approach can be run by setting `method` to ‘trace’ in `fChange()`.

```
R> fchange(rates, method = 'trace')[c('location', 'pvalue')]
```

For the rates data, no significant trace changes were detected,  $p > 0.1$ .

Covariance Kernel: [Horváth et al. \(2022\)](#) examine changes in the covariance kernel by studying the null hypothesis,

$$H_0 : c_1(\cdot) = \dots = c_n(\cdot) \quad (4.23)$$

with  $c_1(\cdot); \dots; c_n(\cdot)$  as defined in Eq (4.20).

Eq (4.23) can be tested by using a **CUSUM** process similar to Eq (4.4) but applied to the covariance function, i.e.

$$Z_n^{(C)}(x; t; s) = \frac{1}{n} \left( \sum_{i=1}^{bnxc} [X_i(t) \quad X_n(t)] [X_i(s) \quad X_n(s)] \right. \\ \left. - \frac{bnxc}{n} \sum_{i=1}^n [X_i(t) \quad X_n(t)] [X_i(s) \quad X_n(s)] \right)$$

This process can be used to define the test statistic

$$T_n^{(C)} = \iiint_0^1 [Z_n^{(C)}(x; t; s)]^2 dx dt ds$$

Under the null hypothesis in Eq (4.23) of no changes in the covariance kernel, it can be shown that

$$T_n^{(C)} \xrightarrow{P} \sum_{k=1}^g \frac{1}{(\lambda_k)^2} N_{k,1}^2$$

where  $N_{k,1} \sim N(0, 1)$  are **IID** standard normal random variables and  $\lambda_1; \lambda_2; \dots; \lambda_g$  are eigenvalues as defined in Eq (4.5) but for the empirical covariance kernels defined as  $[X_i(t) \quad X_n(t)][X_i(s) \quad X_n(s)]$  for  $i = 1; \dots; n$ . The covariance kernel change point detection method can be run in `fchange()` when method is set to 'covariance'.

```
R> fchange(electricty, method = 'covariance')[c('location', 'pvalue')]
```

For the electricty data, a significant covariance change was detected on April 27th with  $p < 0.001$ .

#### 4.3.4.3 Distributional change

When the type of change is distributional or simply unknown, a more general change point detection method may be preferred. Let  $F_i$  indicate the distribution of observation  $X_i(t)$ . Distributional tests consider the null hypothesis

$$H_0 : F_1 = \dots = F_n \quad (4.24)$$

against

$$H_A : F_1 = \dots = F_k \neq F_{k+1} = \dots = F_n; \quad (4.25)$$

for some change point  $k = \text{bnxc}$ ,  $x \in [0; 1]$ .

Projection: Let  $y_{i,d}$  indicate the principal component scores defined in Eq (4.10) where the number of projections  $d$  can be selected by Eq (4.11).

Extending the integrated-type test statistic by [Hušková and Meintanis \(2006\)](#), the integrated projected distributional statistic is defined as

$$T_{n,d}^{(PD)} = \int_0^1 \left( \left( \frac{\text{bnxc}(n - \text{bnxc})}{n^2} \right) \frac{\text{bnxc}(n - \text{bnxc})}{n} \sum_{j=1}^d \int_0^1 j_{x,d}(t) \cdot \sum_{i=1}^{\text{bnxc}} y_{i,d}(t)^2 w(t) dt \right) dx;$$

where  $x \in [0; 1]$ ,  $w(\cdot)$  is a nonnegative weight function with positive integral, and  $j_{x,d}(t)$  and  $\sum_{i=1}^{\text{bnxc}} y_{i,d}(t)$  are partial sums of empirical characteristic functions based on  $y_{1,d}, \dots, y_{x,d}$  and  $y_{x+1,d}, \dots, y_{n,d}$ . Common choices for  $w(\cdot)$  include  $w(t) = \exp(-ajt)$  and  $w(t) = \exp(-at^2)$  for  $a > 0$ . Using a partial sum, let

$$j_{x,d}(t) = \frac{1}{\text{bnxc}} \sum_{j=1}^{\text{bnxc}} \exp(it \cdot y_{j,d}); \quad \text{and}$$

$$\sum_{i=1}^{\text{bnxc}} y_{i,d}(t) = \frac{1}{n - \text{bnxc}} \sum_{j=\text{bnxc}+1}^n \exp(it \cdot y_{j,d});$$

for  $\mathbf{i} = \mathbf{1}^{\text{bnxc}}$ .

The simulation-based approach to determining the critical values for this test statistic for this method has yet to be developed and instead the resampling method to threshold the statistics is used, i.e. set critical to 'resample'. The project-based distributional change point detection approach can be used in `fchange()` when method is 'projdistribution'.

```
R> fchange(electr icity, method='proj di stri buti on',
> critical = 'resampl e')[c('l ocati on', 'pval ue')]
```

For the electr icity data, a significant distributional change was detected on May 16th with  $p < 0.001$ .

Fully Functional: [Horváth et al. \(2025\)](#) propose using empirical characteristic functionals to evaluate for general changes in the distributions of [FTS](#). For a random element  $X \in L^2[0; 1]$ , the *characteristic function* or *characteristic functional* of  $X$  is

$$f_X(v) = E \exp(ihX; v); \quad v \in L^2[0; 1];$$

where  $i$  is the imaginary unit. See [Bosq \(2000\)](#), pg. 32.

The integrated-type test statistic to test Eq (4.24) against Eq (4.25) is defined as

$$T_n^{(D)} = \int_0^1 \int_0^1 jZ_n^{(D)}(v; x)^2 dQ(v) dx;$$

where  $Q$  is the measure on  $L^2[0; 1]$  that, by default, is generated by Brownian motion. The [CUSUM](#) process  $Z_n^{(D)}$  is defined as the [CUSUM](#) of empirical characteristic functionals

$$Z_n^{(D)}(v; x) = \rho_{\bar{n}} \left( \hat{f}_n(v; x) - \frac{bnxc}{n} \hat{f}_n(v; 1) \right);$$

for partial sample estimates of  $f_X$ ,

$$\hat{f}_n(v; x) = \frac{1}{n} \sum_{j=1}^{bnxc} \exp(ihX_j; v); \quad v \in L^2[0; 1]; \quad x \in [0; 1];$$

The test statistic  $T_n^{(D)}$  can be approximated, for sufficiently large  $B$ , via Monte-Carlo integration, i.e.

$$T_n^{(D)} \approx \frac{1}{B} \sum_{i=1}^B \int_0^1 jZ_n^{(D)}(W_i; x)^2 dx;$$

where  $W_1; \dots; W_B$  are [IID](#) stochastic processes with distribution  $Q$ , e.g. iid standard Brownian motions, independent of  $fX_i g_{i2Z}$ .

It can be shown that under mild assumptions, there exists a mean zero, complex valued Gaussian process  $W : L^2[0;1] \rightarrow \mathbb{C}$  such that

$$T_n \stackrel{P}{\rightarrow} \int_0^1 \int_0^1 j(v; x) j^2 dQ(v) dx :$$

Details for the estimation of  $T_n$  are left to [Horváth et al. \(2025\)](#) for brevity.

The characteristic functional approach is run in `fchange()` when parameter `method` is set to ‘characteristic’.

```
R> fchange(rates, method = 'characteristic')[c('location', 'pvalue')]
```

For the rates data, a significant distributional change was detected on 16 January 2008 with  $p < 0.001$ .

#### 4.3.4.4 Multiple changes

We often wish to segment a given FTS into approximately homogeneous subsegments by estimating multiple change points. For example in the cancer data, significant changes were detected using several of the proposed metrics; yet a single change is clearly insufficient to explain all the variation in the data; see [Figure 4.2](#). Testing for a multiple change alternative such as

$$H_A : X_1 = \dots = X_{k_1} \neq X_{k_1+1} = \dots = X_{k_2} \neq \dots \neq X_{k_m} = \dots = X_n \quad (4.26)$$

rather than a single change alternative could create a more parsimonious model. [Binary segmentation](#) and elbow plot methods are included in `fChange` to extend the single change methods to multiple change methods.

**Binary Segmentation:** Intuitively, [binary segmentation \(BS\)](#) refers to successive single segmentation of a data set ([Scott and Knott, 1974](#); [Vostrikova, 1981](#); [Venkatraman, 1992](#)). Yet directly repeating a single change procedure may result in a test that is poorly sized. Conducting individual change point tests usually entails comparing a change point detector to a threshold. [Rice and Zhang \(2022\)](#) suggest the threshold  $\hat{\tau}_n \sqrt{2 \log(n)}$ , where  $\hat{\tau}_n^2 = \text{median}(jjX_{i+1} - X_{ij}j^2=2; i = 2; \dots; n)$ . A complementary improvement is adding a final refinement step where segments that are presumed to only contain a single change are re-examined to refine the change point estimates. This process may be described by the `BINSEG(1, n, n)` algorithm which returns the set of estimated, ordered, change points.

---

Function BINSEG( $l, u, n$ )

---

```
 $l$  start index ;  $u$  end index ;  $n$  threshold  
if  $u - l = 1$  then STOP else  $k_0 := \underset{l < k < u}{\operatorname{sargmax}} jjS_{l,u}^k jj$  ;  $S = jjS_{l,u}^{k_0} jj$   
if  $S > n$  then  
  Add  $k_0$  to the set of estimated change points  
  run BINSEG( $l, k_0, n$ ) and BINSEG( $k_0, u, n$ )  
else STOP  
end if
```

---

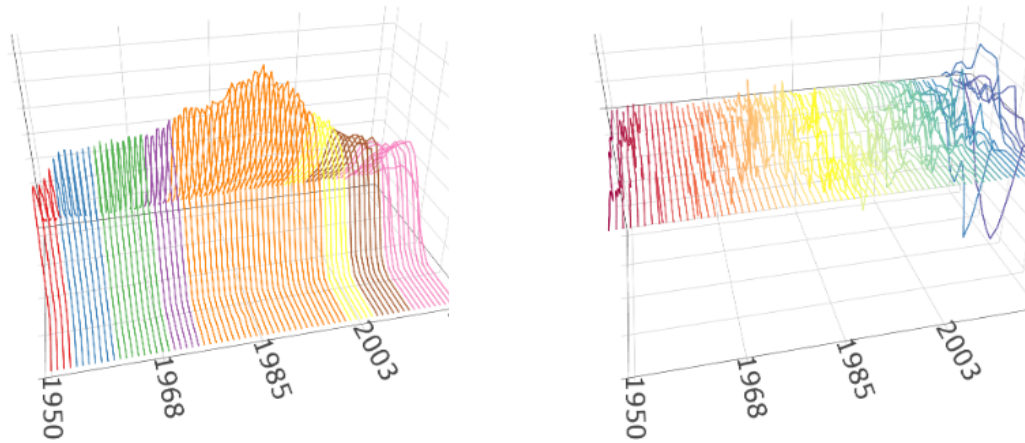
**BS** is implemented in `fchange()` by setting `type` to ‘segmentation’. Parameters such as `method`, `statistic`, and `critical` can be used to specify the details as before. Additional parameters such as methods to trim the data can be found in the function’s documentation.

Using the mean change point detection method, **BS** without trimming is employed on the cancer data and the cancerresiduals data.

```
R> cancer_changes <- change(cancer, method = 'mean',  
> type = 'segmentation')  
R> cancerresiduals_changes <- change(cancerresiduals, method = 'mean',  
> type = 'segmentation')  
R> plot(cancer, cancer_changes)  
R> plot(cancerresiduals, cancerresiduals_changes)
```

Figure 4.5 shows 7 mean changes were detected in the cancer data. The changes were detected at 1953, 1960, 1969, 1974, 2000, 2005, and 2012 with corresponding  $p$ -values 0.036, 0.007, 0.023, 0.000, 0.005, 0.035, and 0.005. Corresponding events in cancer research occurred near each of these changes. For example, in 1953 the first complete cure of a human solid tumor by chemotherapy was completed. An overview of major milestones in cancer research can be found in [Zurrida and Veronesi \(2015\)](#) and [NCI \(2025\)](#). No mean changes were detected in `cancerresiduals`, where the  $HU(5)$  model was used to capture underlying trends in cancer.

**Elbow Plot:** The elbow plot method sequentially suggests  $K$  changes points based on the largest reduction in total variability of the data. The total variability in the data is then given as a function of the number of change points and this is used to chose the number of significant change points. A change point  $i, i = 1; \dots; K$ , is selected by examining the entire data set and finding the point which obtains the largest the variability reduction, subject to the previous  $i - 1$  change points. Once  $K$  change points have been selected, the



(a) Segmented cancer

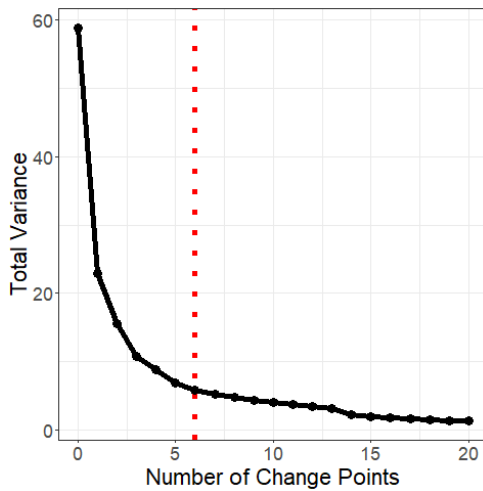
(b) Segmented cancer residuals

Figure 4.5: Cancer segmentation. Segmentation for cancer and cancer residuals using the fully functional mean change model with [binary segmentation](#). (a) The 7 detected mean changes segment the cancer data. Colors indicate approximately homogeneous regions. (b) No mean changes were detected in cancer residuals. A rainbow spaghetti plot is shown.

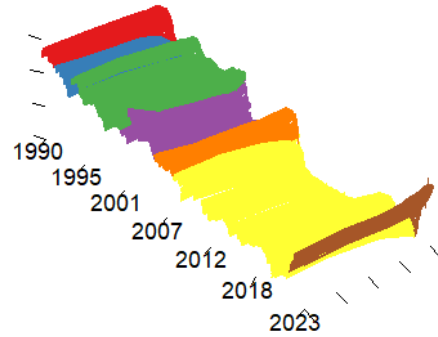
changes are studied as a function of the total variability, e.g. using  $L^2$  norm. The number of significant changes points in the data is determined by finding a plateau, or elbow, in the variability-change function. More exactly, the elbow is the point at which relatively large variability gains are no longer obtained by including additional change points to segment the data. The elbow can be determined algorithmically or via a plot.

The elbow plot method is implemented in `fchange()` by setting `type` to 'elbow'. Details for the algorithmic approach can be found in the corresponding `fChange` documentation.

Figure 4.6a shows the elbow, or variability-change, plot of rates. The red vertical line indicates the algorithmically suggested number of change points. Using that elbow, Figure 4.6b was created. There are 6 changes recommended. In chronological order, the first is detected on 29 November 1991 related to the recession following the Gulf War. This is followed by a change on 15 April 1994 at the end of Global Currency crisis and starting the dot-com boom. The change on 29 December 2000 around the bursting of the dot-com bubble and start of the housing market boom. A change is detected on 7 July 2005 as the housing correction and continued Fed rate hikes. The 3 January 2008 change is related to the global financial crisis. The final change on 31 March 2022 is present when global



(a) Variability-change plot



(b) Segmented data

Figure 4.6: Rates segmentation. Segmentation for rates using the elbow plot method. (a) Variability-change plot showing the total variability as a function of the number of changes. The red vertical line indicates 6 change points are algorithmically suggested. (b) Spaghetti plot for the rates data with 6 change points. Approximately homogeneous regions are indicated using the same color.

inflation was rampant following the COVID-19 pandemic.

#### 4.3.4.5 Confidence intervals

Following the detection of a change point, it is natural to consider how confident one can be in the placement of that change point. Inference on the change point location is dependent on both the data and the nature of the change point. [Horváth and Rice \(2024\)](#) discuss a method for computing the confidence interval of mean changes. This can be run in `fChange` by using `confidence_interval()`. The results can be visually represented using `plot()`.

Let the change point splitting the data be denoted  $\hat{k}_n$ . The mean difference between the two data segments is defined as

$$\hat{e}_n(t) = X_{\hat{k}_n,2}(t) - X_{\hat{k}_n,1}(t);$$

where

$$X_{\hat{k}_n,1}(t) = \frac{1}{\hat{k}_n} \sum_{i=1}^{\hat{k}_n} X_i(t) \quad \text{and} \quad X_{\hat{k}_n,2}(t) = \frac{1}{n - \hat{k}_n} \sum_{i=\hat{k}_n+1}^n X_i(t);$$

The variance  $\sigma^2$  is defined as

$$\sigma^2 = \sum_{j=1}^p E g_{0j} g_{0j};$$

with

$$g_{0j} = \int_{-\infty}^{\infty} \psi_j(t) \frac{e_n(t)}{j j e_n j j} dt;$$

for errors  $\psi_j(\cdot)$ . A natural estimator for the residuals of the data under the proposed mean change is

$$\hat{\psi}_i(t) = \begin{cases} X_i(t) - X_{\hat{k}_n,1}(t) & \text{if } 1 \leq i \leq \hat{k}_n \\ X_i(t) - X_{\hat{k}_n,2}(t) & \text{if } \hat{k}_n + 1 \leq i \leq n \end{cases};$$

These residuals can be used to estimate a **LRV** using kernel  $K$

$$\hat{\sigma}_n^2 = \sum_{j=1}^p \sum_{i=1}^n K\left(\frac{\cdot}{h}\right) \hat{\psi}_i(\cdot);$$

where

$$\hat{\psi}_i(\cdot) = \begin{cases} \frac{1}{n} \sum_{i=1}^n \hat{g}_i \hat{g}_{i+\cdot} & \text{if } 0 \leq \cdot < n; \\ \frac{1}{n} \sum_{i=1}^n \hat{g}_i \hat{g}_{i+\cdot} & \text{if } n < \cdot < 0; \end{cases}$$

for bandwidth  $h$  and  $\hat{g}_j = \int \hat{\psi}_i(t) \hat{e}_n(t) dt$ .

Define a triangular drift and two-sided Brownian motion at break fraction  $\alpha = k/n$  as

$$m(t) = \begin{cases} (1 - \alpha)(1 - \alpha) + \alpha & ; \quad \text{if } t < 0 \\ 0 & ; \quad \text{if } t = 0 \\ (1 - \alpha) + \alpha(1 - \alpha) & ; \quad \text{if } t > 0; \end{cases}$$

and

$$W_2(t) = \begin{cases} W_{2,1}(-t); & \text{if } t < 0 \\ W_{2,2}(t); & \text{if } t \geq 0 \end{cases}$$

where  $W_{2,1}$  and  $W_{2,2}$  are independent Brownian motions. Define the random variable  $\tau$ ,

$$\tau = \tau(\omega) = \operatorname{argmax}_t [W_2(t) - \int_0^t m(t)] :$$

The distribution of  $\tau$  can be computed through Monte Carlo simulation or through its density function

$$f_\tau(t) = \begin{cases} h(-t; (1-\alpha)(1-\beta) + \alpha; (1-\alpha) + (1-\beta)); & t \leq 0 \\ h(t; (1-\alpha) + (1-\beta); (1-\alpha)(1-\beta) + \alpha); & t > 0 \end{cases}$$

where

$$h(u; x; y) = \frac{2x(x+2y) \left[ 1 - \Phi\left(\frac{(x+2y)u^{1/2}}{\sqrt{2x^2 + (x+2y)^2}}\right) \right]}{\exp[2y(x+y)u] - 2x^2 \left[ 1 - \Phi\left(\frac{xu^{1/2}}{\sqrt{2x^2 + (x+2y)^2}}\right) \right]}; \quad u \geq 0;$$

with  $\Phi(x)$  denoting the standard normal distribution function.

The approximate  $1 - \alpha$  confidence intervals around change  $\hat{k}_n$  can be defined as

$$\left( \hat{k}_n - \frac{\hat{\sigma}_n^2(\tau)_{1-\alpha/2}}{jj\hat{\sigma}_n j_2^2}; \hat{k}_n + \frac{\hat{\sigma}_n^2(\tau)_{\alpha/2}}{jj\hat{\sigma}_n j_2^2} \right);$$

where  $(\tau)_q$  is the  $q$  quantile for the distribution of  $\tau$ .

Using the characteristic functional-based distributional change point detection method with BS, 6 change points were detected on electricity. The changes were on March 8th, May 2nd, May 22nd, August 26th, September 21st, and October 30th. At an  $\alpha = 0.1$  significant level, the corresponding confidence intervals could be constructed as follows.

```
R> electricity_interval <- confidence_interval(electricity,
> changes = c(67, 122, 142, 238, 264, 303))
R> plot(electricity, intervals = electricity_interval)
```

Figure 4.7 shows the change points and confidence intervals for the electricity data. In the figure, the data are lightly colored, while the segment means are darkly colored. Both tight and wide bands are present, emphasizing differing levels of confidence for the changes, some of which may not be primarily mean changes.

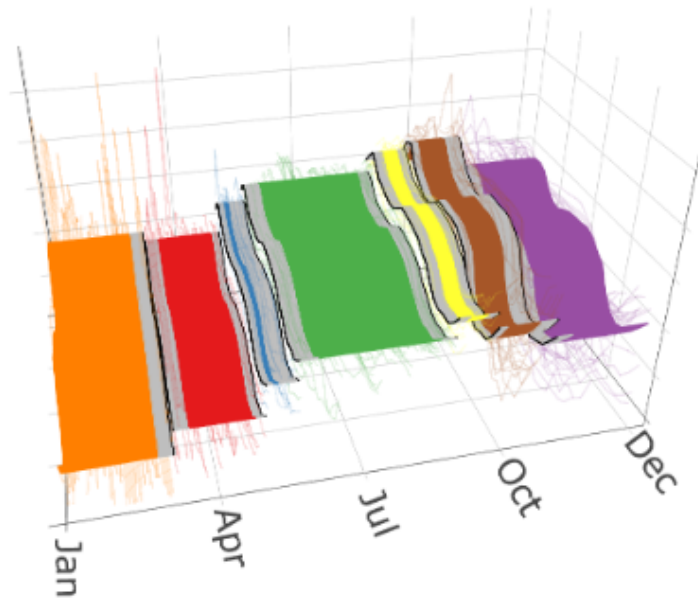


Figure 4.7: Electricity segmentation and intervals. Segmentation for electricity from the fully functional distributional change point detection method based on characteristic functionals. Confidence intervals are computed. Colors indicate approximately homogeneous regions. The black curves indicate the observations immediately surrounding change points. The gray curves indicate the confidence intervals for change points. The observed electricity prices are displayed in light colors while the mean of the region is displayed with dark colors.

## 4.4 Summary and discussion

With the increased prevalence of [FTS](#), functional change point analysis is an essential analysis tool. We present `fChange`, an R package with a focus on functional change point analysis to match the emerging developments. The package contains methods to store, summarize, and visualize discretely observed [FTS](#) along with many change point methods. The package is under active development. It will be improved and new methods continue to be incorporated.

## Chapter 5

# Using Random Forests to Uncover the Predictive Power of Distance–Varying Cell Interactions in Tumor Microenvironments

### 5.1 Introduction

Recent advances in cancer treatment, such as immune checkpoint inhibition and other cancer immunotherapies, have sparked a growing interest in studying the cellular composition and spatial organization of the tumor microenvironment (TME). The latest innovations in imaging technologies allow for single cell resolution of specific proteins, facilitating in-depth study of the spatial arrangement of cell types within the TME. A wide variety of technologies are available for this purpose, each with different benefits and trade-offs ([Toth and Mezey, 2007](#); [Angelo et al., 2014](#); [Giesen et al., 2014](#); [Lin et al., 2015](#); [Goltsev et al., 2018](#); [Herbel et al., 2019](#); [Saka et al., 2019](#)). For a review of the available technologies see [Lewis et al. \(2021\)](#).

In comparing TME data, different spatial relationships between cell types, e.g. between tumor cells and specific immune cell populations, and/or individual proteins, often appear predictive of patient outcomes and may guide therapeutic interventions; see for example [Gaglia et al. \(2023\)](#). Comparisons between cancer subtypes, e.g. hormone-positive versus hormone-negative breast cancers, or lung squamous cell carcinoma vs lung adenocarcinoma, may provide novel insight into tumor biology and guide the development of treatments.

A further goal is to identify specific spatial relationships observed in particular patient’s tumor that are useful in predicting an outcome, such as patient survival or response to therapy. Recent results demonstrate that TME data can be used for such prediction in a variety of tumor types (Wang et al., 2021; Lin et al., 2021).

We consider such prediction problems for data sets generated from tumors imaged with Multiplexed Ion Beam Imaging (MIBI) by means of the MIBIScope in this chapter. The MIBIScope uses ion-beam ablation and time-of-flight mass spectrometry to detect up to approximately 40 protein markers on formalin-fixed, paraffin-embedded (FFPE) tissue. Thus, it provides comprehensive data on cell characteristics and their localisation at a single-cell resolution of around 250-400nm (Keren et al., 2018, 2019). Such data collected on the TME can be considered as marked spatial point patterns (Illian et al., 2008; Diggle, 2013; Baddeley et al., 2015). The cell locations can be considered as points within the pattern, with cell phenotypes and/or protein markers giving the “marks”. An example of such a point process generated from a tumor imaged using MIBIScope is shown in Figure 5.1a (Keren et al., 2018).

Existing methods developed to study cellular interactions in the TME have exploited cell neighborhood analysis in which the spatial relationship between a cell of interest and its neighboring cells can reveal particular cell-cell interactions associated with a disease state or changes associated with response to therapy; see e.g. Schürch et al. (2020). Pairwise cell-to-cell distance calculations over iterations of randomized permutations has also been used to identify relevant cell-cell interactions (Jiang et al., 2022). However, the substantial number of cell types present in the TME leads to a very large number of potential pairwise interactions creating a major challenge in finding interactions that may be meaningful and statistically significant in predicting outcomes of interest. There is ongoing interest in applying spatial statistics methods to similar biological data sets, e.g. Damond et al. (2019), Jackson et al. (2020), and Färkkilä et al. (2020).

A common method of analysing spatial point patterns, such as those that arise in TME imaging, is to consider Ripley’s  $K$  function (Bartlett, 1964; Ripley, 1976, 1977). The  $K$  function describes the distribution of inter-point distances in a given point pattern, giving an indication as to whether points in the pattern (e.g. cells) are clustered or dispersed with respect to one another. The  $K$  function, along with other summary functions from spatial statistics, has previously been employed in the analysis of the TME (Carstens et al., 2017; Barua et al., 2018; Bull et al., 2020; Parra et al., 2021; Parra, 2021; Canete et al., 2021; Vu et al., 2022). An example of a  $K$  function showing the relative distribution of a specific immune cell type around tumor cells within a MIBIScope image from a triple negative breast cancer patient is shown in Figure 5.1b.

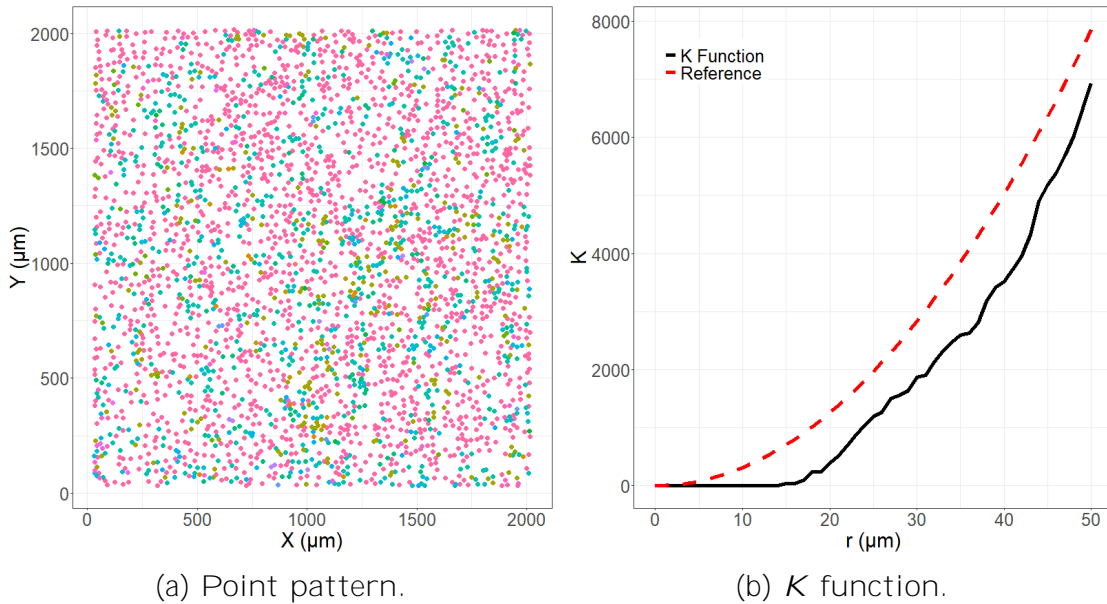


Figure 5.1: An example of point pattern data and an associated  $K$  function. (a) Point pattern data associated with a tumor imaged using MIBIScope from a triple negative breast cancer patient with multiple identified phenotypes. The  $x$ - and  $y$ -axes represent the spatial dimensions, with the points giving individual cell locations, and the colour of the points indicating one of 15 unique phenotypes, e.g. tumor (red), NK cells (purple), and monocytes/neutrophils (cyan). (b) The associated cross  $K$  function (black) for two cell types in the image: tumor and monocytes/neutrophils. The  $x$ -axis indicates the radius,  $r$ , and the  $y$ -axis gives the value of the  $K$  function. The estimated  $K$  function can be compared to  $r^2$  (red dashed line), which is the theoretical  $K$  function associated with complete spatial randomness.

In this chapter, we present a general framework for analyzing and identifying useful spatial relationships in the TME through predicting an outcome of interest. The method we propose uses a novel combination of spatial statistics and [functional data analysis](#), in conjunction with methods in ensemble machine learning. The application of [functional data analysis](#) to spatial point pattern data is a recent development ([Hywood et al., 2016, 2021](#); [Vu et al., 2022](#); [Seal et al., 2022](#); [Yi et al., 2023](#)).

Our approach begins by producing  $K$  functions for the different cell-cell, (or alternatively marker-marker), interactions within images. After performing dimension reduction using functional principal components analysis ([Ramsay and Silverman, 2005](#)), these data are combined with non-functional patient meta-data, such as age or sex, and a modified

random forest model is used to predict the patient outcome. Motivated by [Barber and Candés \(2015\)](#), in order to evaluate the predictive power of the spatial interactions, “knock-off” point patterns that mimic the spatial data in the TME are generated, via permutation, independently of the responses. The importance of specific spatial interactions in predicting the response are evaluated by comparison to the predictive power of the knock-off spatial patterns. This approach overcomes the challenge of distinguishing important spatial interactions among many potential interactions of interest, i.e. it controls the false discovery rate. It also lends itself to the generation of easy-to-interpret plots showing which interactions are useful in predicting the response. Moreover, it grants high power for even a relatively few number of cells due to the robustness of the  $\mathcal{K}$  functions, and high power even with small sample sizes. Small sample sizes are a relatively common obstacle for this type of data.

To our knowledge the method we propose is the first to evaluate for the importance of spatial interactions in the TME that may vary over distance with a large number of potential cell interactions. Many statistical methods proposed to date rely on simple summaries of the images, such as the average number of cells observed, or otherwise simplify the data by, for example, initially clustering cells into groups (e.g. [Bottai et al. \(2016\)](#), [Lehmann et al. \(2016\)](#), [Sugie et al. \(2020\)](#), [Patwa et al. \(2021\)](#), [Wang et al. \(2021\)](#), and [Jiang et al. \(2022\)](#)). In other comparable methods available in the literature, a single distance of interest is considered (e.g. [Keren et al. \(2018\)](#), [Barua et al. \(2018\)](#), and [Greenbaum et al. \(2023\)](#)). [Canete et al. \(2021\)](#) provides an approach for detecting differences across multiple images between cell-cell interactions by comparing the integrated difference between a spatial summary function similar to the  $\mathcal{K}$  function and its expected value under complete spatial randomness. However, this approach only considers individual spatial interactions, while we wish to consider all interactions in a single model. Additionally, a possible drawback of the approach in [Canete et al. \(2021\)](#) is that integration over the summary statistic may lose valuable information relating to differences in the shape of the functions, since differently shaped spatial summary functions may still have the same integral. [Vu et al. \(2022\)](#) uses a related functional linear model based on interaction distance functions, although this approach is not tailored to examine a large number of interactions simultaneously.

Some methods may predict well, but do not lead to interpretable results that allow for the importance of individual spatial characteristics to be compared. Other methods consider only a single image or an equal number of images per patient. This is further complicated by the fact that some images do not contain all cell types. However, our approach is interpretable and able to be used to analyse data with multiple, possibly differing, numbers of observations per patient. Additionally, images may have different

shapes or sizes. In that sense, our approach allows for the complete use of the data.

We apply the proposed methods to two MIBIScope data sets; a data set of triple negative breast cancer (TNBC) patients, and a data set consisting of both lung squamous cell carcinomas (LUSC) and lung adenocarcinomas (LUAD). Regarding the TNBC data, our method was accurately able to identify clustered versus dispersed tumors when compared to [Keren et al. \(2018\)](#), and was additionally able to identify important cell spatial interactions in making that determination. Our method also indicated that there did not appear to be measurable differences in the spatial arrangement of tumor and immune cell types, as measured by homogeneous K functions, between the LUSC and LUAD groups.

Whilst the methodology presented here is motivated by, and applied to, MIBIScope data, it can be applied to similar data generated by other technologies, e.g. OPAL, Phenocycler Fusion, Merscope, Xenium and Cosmx ([Wang et al., 2012](#); [Black et al., 2021](#); [Chen et al., 2022](#); [Williams et al., 2022](#); [Li et al., 2022](#); [Canning et al., 2022](#); [Sadeghirad et al., 2023](#)). Furthermore, the methodology can easily extend beyond two-dimensions to higher-dimensional images, another area of active research ([Zanotelli et al., 2020](#)).

The rest of the chapter is organized as follows. In section 5.2, we give a detailed description of the data we consider and the methods to analyze them, including subsections on how we fit a modified random forest in this setting, and how we evaluate the statistical significance and uncertainty in measuring the variable importance of spatial interactions of cell types as encoded by  $K$  functions. We also introduce the R package `funkyCellS` (shortened version of functional data analysis of K functions for multiplexed images of cells), an open-source implementation of our approach in that section. Section 5.3.1 details the results of simulation experiments in which we found that the proposed method performed well when applied to synthetic data built to mimic the TNBC data. We report the results when this approach was applied to the TNBC and LUSC vs. LUAD data sets in section 5.3.2. Some concluding remarks and directions for future work are collected in section 5.4.

## 5.2 Materials and methods

The raw spatial data that we consider take the form of 2-dimensional point patterns, as generated using MIBIScope. We denote the cell spatial data, as

$$\mathbb{C} = \{ (x_{c;t}^{(p;i)}, y_{c;t}^{(p;i)}, a_{c;t}^{(p;i)}) ; p = 1, \dots, N; i = 1, \dots, I_p; t = 1, \dots, T; c = 1, \dots, n_{p;t} \} \quad (5.1)$$

where  $(x_{c;t}^{(p;i)}, y_{c;t}^{(p;i)}, a_{c;t}^{(p;i)})$  denotes the  $x$  and  $y$  coordinates and the associated cell properties  $a$  of a given cell. Specifically, the term refers to the the  $c^{\text{th}}$  cell of type  $t$  (of which there

are  $T$  total types), for the  $i^{\text{th}}$  image of the  $p^{\text{th}}$  patient, with  $n_{p:i:t}$  giving the number of cells of type  $t$  in image  $i$  of patient  $p$ . The properties in  $a_{c:t}^{(p:i)}$  may (typically) simply give the cell’s phenotype (and is therefore redundant due to the term  $t$ ), or may be more general, such as a vector describing individual protein expression—allowing use of this method for processed or raw data. For example, the vector could be composed of binary indicators as to whether a protein is expressed or not. For notational clarity, and since we here only consider data consisting of cells and their associated phenotype, we drop the  $a$  term throughout the chapter.

Since the applicability of our method extends beyond this example, we designate several general terms for use throughout the chapter. We refer to point patterns such as in Figure 5.1a as “images”. We interchangeably use the terms cell phenotype and cell type. We also interchangeably use the terms cross-over K function and K function. We assume a single response variable,  $Z_i$ , for each of the  $N$  patients (e.g. tumor type, response to therapy, etc.). The set of outcomes for the  $N$  patients is denoted  $Z = (Z_1; \dots; Z_N)$ . In the real data examples we consider below,  $Z_i$  is a binary response, e.g. “compartmentalized” versus “mixed” tumors for the TNBC data, or LUSC versus LUAD for the lung cancer data, in which case we can encode the outcomes as taking the values 0 and 1. These methods may easily be adapted for more general class responses, e.g. different types of tumors, or numeric responses, e.g. survival time.

In addition to the spatial data, we assume that we may have access to non-spatial data on the patients. We refer to this data as patient “meta-data”, and we assume that it takes the form  $M = (\mathbf{m}_1; \dots; \mathbf{m}_N)$ , where each  $\mathbf{m}_i$  is a vector of patient attributes, for example age or sex.

With both the cell spatial data  $C$  and meta-data  $M$ , our goals are to (1) investigate to what extent these data are useful in predicting the outcomes  $Z$ , and (2) to identify which specific spatial relationships and/or components of the meta-data from the full data set are useful in predicting  $Z$ . We deem data on a spatial relationship or component of the meta-data “useful” if their importance in predicting the outcome exceeds, to a statistically significant degree, that of similar variables that are known to be unrelated to the outcome. For reference throughout the chapter, high-level schematics of our proposed method are presented in Figs 5.2 and 5.3. Figure 5.2 overviews the major steps for processing data in our model while Figure 5.3 focuses on the steps of the statistical analysis of the model.

Towards answering these questions, we build a model of the outcomes  $Z$  in terms of the image spatial information  $C$  and meta-data  $M$ . In doing so, we must address how we incorporate the complex image data into such a model. Motivated by the expectation that patient outcomes are influenced by the relative distribution of various immune cells or

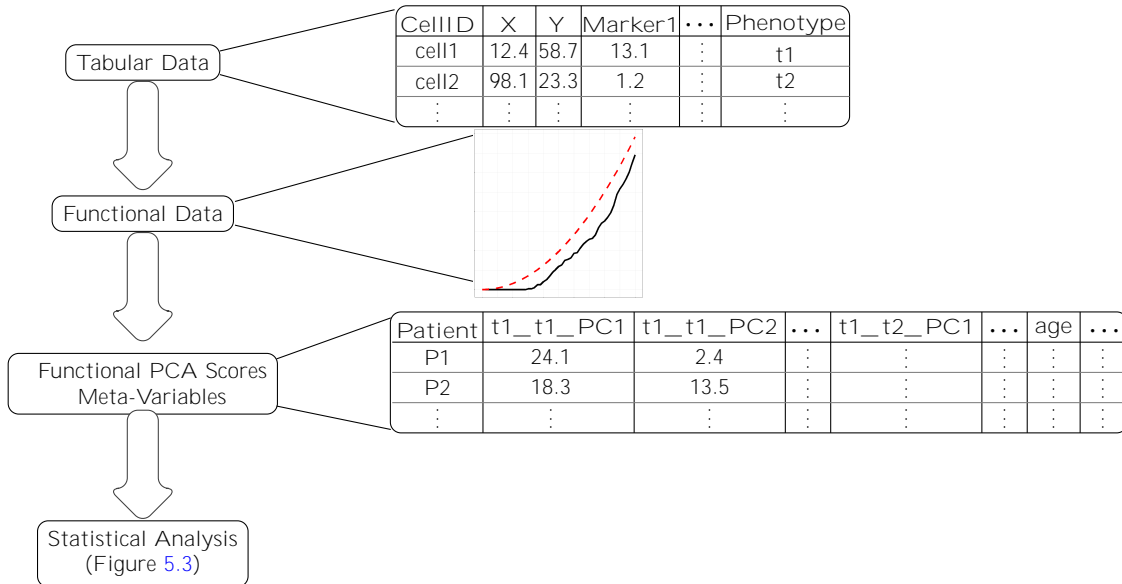


Figure 5.2: Flow chart for data processing. The methods presented here begin with tabular data obtained after pre-processing multiplex images (steps that include cell segmentation, phenotyping, etc.). For a given image the tabular data consists of rows for each imaged cell, giving the associated x-y position, marker intensities, and cell phenotype. Next, the tabular data are converted into spatial  $K$  functions for each interaction of interest (this can be exhaustive, and include all possible interactions between phenotypes, or selective, with only a subset of interactions analysed). Next,  $K$  functions are converted into functional principal component scores. Patient meta-variables are added at this stage. The resulting data is then used in the statistical model, as described in Figure 5.3.

protein markers around each other, we begin by computing spatial “cross-over”  $K$  functions from the image data, which summarize the spatial distribution of cells with respect to one another as in Figure 5.1b.

We provide an open source implementation of our approach in R ([R Core Team, 2022](#)) on CRAN, in the package, `funkycell`s, or at [github.com/jrvanderdoes/funkycell](https://github.com/jrvanderdoes/funkycell)s. This implementation is runnable on personal computers, and includes the code and data used in the presented simulations and data analyses.

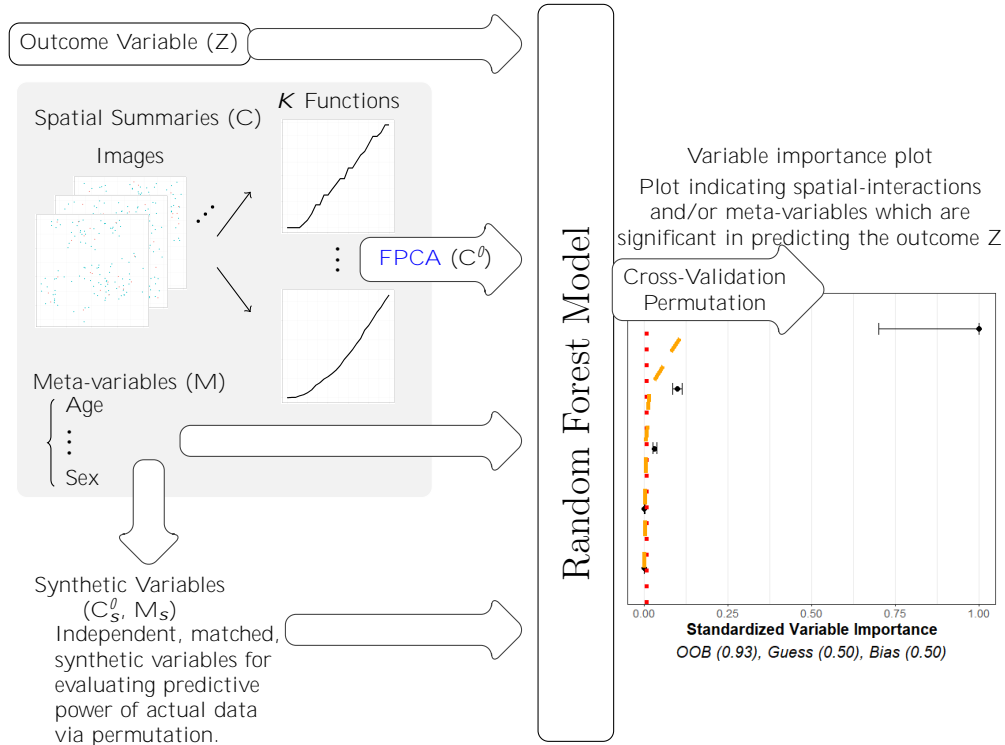


Figure 5.3: Flow chart of model. When modeling using funkycell<sub>s</sub>, there are several major steps: organizing data, generating synthetic data, and modeling using random forests. The spatial data is organized into functional summaries ( $K$  functions) that are projected into finite dimensions (FPCA) and used with meta-variables to predict the outcome variable. The spatial data and meta-variables are permuted to create synthetic variables with similar properties but independent of the outcome. These synthetic variables are then added to the model, and used to quantify the strength of the relationships between the spatial and meta-data with the response. The model processes the data, employing cross-validation and permutation to return a variable importance plot (with predictive accuracy estimates) indicating spatial interactions and/or meta-variables which are significant in predicting the outcome  $Z$ .

### 5.2.1 Summarizing the image data using $K$ functions

The cross-over  $K$  function for image  $i$  of person  $p$  and cell types  $t$  and  $t^\theta$  is defined as

$$K_{t;t^\theta}^{(p;i)}(r) = \frac{1}{t^\theta} \text{E}(\text{Number of cells of type } t^\theta \text{ within distance } r \text{ of a cell of type } t) ; \quad (5.2)$$

where  $E$  denotes mathematical expectation, the radius  $r$  ranges from  $0 \leq r \leq R$ , with the max radius  $R$  being a user specified parameter that we discuss below, and  $t^\theta$  gives the density of cells of type  $t^\theta$  (Illian et al., 2008; Diggle, 2013; Baddeley et al., 2015).

By examining this function for varying radii, we may infer how cell types are distributed around each other. For example, if cell types are distributed around each other entirely at random, then  $K_{t;t^\theta}^{(p;i)}(r)$  is equal to the area of a circle of radius  $r$ ,  $\pi r^2$ . Regularity or dispersion of the cells around each other tends to reduce  $K_{t;t^\theta}^{(p;i)}(r)$  while clustering tends to increase it. An example of a  $K$  function computed between the tumor and monocytes/neutrophils phenotypes for a given tumor in the TNBC data set is presented in Figure 5.1b, which indicates a degree of dispersion with respect to monocytes/neutrophils around tumor cells across the given  $r$  values compared to that expected for cells distributed around tumor cells with complete spatial randomness. Cross-over  $K$  functions can be used to summarize all two-way interactions between cell types for a given image.

In practice, estimation is based on an empirical average replacing the expectation. The estimated cross-over  $K$  function for image  $i$  of person  $p$  is given by

$$K_{t;t^\theta}^{(p;i)}(r) = \frac{jij}{n_{p;i;t^\theta}} \sum_{c=1}^{n_{p;i;t}} \sum_{c^\theta=1}^{n_{p;i;t^\theta}} \mathbf{1} \left( \sqrt{[x_{c;t}^{(p;i)} - x_{c^\theta;t^\theta}^{(p;i)}]^2 + [y_{c;t}^{(p;i)} - y_{c^\theta;t^\theta}^{(p;i)}]^2} \leq r \right); \quad (5.3)$$

$0 \leq r \leq R$ , where indicator  $\mathbf{1}(A)$  takes the value one if the condition  $A$  is satisfied, and is zero otherwise, and  $jij$  indicates the area of the image. When patients have multiple images, we combine their cross-over  $K$  functions by computing a weighted average,

$$K_{t;t^\theta}^{(p)}(r) = \sum_{i=1}^{I_p} \frac{n_{p;i;t}}{n_{p;i;t}} K_{t;t^\theta}^{(p;i)}(r); \quad 0 \leq r \leq R; \quad (5.4)$$

with  $I_p$  giving the number of images for a given patient  $p$ . In other words, the  $K$  functions from each image are weighted according to the prevalence of the cells of the type under consideration. We note that if there is one image per patient (so that  $I_p = 1$ ), then  $K_{t;t^\theta}^{(p)}(r) = K_{t;t^\theta}^{(p;1)}(r)$ , and further that the weights in Eq (5.4) vanish if the cell types  $t$  and  $t^\theta$  are missing in an image. We use a standard isotropic edge correction in this chapter (see Appendix B.2).

In computing these  $K$  functions for each cell type, we can transform the spatial data  $C$  into a collection of  $T^2$  different  $K$  functions for each patient,  $fK_{t;t^\theta}^{(p)}(r); t, t^\theta = 1; \dots; T; 0 \leq r \leq Rg$ . The  $K$  functions are then treated as **functional data** objects; see e.g. Ramsay and Silverman (2005). Since even moderate values of  $T$  lead to a large number of  $K$  functions

to consider, user input is often helpful in determining a smaller subset of interactions (and hence  $K$  functions) of particular interest for analysis.

Although informative, these  $K$  functions are unwieldy to directly use in a model, and we further transform the functions using the dimension reduction technique of [functional principal component analysis \(FPCA\)](#). FPCA is a common technique in [functional data analysis](#) that decomposes the leading sources of variability among the curves  $K_{t,t^\theta}^{(p)}(r)$  into a set of finite-dimensional, approximately uncorrelated principal components (PCs); see [Ramsay and Silverman \(2005\)](#). To do so, for each pair of cell types  $t$  and  $t^\theta$ , we define the empirical covariance kernel as

$$\hat{C}_{t,t^\theta}(r; s) = \frac{1}{N} \sum_{\rho=1}^N \left[ K_{t,t^\theta}^{(\rho)}(r) \quad K_{t,t^\theta}(r) \right] \left[ K_{t,t^\theta}^{(\rho)}(s) \quad K_{t,t^\theta}(s) \right]; \quad (5.5)$$

where  $K_{t,t^\theta}(r) = \frac{1}{N} \sum_{\rho=1}^N K_{t,t^\theta}^{(\rho)}(r)$ .

The eigenvalues and eigenfunctions of the kernel  $\hat{C}_{t,t^\theta}$  are then computed to satisfy the functional equation

$${}_{i;t,t^\theta} \hat{C}_{t,t^\theta}(r; s) {}_{i;t,t^\theta} = \int_0^R \hat{C}_{t,t^\theta}(r; s) {}_{i;t,t^\theta}(s) ds; \quad (5.6)$$

The  $K$  function  $K_{t,t}^{(p)}$  is summarized using the  $d$  coefficients (PCs)

$$k_{t,t^\theta}^{(p;d)} = \left( \int_0^R K_{t,t}^{(p)}(r) {}_{1;t,t^\theta}(r) dr; \dots; \int_0^R K_{t,t}^{(p)}(r) {}_{d;t,t^\theta}(r) dr \right)^>; \quad (5.7)$$

The  $d$  coefficients comprising  $k_{t,t^\theta}^{(p;d)}$  describe the projection of the  $K$  function  $K_{t,t}^{(p)}$  onto the finite dimensional linear space spanned by  ${}_{1;t,t^\theta}; \dots; {}_{d;t,t^\theta}$ , which are optimal in terms of capturing the variability among the curves  $K_{t,t}^{(p)}$ ,  $\rho = 1; \dots; N$ , with a  $d$ -dimensional summary. An advantage of summarizing the curves in this way is that, when differences in the  $K$  functions across the population are present due to differences in the outcome(s) of interest, the PCs are expected to capture these differences.

As such, we summarize the spatial data using the principal components  $C^\theta = f K_{t,t^\theta}^{(p;d)}; \rho = 1; \dots; N; t; t^\theta = 1; \dots; Tg$ , which we then incorporate with the meta-data  $M$  into a model for  $Z$  of the form

$$\hat{Z} = f(C^\theta; M); \quad (5.8)$$

Since our ultimate goal includes evaluating which spatial interactions or elements of the meta-data are useful in predicting the outcomes, we use a random forest model for

*f.* Random forest models are tree-based ensemble machine learning methods in which decision trees are built, after sampling with replacement the patient data and discarding some covariates at random, by sequentially splitting on variables to minimize a metric for predicting  $Z$  (Breiman, 2001). The main reasons for the sampling procedures for the patient data and covariates in building each tree is to build nearly independent trees and also address overfitting, common in many machine learning applications. When the trees are combined to create a forest, increased statistical power is observed. Additionally, the computational complexity is similar to a traditional random forest model, see section 5.3.

## 5.2.2 Variable importance

Random forest models are useful in achieving our goals since they have strong predictive power while still allowing for a quantification of the usefulness of individual covariates in predicting the response through various “variable importance” measures. We now describe the computation of variable importance metrics for the random forest models introduced Breiman (2001). There are multiple methods for calculating variable importance values, such as permuting variables in data and comparing the difference in loss metrics, or measuring node purity. We tested several methods and found similar results regardless of the variable importance metric used. Due to computational considerations, we implement a node purity metric, sometimes called recursive partitioning, as described in Breiman et al. (1984) and Therneau and Atkinson (2022).

In the following explanations for node purity variable importance, it is perhaps useful to imagine the scenario where the outcomes  $Z_i$  take the values 0 and 1. When each constituent decision tree is formed in producing the random forest model, nodes are split based on some impurity metric relating to the outcomes  $Z_i$  (Breiman et al., 1984). For a given node,  $A$ , node impurity is defined as

$$I(A) = \sum_{i \in O} g(p_{i;A}) ; \tag{5.9}$$

where  $g$  is an “impurity” function,  $O$  is the set of possible outcomes, e.g. 0 or 1, and  $p_{i;A}$  is the proportion of data in  $A$  which belong to outcome class  $i$  of  $Z$ , i.e. 0 or 1. Typical choices for  $g$  are the information index ( $g(p) = -p \log(p)$ ) or the Gini index ( $g(p) = p(1 - p)$ ). At a given stage of the decision tree, the splitting variable, and its value, is chosen to maximize the impurity reduction

$$I_A = \rho(A)I(A) - \rho(A_L)I(A_L) - \rho(A_R)I(A_R) \tag{5.10}$$

where  $A_L$  and  $A_R$  are respectively the left and right resulting nodes and  $p(A)$  is the probability of  $A$  (for future observations) (Therneau and Atkinson, 2022). A variable  $v$ 's importance in a single tree can be computed as

$$\sum_{i \in P_v} I_i \tag{5.11}$$

where  $P_v$  is the set of splits for (i.e. nodes that split on) the variable.

Variable importance (for the entire forest) can similarly be calculated by considering nodes across all decision trees, typically standardized by the number of trees fit or the number of trees where the split was present,

$$VI(v) = \frac{1}{\text{Number of Trees}} \sum_{i \in P_v} I_i \tag{5.12}$$

Additional modifications, such as the use of surrogates, can also be added to improve variable importance metrics. The technical details are left to more complete works on random forests, e.g. see Breiman et al. (1984), and Appendix B.3.

### 5.2.3 Variable importance comparison

Although the computed variable importance is a helpful summary statistic for ordering variables in terms of their expected usefulness in predicting the outcome, there are several challenges in using these values to determine which variables are significantly more useful than others. One is that the variables in  $C^\theta$  are  $d$ -dimensional proxies of the information derived from the spatial image data. When multiple components are used to describe a single function, i.e.  $d > 1$ , we must take into account that each individual component in  $C^\theta$  describes only a portion of the associated  $K$  function. Therefore, the importance of each component must be combined to describe the importance of each spatial interaction. Yet this importance must be made comparable to that of the meta-variables. Also, we wish to identify spatial interactions and meta-variables that are of “significant importance”, which we take to mean that their importance exceeds to a statistically significant degree that of similar variables that are unrelated to the response. This task is complicated by the fact that we are often faced with such a large number of spatial interactions. Given the large number of variables, we expect some to have anomalously large variable importance even when they are independent from the response.

To allow for such comparisons, we standardize the variable importance metrics by adding matched synthetic variables. These synthetic variables are generated by permuting

the true data in order to maintain identical distributions, but are independent from the outcome. We denote the synthetic spatial components as  $C_s^\ell$  and the synthetic meta-variables  $M_s$ .

A random forest model is fit using both the true and synthetic variables,

$$\hat{Z} = f(C^\ell; M; C_s^\ell; M_s) : \quad (5.13)$$

To build  $C_s^\ell$ , a random functional variable is selected for each  $b$ 'th iteration,  $1 \leq b \leq B$ , where  $B$  is a user specified parameter. The  $d$  principal components associated with each patient are permuted across patients, resulting in assignment of the  $d$  components to a random patient and hence outcome. Note that in doing so, the  $d$  components are kept together.

One could use  $B$  synthetic variables for each functional variable. However, investigations into the model through extensive simulations has shown that a single group of  $B$  synthetics is generally sufficient for the  $K$  functions, and additional synthetic  $K$  variables do not improve power to identify important variables beyond this.

Although one synthetic group for meta-variables could be used, previous work has shown a tendency for random forests to favor continuous predictors over discrete predictors (Strobl et al., 2007). The model accounts for this tendency through unique synthetic meta-variable groups. Therefore,  $M_s$  is created by permuting each meta-variable across patients  $B$  times.

We use these synthetic variables,  $C_s^\ell$  and  $M_s$ , to standardize the variable importance values of the true data and build noise thresholds. In doing so we are able to infer which spatial interactions and meta-variables lead to significant improvements in the model accuracy. The details of this are left to section 5.2.4.

Due to the innate randomness in the models, the variable importance values fluctuate between runs and model fits of the random forests. To quantify this, we employ [cross-validation \(CV\)](#).

Let the data be randomly assigned to the  $F$  folds such that an indexing function  $f: \{1, \dots, N\} \rightarrow \{1, \dots, F\}$  indicates the partition to which each  $p^{\text{th}}$  patient's data are allocated. Denote the fitted model, using the true variables as in the model from Eq (5.13), with the  $j^{\text{th}}$  fold removed  $\hat{Z}^{(-j)}(C^\ell; C_s^\ell; M; M_s)$ . The estimated [CV](#) variable importance for each functional variable  $c$  (in both  $C^\ell$  and  $C_s^\ell$ ), which are described by the previously discussed  $d$  dimensional principal components  $k_c^{(d)}$ , is computed as

$$\text{VI}^{(CV)}(c) = \frac{1}{F} \sum_{j=1}^F \sum_{i=1}^d \text{VI}_j^{(CV)}(k_c^{(i)}); \quad (5.14)$$

where  $\text{VI}_j^{(CV)}(k_c^{(i)})$  denotes the variable importance estimate from the  $j^{\text{th}}$  fold-removed model  $\hat{Z}^{(j)}(C^\theta; C_s^\theta; M; M_s)$  for component  $k_c^{(i)}$ , which is the  $i^{\text{th}}$  component relating to the  $c^{\text{th}}$  function. For meta-variable  $m$ , the variable importance is computed as

$$\text{VI}^{(CV)}(m) = \frac{1}{F} \sum_{j=1}^F \text{VI}_j^{(CV)}(m) : \quad (5.15)$$

We quantify the uncertainty in the estimate of the variable importance measure for each variable  $v$ , both functional and meta, by calculating its standard deviation across the  $F$  folds.

$$\text{SD}(v) = \sqrt{\frac{1}{F-1} \sum_{j=1}^F \left( \text{VI}_j^{(CV)}(v) - \text{VI}^{(CV)}(v) \right)^2} : \quad (5.16)$$

This uncertainty estimate is used with variable importance estimates created from the mean of non-cross-validated models. Let  $\text{VI}_j(x)$  indicate the variable importance metric from the model in Eq (5.13), iterate  $j$  (where we take  $1 \leq j \leq F$  for ease, but each run is on the full data set), then the estimates are computed respectively for functional and meta-variables as

$$\text{VI}(c) = \frac{1}{F} \sum_{j=1}^F \sum_{i=1}^d \text{VI}_j(k_c^{(i)}) ; \text{ and } \text{VI}(m) = \frac{1}{F} \sum_{j=1}^F \text{VI}_j(m) : \quad (5.17)$$

#### 5.2.4 Comparing variable importance of spatial interactions and meta-variables to noise

As mentioned above, due to the use of  $d - 1$  principal component summaries, we expect that for spatial interactions and meta variables that are independent of the outcome the variable importance of the spatial interactions will typically be larger. As such we use the estimated variable importance values for the synthetic variables to calibrate the variable importance between the spatial interactions and meta-variables. We compute for each of the spatial and meta-variables, respectively  $c$  and  $m$ , the empirical quantiles of the variable importance values of the synthetic variables. If  $C^{(s)}$  indicates the set of synthetic functions, i.e. the combined synthetic components, and  $M^{(s;m)}$  indicates the set of synthetic

meta-variables matched to meta-variable  $m$ , then we set

$$Q_C = \inf \left\{ q : \frac{1}{B} \sum_{c_{\text{syn}} \in \mathcal{C}^{(s)}} \mathbf{1}(\text{VI}(c_{\text{syn}}) \geq q) > \alpha \right\}; \text{ and} \quad (5.18)$$

$$Q_{M;m} = \inf \left\{ q : \frac{1}{B} \sum_{m_{\text{syn}} \in \mathcal{M}^{(s;m)}} \mathbf{1}(\text{VI}(m_{\text{syn}}) \geq q) > \alpha \right\} :$$

Let  $Q_M = (Q_{M;1}; \dots; Q_{M;jMj})$  where  $jMj$  is the total number of meta-variables. Below we always set  $\alpha = 0.95$ . Letting  $Q_{\text{noise}} = \max_{\mathcal{C}; Q_M} Q_C$ , we calibrate the variable importance of each true variable computed from the model in Eq (5.13) that includes synthetic variables, denoted  $v_{\text{true};(5.13)}$ , as

$$\text{VI}_{\text{adj}}(v_{\text{true};(5.13)}) = \frac{Q_{\text{noise}}}{Q_C \mathbf{1}(v_{\text{true};(5.13)} \geq C) + \sum_{m \in \mathcal{M}} Q_{M;m} \mathbf{1}(v_{\text{true};(5.13)} \geq m)} \text{VI}(v_{\text{true};(5.13)}) : \quad (5.19)$$

The cross-validated standard deviations of the variable importance values are similarly adjusted based on the model from Eq (5.8) as

$$\text{SD}_{\text{adj}}(v_{\text{true};(5.8)}) = \frac{Q_{\text{noise}}}{Q_C \mathbf{1}(v_{\text{true};(5.8)} \geq C) + \sum_{m \in \mathcal{M}} Q_{M;m} \mathbf{1}(v_{\text{true};(5.8)} \geq m)} \text{SD}(v_{\text{true};(5.8)}) : \quad (5.20)$$

These adjusted variable importance values may then be compared to  $Q_{\text{noise}}$ . Estimates that go below the threshold  $Q_{\text{noise}}$  have variable importance values that appear to show no statistically significant relationship with the outcome. Plotting  $Q_{\text{noise}}$ , the adjusted variable importance values,  $\text{VI}_{\text{adj}}(v)$ , along with their adjusted standard deviations gives a simple way to evaluate at a glance which variables appear to be important in predicting the response. An example plot of this is shown in the right hand panel of Figure 5.3, where  $Q_{\text{noise}}$  is plotted as a vertical, red-dotted line, and the variable importance values are plotted (black dots) along with their adjusted standard deviations (surrounding bands).

While  $Q_{\text{noise}}$  estimates the  $\alpha$  percentile of the variable importance values corresponding to outcomes that are unrelated to the outcome, the variable importance values themselves can be distorted due to overfitting. In essence as more variables are observed, it is increasingly likely to find a variable that seems to have high importance, despite no true relationship to the response.

To account for this potential effect, we employ an additional variable permutation. In this step, we permute the true variables in addition to the synthetic variables,  $H$  times.

In each forest, the synthetic variables are again used to align the variables (using the previously computed  $Q_{\text{noise}}$ ), but the remaining, aligned “true” variables do not have any relation to the outcome. In each of these forests based on variables that are independent from the response,  $1 \leq h \leq H$ , the resulting, adjusted variable importance estimates are ordered,

$$VI_{adj;h;\cdot} = f^{\text{th}} \text{ largest adjusted variable importance value in forest } hg; \quad (5.21)$$

ensuring  $VI_{adj;h;1} > VI_{adj;h;2} > \dots > VI_{adj;h;V}$  where  $V$  denotes the total number of variables between the spatial interactions,  $C$ , and the meta-variables,  $M$ . When the number of variables  $V$  is large, especially in relation to the number of patients  $N$ , we might expect even when the spatial variables and meta-variables are independent of  $Z$  that the largest variable importance values will (far) exceed  $Q_{\text{noise}}$ . As such, we also compute the  $q$  quantile of the variable importance values in each ordered position  $\cdot$  for the random forests fit to the permuted data,

$$Q_{int;\cdot} = \inf \left\{ q : \frac{1}{H} \sum_{h=1}^H 1(VI_{adj;h;\cdot} \leq q) > \right\}; \quad (5.22)$$

We let  $Q_{\text{int}} = (Q_{int;1}, \dots, Q_{int;V})$ , and call it the “interpolation threshold”.

We include  $Q_{\text{int}}$  with  $Q_{\text{noise}}$  in order to quantify overfitting and thereby evaluate the significance of values of the variable importance values  $VI_{adj}(v)$ ; this is the orange-dotted line in the right hand panel of Figure 5.3.

In summary, variables with adjusted importance values,  $VI_{adj}(v)$ , that are larger than both  $Q_{\text{noise}}$  and  $Q_{int;\cdot}$  exhibit importance that significantly exceeds (at the 1 level) what we might expect from similar variables that are unrelated to the outcome. This holds taking into account the inflation in the variable importance values that arise from fitting the random forest to a large number of spatial interactions and meta-variables.

### 5.2.5 Choice of parameters $B$ and $H$

The parameters  $B$  and  $H$  must be selected by the user. In constructing  $Q_{\text{noise}}$  and  $Q_{int;\cdot}$ , these values define the number of synthetic variables used to approximate the quantiles in Eqs (5.18) and (5.22), respectively. We have explored several choices of  $B$  and  $H$  in our simulation experiments, and found that when the level of the thresholds is set at  $\alpha = 0.95$ , the values  $B = H = 100$  give satisfactory results for the up to 16 cell types considered. Further details of these simulations are shown in Appendix B.4. In summary,

we have observed that for up to 16 cell phenotypes and with  $\alpha = 0.95$ , values of  $B$  and  $H$  exceeding 50 behaved almost identically in terms of their false positive and false negative rates for identifying important variables. There was no apparent advantage, or apparent disadvantage other than additional computational burden, for choosing larger values of  $B$  and  $H$  for the number of cell interactions considered. We note that we only considered up to 16 cell phenotypes. We recommend using larger values of  $B$  and  $H$  for a larger number of cell phenotypes.

### 5.2.6 Predictive accuracy estimates

In weighing the significance of the computed variable importance values, one should also consider the overall predictive accuracy of the final model for the outcomes. For example, a variable may have a large variable importance value within a model that does not lead to improved predictions of the outcome over naïve models.

We consider out-of-bag accuracy (OOB) to estimate how well our final random model works on unseen data, and compare it to the naïve approaches that we label GUESS, and BIAS.

OOB is computed by predicting the data left out during each CV iteration,

$$\text{OOB} = \frac{1}{N} \sum_{p=1}^N \text{Diff} \left( \hat{Z}^{(\rho)} \left( X^{(\rho)}; M^{(\rho)} \right); Z_p \right); \quad (5.23)$$

for the  $N$  patients and where Diff indicates a difference function. For classification problems such as with the TNBC or LUSC versus LUAD data, this may be defined using an indicator function,  $\text{Diff}(x; y) = \mathbb{1}(x = y)$ :

GUESS is defined as the probability of correctly guessing a patient's outcome by randomly selecting the outcome based on the outcome's observed frequency in the original data. If for the observed frequencies of outcomes  $1; \dots; n$  in the data are  $p_1; \dots; p_n$ , this is computed as

$$\text{GUESS} = \sum_{i=1}^n p_i^2 \quad (5.24)$$

For example, in the TNBC data in which we wish to predict the “compartmentalised” versus “mixed” result with a proportion  $\rho$  of “compartmentalised” patients, this amounts to computing the rate at which we would accurately guess the outcome by flipping a coin independently for each patient with probability  $\rho$  of heads, and guessing the outcome

is “compartmentalised” for heads, and is “mixed” otherwise. The GUESS value would therefore have a success probability of  $p^2 + (1 - p)^2$ .

BIAS is built by always guessing the patients outcome that is the most prevalent outcome in the sample. For classification problems, the most likely outcome can naturally be defined by the data mode. We compute

$$\text{BIAS} = \frac{1}{N} \sum_{p=1}^N \text{Diff}(\text{mode}(Z) - z_p) : \quad (5.25)$$

For two outcome data sets like the TNBC or lung cancer examples, this will have success probability for a random patient drawn from the sample of  $\max\{p; 1 - p\}$ .

When no signal is in the data, we expect OOB to perform similarly to GUESS or BIAS. However, if the covariate information available is useful in predicting the response, we expect OOB to far exceed the naïve estimates. The variability of OOB estimates is further analyzed and discussed in Appendix B.5.

### 5.2.7 Variable importance plot

The variable importance plot summarizes the variable importance values of both the spatial interactions and meta-variables. It shows how they compare to what we might expect from similar noise variables which are unrelated to the outcomes, and also summarizes the overall efficacy in predicting the outcomes using the random forest model. Figure 5.4 is the variable importance plot created from simulated data with two cell types,  $A$  and  $B$ , and two meta-variables with differing distributions, age and sex. Variables are displayed according to their adjusted variable importance estimates (black dots), standardised with respect to the variable with the largest variable importance (such that the standardised variable importance values range from 0 to 1). The variables are listed on the y-axis in order of decreasing variable importance with the largest values being at the top.

The plot visualizes uncertainty in computing the VIs through (gray) intervals representing one standard deviation on either side of the estimate. These intervals are the adjusted standard deviations estimated through CV, as given in Eq (5.20).

The noise threshold  $Q_{\text{noise}}$  as well as the interpolation threshold  $Q_{\text{int}}$  at 95% levels ( $\alpha = 0.05$ ) are also shown. Figure 5.4 presents the noise threshold as a red, dotted, vertical line and the interpolation threshold as a orange, dashed, curved line.

Estimates for predictive accuracy are given at the bottom of the figure. The estimates include OOB, GUESS, and BIAS, respectively defined in Eqs (5.23), (5.24), and (5.25).

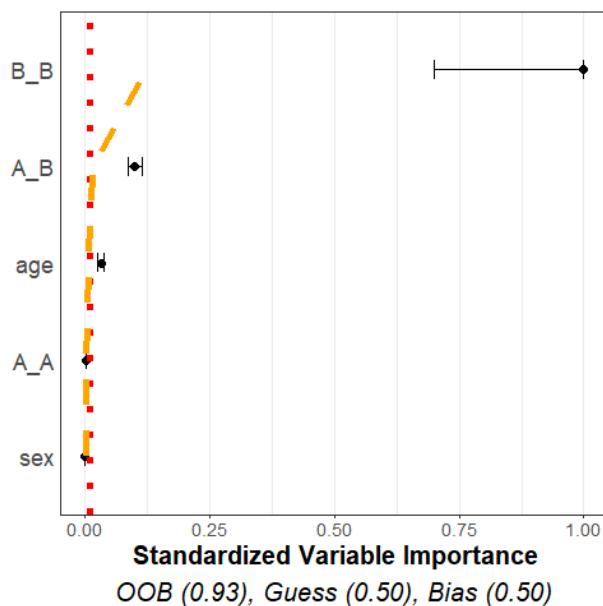


Figure 5.4: Sample variance importance plot. This sample variable importance plot uses simulated data with a binary outcome, two cell types, and two meta-variables. The data was simulated with significant differences between the outcomes in the  $B\_B$ ,  $A\_B$  spatial interactions, and age meta-variable, but no significant difference across sex and the  $A\_A$  spatial interaction. The point estimates of the variable importance values are the black dots, with accompanying intervals indicating the uncertainty. The red dotted straight line is the noise threshold and the orange dashed curved line is the interpolation threshold. Both thresholds are used to indicate if a variable is predictive of the outcome beyond that of random noise. The variable importance values of the known significant variables are shown to exceed that of the noise and interpolation thresholds.

Further interpretation of the variable importance plots are given in sections 5.3.1 and 5.3.2. These sections consider simulations for which there is a known solution and real data examples that can be interpreted.

## 5.3 Results

In this section we present results on simulation experiments and analyses of two real experimental data sets using the proposed methodology. All investigations were completed

on a standard personal laptop (with an Intel i7 processor), and did not require high-end computational resources. In fact, the computational complexity can be readily shown to be

$$O\left(N \log(N)[dc + m][F + H]BT\right); \quad (5.26)$$

and notably increases only linearly in each variable aside from the sample size, which includes an additional logarithmic factor. Additional details on the computational complexity of random forests models can be found in e.g. [Breiman et al. \(1984\)](#).

Many additional (unreported) simulations were also conducted to test the robustness of the method against various patterns, image shapes, sample sizes, and parameter choices. We found robustness to sample size and variable count (See Appendix B.6), informative and non-informative patterns, image shape and size (provided images are large enough to capture the desired relationships), and parameter choices (See Appendix B.4).

### 5.3.1 Simulation study

We present the results here of simulation experiments in which we applied the proposed methods to simulated spatial point patterns. In particular, we produced simulated point patterns with properties, such as cell counts, numbers of phenotypes, etc., similar to that of the TNBC data set in [Keren et al. \(2018\)](#). The real data motivating this simulation experiment are analysed in the following section. The primary goals of the simulation experiment were to demonstrate that the proposed method is effective at controlling the false discovery rate, i.e. the rate at which we mistakenly identify spatial interactions or meta-variables with no relationship to the response of interest as being informative or important, and also that it has the power to detect important spatial interactions.

We considered simulated data from 34 “patients”. These patients were defined as being negative or positive for a binary outcome  $Z$  (note that positive/negative here refers to the arbitrary outcome  $Z$ , and is not related to hormone receptor status as it does in the term TNBC). We let there be 17 positive and 17 negative patients and simulate one image per patient. Each image consists of a point pattern with 16 cell phenotypes. Patients were also ascribed a single meta-variable, which we call age. We developed a random forest model as described in section 5.2 to predict patient outcome using interactions between the 16 different cell types in the images and the additional meta-variable.

We considered two main settings: (1) a simulation with only non-informative variable-outcome relations and (2) a simulation with both informative and non-informative variable-outcome relations.

To generate the images, the cells were placed according to multiple, potentially nested, (modified) Thomas processes, which are constructed iteratively (Thomas, 1949). For a given image, a Thomas process first places cells (of a given type, say  $c_1$ ) at random, according to a Poisson process. As such the number of cells to place,  $n_{c_1}$ , is determined at random based on a Poisson distribution, where the distributional parameter is user selected. For our simulation study, these are selected to correspond to the mean number of cells in the TNBC dataset. We standardize the images to unit length in both the  $x$  and  $y$  directions.

Around each cell  $c_{a,c_1}$ ,  $a \in \{1, \dots, n_{c_1}\}$ , cells of a different type, say  $c_2$  may be placed. Again the number of  $c_2$  cells are randomly selected based on a Poisson distribution with a user selected parameter. The coordinates of the  $c_2$  cells are placed according to another distribution in such a way that they either cluster or disperse around the  $c_1$  cells. In our experiments, a bivariate Normal distribution is used to place the  $c_2$  points, so that the mean coincides with the location of a randomly selected  $c_1$  cell, and the covariance matrix is a scalar multiple  $\sigma^2$  times the identity matrix. By varying  $\sigma^2$  in the outcome groups, clustering or repulsion in the cell interactions can be introduced.

Additional cells can be simulated around the points of  $c_1$  or  $c_2$ , and so on. Moreover, the original cells can be removed such that the new cells appear to exhibit self-clustering. Compilations of such placement patterns, with use of potentially different distributions, can achieve images with varying degrees of clustering or regularity.

In our simulations, 16 cells were iteratively placed according to this modified Thomas pattern. Some cells were placed completely at random ( $c_1, c_4, c_5, c_6, c_7, c_{13}, c_{14}, c_{15}$ ), some were placed exhibiting self-clustering ( $c_8, c_9, c_{10}, c_{11}, c_{12}$ ), and some were placed exhibiting clustering around  $c_1$  ( $c_2, c_3, c_{16}$ ). In what we call the “no-relation” simulation, the cells locations were simulated in the same way for both the positive and negative patients. However in the “relation” simulation,  $c_2$  exhibited increased clustering around  $c_1$  while  $c_3$  exhibited repulsion from  $c_1$ , for positive outcomes. Similar to the true data, some cell types were present hundreds of times per image while others only rarely appeared. Each of the synthetic cell types were generated to mimic behaviors and frequencies seen in the TNBC data. Figure 5.5 presents two images, an image from the TNBC data set and a simulated image.

In this way a single image for each of the 17 positive and 17 negative patients were simulated. The meta-variable age was simulated as a Normal random variable with unit variance. While in the no-relation case the mean age was constant, 25 for both outcomes, in the relation-case the mean age was set to be 25 for negative outcome patients and 27 for positive outcome patients. Moreover, in the relation-case, 2 of the 17 positive patients

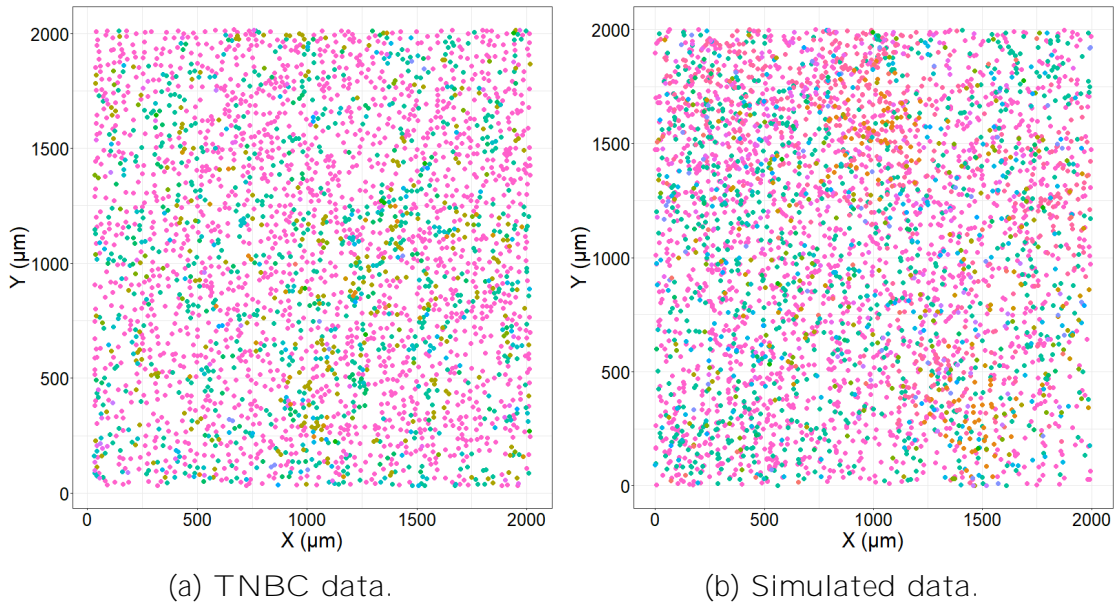
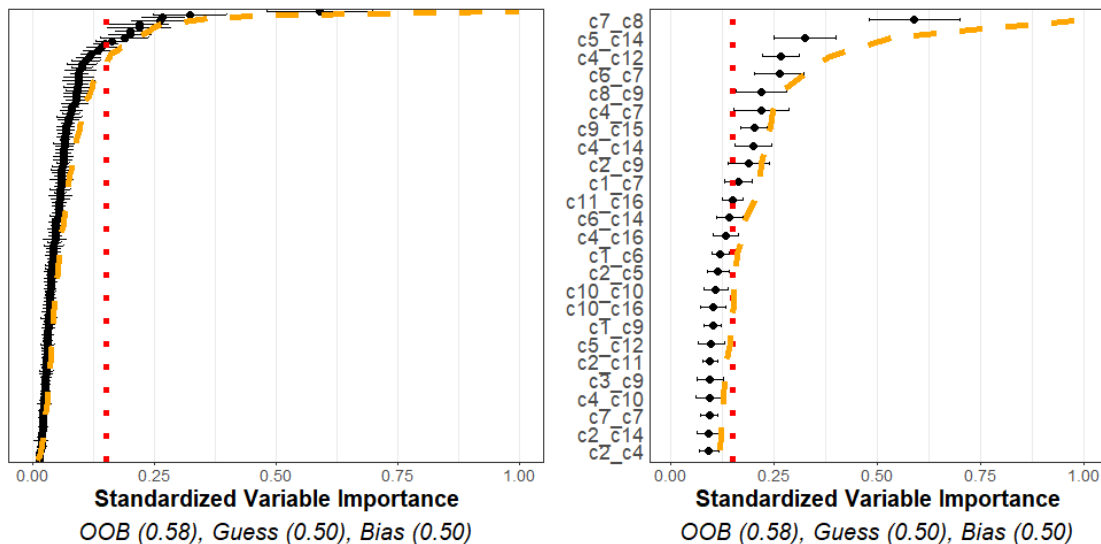


Figure 5.5: Comparison of TNBC and simulated data. (a) An image from the TNBC data and (b) an image from the simulated data. Different colors indicate one of the 16 different cell phenotypes, showing the comparability of the simulations and true data.

were given no-relation to the outcome as additional “noisy” images.

When modeling both cases, there were several tuning parameters selected. We used the standard choices of 500 trees for each random forest, each tree using 80% variable selection and full data bootstrap, 10 folds in cross-validation, and a standard significance level of  $\alpha = 0.05$ . We used 100 interactions for the permuted random forest in creation of the interpolation cut-off. Appendix B.4 highlights several numerical investigations of these values, showing robustness in the results for these choices. Creation of the  $K$  functions also required selection of the maximum radius,  $R$ . Although we investigated the effect of the choice, we saw little variation in the results and used a traditionally recommended 25% of the side length of the image, along with the previously discussed isotropic edge corrections for the simulations. Moreover, the  $K$  functions were summarized using 3 principal components, which were selected to match the TNBC analysis. In the TNBC case 3 components explain at least 95% of the total variance explanation for each  $K$  function.

Figure 5.6a shows representative variable importance for all variables and Figure 5.6b shows the top 25 in the “no-relation” case. One may see from the plot that some variable importance estimates exceeded the red noise threshold, but they all were observed to be

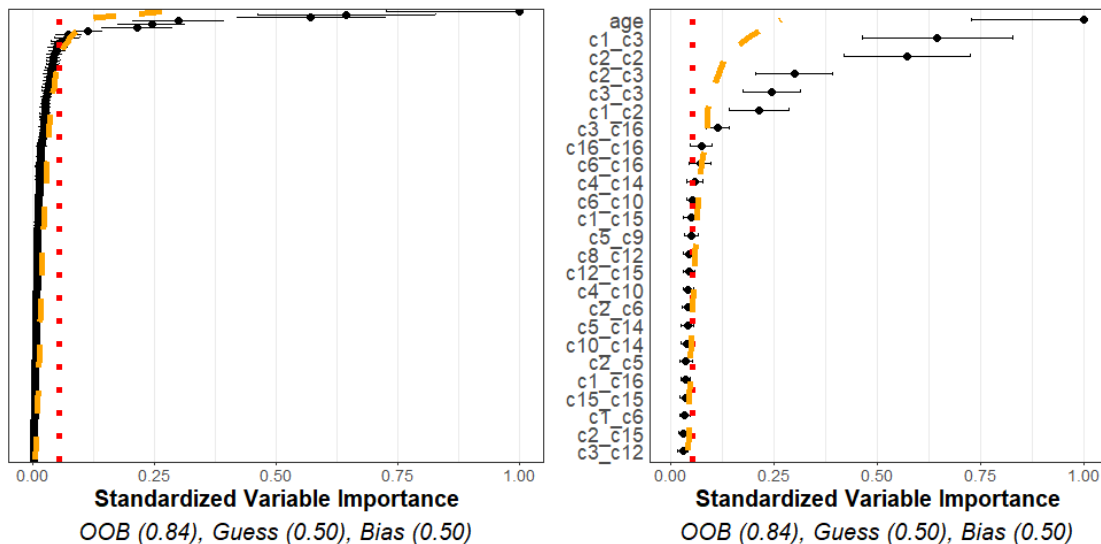


(a) Full variable importance plot. (b) Top 25 variable importance plot.

Figure 5.6: No relationship simulation. Simulation of 16 cell types for 34 patients with meta-variable age. (a) Figure with the variable importance values for all variables. (b) Figure with only the top 25 largest variable importance values. All variables were generated with no-relationship to the outcome and all were determined to have no relation to the outcome beyond noise as the variable importance estimates are below noise and interpolation thresholds.

below the orange interpolation threshold. This may be interpreted that the observed variable importance values did not exceed, to a significant degree, what we would expect to see from similar variables that are known to be independent of the response. Moreover, the prediction accuracy estimate (OOB) indicates the model performs similarly to a naïve guessing approach. Taken together, the plot indicates that none of the spatial interactions appeared to be important in predicting the outcome, as was expected in this case.

On the other hand, Figure 5.7 shows a representative variable importance plot computed from a single simulation run in the “relation” simulation. All variables are shown in Figure 5.7a, and only top 25 are shown in Figure 5.7b. Although many variables are still below the noise and interpolation thresholds, the known related variables are found to have a significant relationship with the outcome variable. Moreover, the OOB estimate far exceeds that of the GUESS or BIAS estimates. This plot indicates that interactions between the  $c1$ ,  $c2$ ,  $c3$  cell types, and the age variable, appeared to be useful, to a statistically significant degree, in predicting the outcome, once again as expected. We note that since  $c2$  and  $c3$



(a) Full variable importance plot. (b) Top 25 variable importance plot.

Figure 5.7: Relationship simulation. Simulation of 16 cell types for 34 patients with meta-variable age. (a) Figure with the variable importance values for all variables. (b) Figure with only the top 25 largest variable importance values. Most cell types were generated with no relationship to the outcome. However, age,  $c1\_c2$ , and  $c1\_c3$  were designed to have a relationship with the outcome (which naturally means  $c2\_c2$  and  $c2\_c3$  would also have relationships to the outcomes). These variables are seen with significantly larger variable importance values than the thresholds and other variable importance values.

cells are distributed around  $c1$  cells, any change in these distributions will necessarily lead to differences in the cross  $\mathcal{K}$  functions between these two cell types as well, as observed.

In order to verify that each of the orange and red lines appeared to be appropriately calibrated, we performed an additional simulation experiment. We considered two cases (1) with 4 cell types and (2) with 16 cell types. Each cell type other than cells  $c1$  and  $c2$  are generated such that they have no relationship to outcome. We modify the clustering of  $c2$  around  $c1$  in the positive group, and examine the rate at which the variable importance estimated for the  $c1\_c2$  spatial interaction exceeded the various thresholds (red/orange lines). By changing the standard deviation in the Thomas process for placing  $c2$  points around  $c1$  points, we were able to investigate whether the approach is able to detect the presence of a relationship when the cells either cluster or are more dispersed across the binary outcomes. The resulting power curves, based on 100 simulations for each setting, are shown in Figure 5.8. These show the rate at which the variable importance estimates exceed

Table 5.1: False discovery table. Table showing empirical false discovery rate based on 100 independent simulations for both  $T = 4$  and  $T = 16$  cell types with 34 patients split into positive and negative groups. The false discovery rate is computed as the percentage of simulations for which any VI exceeded the corresponding threshold or both thresholds (Variable Importance Only), any of the cross-validated one standard error intervals for the VI lies to the right of the corresponding threshold (Variable Importance-1 CVSD), and whether the largest VI computed exceeded the corresponding threshold (Largest Variable Importance).

| Empirical False Discovery Rate:<br>Nominal Rate 5% |       |               |      |
|--|-------|---------------|------|
| Threshold:   | Noise | Interpolation | Both |
| Variable Importance Only                           |       |               |      |
| $T = 4$  | 0.35  | 0.15          | 0.07 |
| $T = 16$   | 1.00  | 0.74          | 0.09 |
| Variable Importance- 1 CVSD                        |       |               |      |
| $T = 4$  | 0.23  | 0.06          | 0.04 |
| $T = 16$   | 0.99  | 0.05          | 0.04 |
| Largest Variable Importance                        |       |               |      |
| $T = 4$  | 0.35  | 0.05          | 0.05 |
| $T = 16$   | 1.00  | 0.03          | 0.03 |

the 95% noise threshold, interpolation threshold, and both the noise and interpolation thresholds. In reference to Figure 5.8, the underlying standard deviation in the Thomas process relating  $c_1$  and  $c_2$  takes the value of 0.025 in the “no relation” case. As such, for this value it is desired that the noise and interpolation thresholds are exceeded no more than  $\alpha = 0.05$  proportion of the time. Otherwise, detecting that standard deviation values smaller/larger than 0.025 lead to increased/reduced clustering of  $c_1\_c_2$  in the positive groups is desired.

We saw that each threshold appeared to be effective in detecting clustering or dispersion relationships, such that even relatively small changes were detected with high frequency, even when considering many variables, see also Appendix B.6.

Table 5.1 further shows the empirical false discovery rate computed from 100 independent simulations for 4 and 16 cell types with a simulated population of 34 patients evenly split between positive and negative response groups. In this case all spatial interactions and meta variables were simulated independently of the response. The false discovery rate

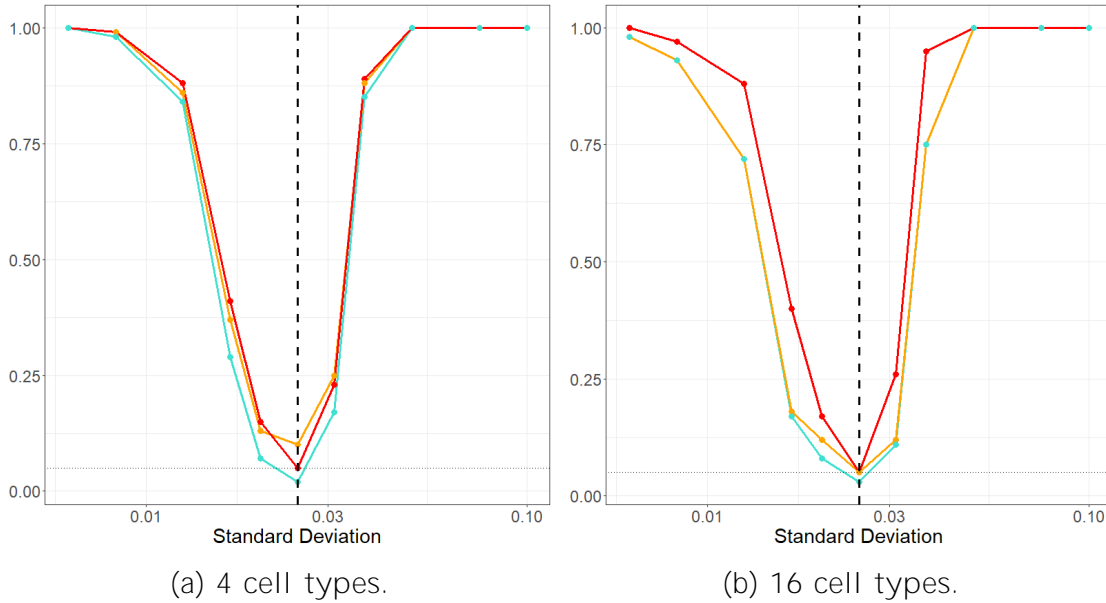


Figure 5.8: Power curves. Power curves showing the empirical rate, from 100 simulations, that the variable importance for the spatial interaction  $c1\_c2$  exceeded the 95% noise threshold, interpolation threshold and both the noise and interpolation thresholds, indicating a significant spatial interaction is detected. The curves indicate the threshold method: above both thresholds (teal), above the curved interpolation threshold (orange), and straight noise threshold (red) (the colours correspond with the noise curves used in the variable importance plots). The  $x$  axis gives the standard deviation parameter controlling the distribution of  $c2$  cells around  $c1$  cells. The vertical black dotted line is the base case, when both classes exhibit the same interactions across all cell types, including the  $c1\_c2$  interaction. To the right of this line,  $c2$  cells are less densely clustered around  $c1$  cells, and to the left of the line there is increased clustering around  $c1$  cells. The light horizontal, dotted gray line indicates the desired mis-classification rate when no signal is present (i.e. 0.05). (a) Spatial data with 4 cell types is used to create the curves. (b) Spatial data with 16 cell types is used to create the curves. Both images show the method is effective at correctly detecting when the important interaction does and does not differ between the patient outcomes. This is seen as all lines quickly climb to 1 (perfect detection of a signal) as the standard deviation parameter moves further from the no effect case (vertical line). The size and power is similar between the simulations despite the large increase in total interactions considered.

is computed as the percentage of simulations for which any VI exceeded the corresponding threshold or both thresholds. We also considered “Variable importance - 1 SDCV”, in which the false discovery rate is computed as the percentage of simulations in which any of the cross-validated one-standard-error intervals for the VI lies to the right of the corresponding threshold, and “Largest Variable Importance”, in which we only considered whether the largest VI computed exceeded the corresponding threshold.

These results suggest that while either of the two thresholds alone are not suitable to control the false discovery rate, comparing the VI along with the one standard error cross-validation interval to *both* thresholds (red and orange lines) yielded a false discovery rate very close to the nominal rate. Moreover, comparing the largest VI to both thresholds controlled the rate of falsely identifying the spatial interaction or meta-variable associated with that as being statistically important.

In Appendix B.6, we provide additional simulation evidence illuminating the effect of increasing the number of cells considered. These results suggested that the false discovery rate is not influenced by the number of cells considered, and remains controlled. The power to detect a single, important interaction was observed to decrease as the number of cells considered increased, as expected.

### 5.3.2 Applications to MIBIScope data

In this section we present two applications of the proposed methods to MIBIScope data sets. The first investigates known clusters in tumors related to triple negative breast cancer tumors, while the second investigates unknown relations in tumors related to lung cancer. Unless otherwise stated, all parameter values for the model discussed previously remain the same, see section 5.3.1.

#### 5.3.2.1 Triple negative breast cancer

We investigate the TNBC data set in [Keren et al. \(2018\)](#). This data was obtained via a MIBIScope, and the authors employ a mixing score to categorise tumors based on their TME. Although we focus on their mixing score, additional information is also available in [Keren et al. \(2018\)](#). The mixing score they use was defined as the proportion of immune cells touching tumor cells, and was calculated as the number of immune-tumor interactions divided by the number of immune-immune interactions for an image. They separate tumors into “compartmentalised” and “mixed” groups (and a “cold” group that we ignore), such that compartmentalised tumors tend to have tumor cells aggregated together with

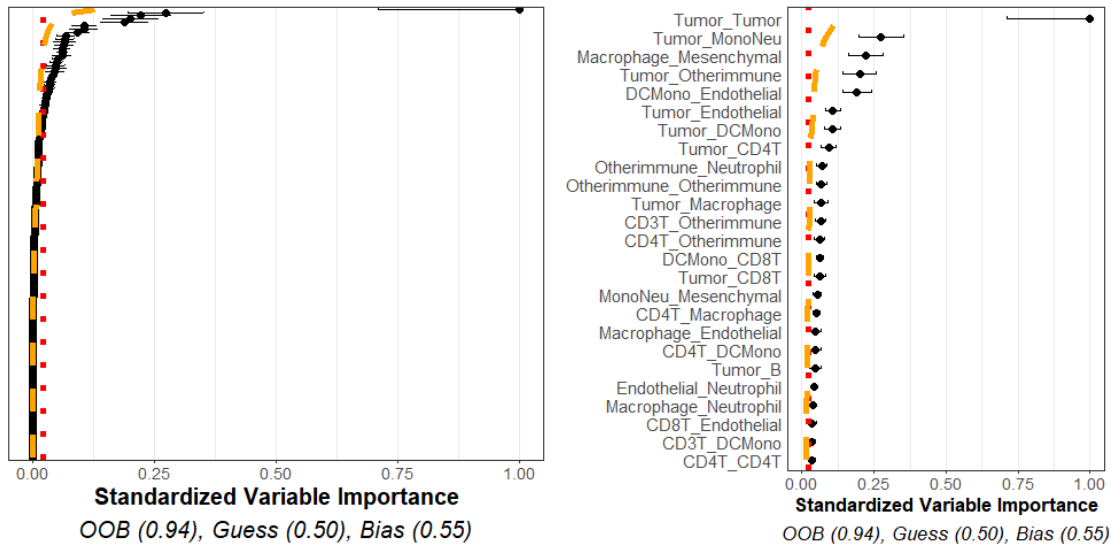
immune cells located around or away from the tumor cells, and mixed tumors tend to have tumor cells and immune cells mixed together. This makes a useful test data set on which to employ our method, as it provides two tumor groups that explicitly have different TMEs. The data set contains 33 patients with a single image per patient, 18 mixed tumors and 15 compartmentalised tumors. We define the outcome  $Z$  as 0 for mixed and 1 for compartmentalised, and we wish to predict the outcome using interactions between 16 cell phenotypes in the images and an additional meta-variable, age.

One tuning parameter with  $K$  functions is the maximum radius  $R$ . For this data we investigated using several options—25, 50, 100, 500, and 1000 micrometers. In all cases the conclusions were comparable, illustrating a robustness to the choice. The only (expected) difference is that as  $R$  increases, additional principal components,  $d$ , are recommended to capture the variations in the functions. With this observation and domain knowledge,  $K$  functions up to a maximum radius of 50 micrometers were used.

The results of this analysis are shown in Figure 5.9. As expected, many interactions are shown to be non-significant. However, there are a reasonable number above both thresholds and the predictive accuracy estimates agree that there are some important relationships in the data. Specifically, tumor cell-tumor cell interactions came as the top interaction, consistent with the characteristics of the “compartmentalised” tumors where few immune cells infiltrate the tumors and tumor cells are densely packed.

We also applied the method introduced in [Vu et al. \(2022\)](#), which is comparable to fitting a functional generalized linear model (logistic regression) to the data and evaluating for the parameter significance corresponding to each interaction. When all interactions are included in such a model, no significant variables at 5% level were identified. After sequentially removing the least important variables based on the variable importance ranking supplied by `funkycells`, this approach was only able to detect the significance of the Tumor\_Tumor interaction when it was the only variable included in the model.

Although the method we present here can indicate important relationships, it does not quantify the type of differences. While meta-variables can be easily compared using traditional statistical methods, the  $K$  functions are more difficult to analyze. To this end, we can also consider plots of the  $K$  functions. Figure 5.10 examines the Tumor\_Tumor interaction, which is found to be significant, and the CD4T\_Endothelial interaction, which was found to be insignificant. The significant  $K$  functions seem to be well separable (i.e.  $K$  functions are clustered with  $K$  functions having the same outcome), while the insignificant  $K$  functions are not easily separable between outcome groups, and have high variability in individual  $K$  functions of the same group.



(a) Full variable importance plot. (b) Top 25 variable importance plot.

Figure 5.9: TNBC variable importance. Variable importance plot and random forest model summary for predicting “compartmentalized” versus “mixed” tumor types with the TNBC data. (a) Figure with the variable importance values for all variables. (b) Figure with only the top 25 largest variable importance values. The OOB far exceeds those of naïve models, and many of the spatial interactions between tumor cells and immune cell populations exhibited significant variable importance values, suggesting important interactions in the data (such as Tumor\_Tumor).

### 5.3.2.2 Lung adenocarcinoma versus lung squamous cell carcinoma

We also applied our approach to attempt to predict different pathological subtypes of non-small cell lung cancer. In this section, we aim to measure to what degree the TME of two of the most common subtypes of lung cancer, LUAD and LUSC, can be differentiated using the spatial relationships between phenotyped cells as characterised by  $K$  functions. Our dataset contained 44 LUAD and 20 LUSC samples stained with antibodies against 35 proteins to enable the phenotyping of tumor cells, fibroblasts, and 10 immune cell subsets after scanning on the MIBIScope. The cells were pre-processed and classified and a summary of this data is given in Appendix B.1.  $K$  functions were computed between each cell type with again a maximum radius of 50 micrometers.

The variable importance plot and model summary when using our method to predict LUAD versus LUSC are shown in Figure 5.11. The results suggest that no spatial inter-

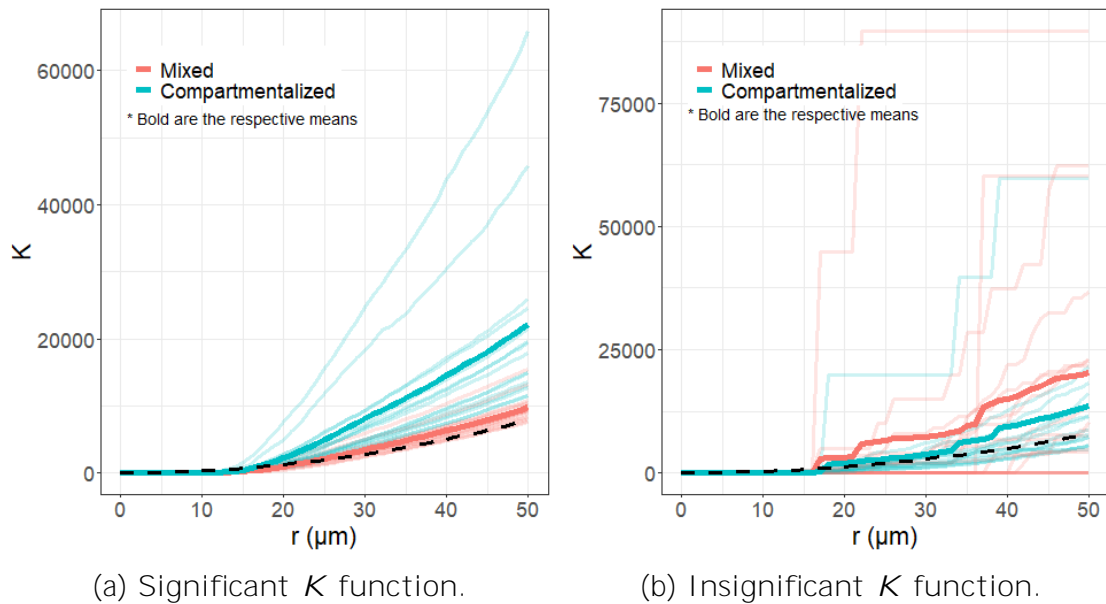
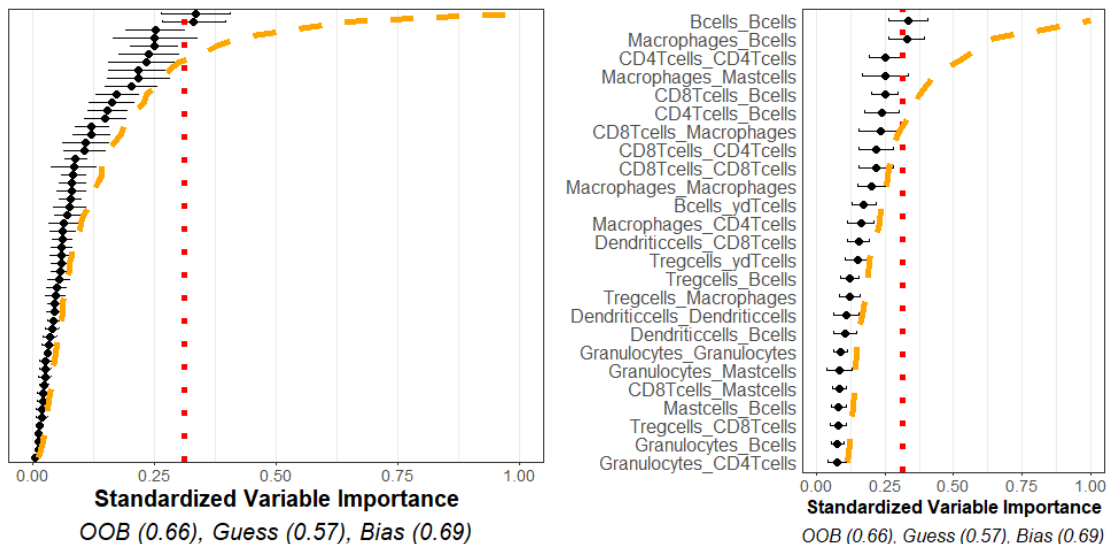


Figure 5.10: Example TNBC  $K$  functions. The  $K$  functions from the different outcomes are compared in these two plots. In both plots, the  $x$ -axis indicates the radial distance,  $r$ , in micrometers and the  $y$ -axis is the value of the  $K$  function. The lightly colored lines are the  $K$  functions for each patient, while the bold lines indicates the average (point-wise mean). In the figures, red indicates the mixed tumors while blue indicates the compartmentalized tumors. The black dashed line indicates the curve of a totally spatially random process for reference. (a) Plots the  $K$  functions for the Tumor\_Tumor interaction, which was found to have significant differences in the outcomes. (b) Plots the  $K$  functions for the CD4T\_Endothelial interaction, which was found have no significant differences between the outcomes. In (a), as expected, the compartmentalized group has relatively larger  $K$  functions—indicating increased clustering—and the functions are well grouped together. Conversely, (b) shows no clear differences between the  $K$  functions of the two groups and  $K$  functions are generally surrounded by  $K$  functions from patients of an assortment of the groups. That is,  $K$  function patterns vary widely even within the same outcome groups.

actions are significantly useful in distinguishing LUAD versus LUSC. The OOB accuracy was observed to be on-par with a naïve method, and none of the variable importance measurements exceeded the 95% quantile of what we would expect from independent point patterns. This indicates our method using homogeneous  $K$  functions applied to these specific cell phenotypes is unable to differentiate between the TME of LUAD and LUSC cancer types using this data.



(a) Full variable importance plot. (b) Top 25 variable importance plot.

Figure 5.11: Lung cancer variable importances. Variable importance plot and random forest model summary for predicting LUAD versus LUSC tumor types. (a) Figure with the variable importance values for all variables. (b) Figure with only the top 25 largest variable importance values. The OOB is similar to a naïve model, and none of the measured variable importance value were statistically significant, indicating no significant variable interactions or meta-variables.

For the purpose of comparison, we also fit functional linear models akin to those proposed in [Vu et al. \(2022\)](#) to this data. When all interactions were included, none of the linear model parameters were significant at the 5% level. After sequentially removing the least important variables down to the top 10 based on the variable importance ranking supplied by `funkyce11s`, we interestingly observed a, apparently spurious, significant interaction associated to spatial distribution of CD8Tcells. In an apparent contradiction, such models based on reducing the number interactions further, including examining a model with only the CD8Tcells\_CD8Tcells interaction, all suggested no significant interactions.

This analysis is presented to demonstrate the application of the above methods to real data, and emphasize that while naïve or repeated applications of existing methods might lead to spurious findings, our approach is built to control the rate of false discovery of important spatial interactions in the TME. The lack of significance in this example suggests though the need to consider a variety of new methodological approaches when analysing complex TME data. It also highlights the potential limitation of modeling cell interactions

through  $K$ -functions, which intrinsically assume a degree of spatial homogeneity in the spatial distribution of cells. These issues are discussed further in the following section.

## 5.4 Discussion

In this chapter, we present a novel method for the analysis of data-rich TME spatial data. We consider the case in which we compare TME and meta data or clinical data from two different patient groups, and develop a model to identify significant differences between these groups in the distribution of cell phenotypes, protein antigens, or a mixture thereof. In addition, the model aims to predict which group a new patient is in using their TME and meta data. Our model employs a combination of spatial statistics and [functional data analysis](#) and is applicable to marked point processes in general. Benefits of the model include general applicability, with few tuning parameters, and easily visualised and interpreted output. We find our model to be robust to the choice of the tuning parameters. The model demonstrates a powerful ability to identify important variables while maintaining good predictive power. It is also resistant to overfitting and manages false discovery rate.

We evaluate our approach on simulated data, demonstrating the effectiveness of our method for marked spatial point patterns with known interactions. Our approach is then applied to TNBC and lung cancer data obtained from multiplexed ion beam imaging. For the TNBC data, we compare two groups of tumors that are separated into “compartmentalised” and “mixed” groups based on the degree of tumor-immune cell interactions. This separation of tumors gives two groups with explicit differences in their TMEs, making a useful data set for the demonstration of our model. The model demonstrates good predictive accuracy, and identifies expected specific cell phenotypes relationships, and interactions of interest. The lung cancer data shows our model can also detect lack of differences in cell phenotypes relationships.

Throughout this chapter we have assumed homogeneity of the underlying spatial point processes in defining the  $K$  functions used in the model. Spatial homogeneity is defined such that the intensity of a given mark is independent of spatial location ([Illian et al., 2008](#); [Diggle, 2013](#); [Baddeley et al., 2015](#)). Whilst such an assumption may be reasonable in some cases, given the complexity of the TME, homogeneity may not always apply. We note, however, that since our approach compares  $K$  functions between different groups, rather than against the theoretical  $K$  functions associated with complete spatial randomness (as is typical in other circumstances), the lack of underlying spatial homogeneity in the TME for a data set may not be overly problematic. Statistical tests for homogeneity exist ([Illian](#)

et al., 2008; Diggle, 2013; Baddeley et al., 2015). Inhomogenous  $K$  functions can be used in an attempt to mitigate the issue of inhomogeneity (Baddeley et al., 2000; Illian et al., 2008; Diggle, 2013; Baddeley et al., 2015). Regardless of how  $K$  functions are defined they are amenable to being used in our methodology. Furthermore, whilst we have focused throughout this chapter on  $K$  functions, we note that the methods presented here can be applied to any summary functions from spatial statistics, or indeed any functions in general. See, for example, suggestions in Ripley (1977), Boots and Getis (1988), and Zhu et al. (2020).

In addition, we note that  $a_{c;t}^{(p;t)}$  in Eq (5.1), may consist of the raw measured protein expression level for each protein, and may also include other cell information (e.g. cell size). Our approach could potentially be adapted for analyzing raw protein expression levels via the use of mark-weighted  $K$  functions (Penttinen et al., 1992; Illian et al., 2008). Consequently, the approach may be useful in methods applied to cell phenotypes based on proteins (OPAL, MIBI, Phenocycler Fusion) or transcripts (Xenium, Cosmx, MERscope) expression (Wang et al., 2012; Black et al., 2021; Chen et al., 2022; Williams et al., 2022; Li et al., 2022; Canning et al., 2022; Sadeghirad et al., 2023). Usage in non-cellular contexts is also possible.

We have also assumed in this chapter outcomes which are categorical, or classes in a group. In this way, we say the model performed classification. Although we considered two classes, a larger number of classes is directly possible. Further extension of this method to real-valued data, e.g. survival time, is likewise natural. Random forest models designed for continuous or survival time analysis exist, e.g. Breiman (2001), Ishwaran et al. (2011), and Pickett et al. (2021), and metrics such as  $L^2$  error can be used in place of the difference function of the OOB and naïve estimates.

# Chapter 6

## The Maximally Selected Likelihood Ratio Test in Random Coefficient Models

### 6.1 Introduction

In this chapter, we focus on scalar data to complement and extend the results in [Horváth and Trapani \(2022\)](#), by studying the maximally selected Likelihood Ratio test to detect changepoints in a (scalar) Random Coefficient Autoregressive model of order 1 (RCA(1) henceforth):

$$y_i = (\alpha_i + \beta_{i,1})y_{i-1} + \beta_{i,2}\epsilon_i, \quad 1 \leq i \leq n \quad (6.1)$$

In Eq (6.1), we assume that the process starts from an initial value  $y_0$  (for the sake of simplicity we assume that  $y_0$  is non random). In particular, we test for the null of no change in the deterministic part of the autoregressive root

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_n \quad (6.2)$$

versus the alternative of [at most one change \(AMOC\)](#)

$$H_A: \alpha_1 = \alpha_2 = \dots = \alpha_k \neq \alpha_{k+1} = \dots = \alpha_n \quad (6.3)$$

where  $k$  is the unknown time of change.

The RCA model was firstly studied by [Anděl \(1976\)](#) and [Nicholls and Quinn \(2012\)](#). It belongs in the wider class of nonlinear models for time series, which have been proposed

“as a reaction against the supremacy of linear ones - a situation inherited from strong, though often implicit, Gaussian assumptions” (Akharif and Hallin, 2003). Model (6.1) allows for the autoregressive “root”  $\rho + \beta_1$  to vary over time, and thus also for the possible co-existence of stationary and nonstationary regimes. The RCA also nests the so-called Stochastic Unit Root (STUR) model, where  $\beta_1 = 1$  and  $\text{Var}(\epsilon_{i,1}) > 0$ , where  $y_i$  can “[...] be stationary for some periods, and mildly explosive for others” (Granger and Swanson, 1997, p. 35); this behaviour may be a more appropriate characterisation for many time series (see, *inter alia*, Leybourne et al., 1996, Lieberman, 2012, Lieberman and Phillips, 2017 and Lieberman and Phillips, 2020). Further, Tsay (1987) shows that the RCA model is a second-order equivalent to (i.e., it has the same conditional variance as) the popular ARCH(1) model by Engle (1982), thus ensuring that Eq (6.1) can accommodate for conditional heteroskedasticity, often found in many datasets. Finally, the RCA model shares many similarities with, and virtually the same probabilistic structure as, the Double Autoregressive (DAR) model, which has an autoregressive structure both in the conditional mean and in the conditional variance (we refer, *inter alia*, to the contributions by Borkovec and Klüppelberg, 2001, Ling, 2004, Chen et al., 2014, and Cavaliere and Rahbek, 2021).

Indeed, there are several examples of datasets which could exhibit, as time elapses, both stationary and nonstationary/explosive behaviours. As a leading example, in the financial econometrics literature explosive regimes may arise (or, if present, collapse) in the case of a bubble, where asset prices, are characterised by explosive growth; interestingly, we note that the bubble component of asset prices is modelled as an explosive RCA(1) - see e.g. Diba and Grossman (1988). Testing for bubbles ex-post, or detecting bubbles on-line, have both been paid significant attention by the financial econometric literature in recent years, and we refer, *inter alia*, to the contributions by Phillips et al. (2011), Homm and Breitung (2012), and Astill et al. (2021). However, whilst testing for a change from a non-explosive towards an explosive regime (corresponding to the inception of a bubble) is well-developed (see, *inter alia*, the contribution by Shi and Phillips, 2022b), testing for a change from an explosive to a different (explosive or non-explosive) regime is a far more complicated task. Phillips et al. (2015a) and Phillips et al. (2015b) propose a methodology to consistently date bubbles collapsing based on a first-crossing principle (see also the related contributions by Phillips and Shi, 2018, and Phillips and Shi, 2019). However, in the context of an AR(1) model, developing a formal test would require studying the asymptotics under explosiveness in an AR model, which is bound to depend on several nuisance parameters. Conversely, as we discuss below, this is not an issue when considering an RCA(1) model. As a second example which speaks to the versatility of the RCA(1) model, in epidemiology, the famous SIR model (Kermack and McKendrick, 1927) stipulates that the number of infections in an epidemic outbreak will display explosive growth, whilst

having mean reverting behaviour when the epidemics subsides. This is regulated by the so-called “(basic) reproduction rate”  $R_0$ , which is greater than 1 during an outbreak, and smaller than 1 when the epidemics is waning off. The insightful analysis in [Magdalinos et al. \(2022\)](#) shows that - in a reduced form framework - there is a one-to-one mapping between  $R_0$  and the first order autoregressive coefficient in an AR(1) model for the number of infections: hence, the importance of entertaining the possibility of having explosive and stationary regimes in an epidemiological time series. Finally, we point out that, in addition to the examples above, Eq (6.1) has been used in several other applied sciences, including biology ([Stenseth et al., 1998](#)), medicine ([Fryz, 2017](#)), physics ([Ślęzak et al., 2019](#)), and economics (see the review by [Regis et al., 2022](#)).

A major advantage of the RCA(1) model - compared to a standard AR(1) - is the fact that, in the context of Eq (6.1), the asymptotic distribution of the estimator of  $\rho_0$  is Gaussian irrespective of whether  $y_i$  is stationary or not. This has been shown in several contributions considering different estimation techniques. To cite just a few, [Hwang and Basawa \(2005\)](#) show that the WLS estimator is asymptotically normal also in the nonstationary case; [Hill and Peng \(2014\)](#) derive a similar set of results for the Empirical Likelihood estimator, and [Aue et al. \(2006\)](#), [Berkes et al. \(2009\)](#), [Aue and Horváth \(2011\)](#) study the Quasi Maximum Likelihood estimator (QML; see also a recent contribution by [Benmoumen and Salhi, 2021](#) on QML estimation for a stationary RCA( $\rho$ ) model). In light of the flexibility of Eq (6.1), which can accommodate for both stationary and explosive regimes, and in light of the empirical relevance of being able to conduct inference irrespective of whether data are stationary or not, it can be envisaged that using the RCA framework is a very convenient way of modelling “long” time series, where the co-existence of stationary and explosive periods is likely to have happened. Hence also the importance of being able to detect changepoints in Eq (6.1), both in order to have consistent estimates in the various regimes of  $y_i$ , and to identify which regimes are stationary and which ones are not. As mentioned above, testing for changepoints in the presence of explosive regimes with an AR(1) model is a highly nontrivial task; however, the RCA model always ensures asymptotically Gaussian inference, and it is therefore immune from the difficulties encountered in the AR(1) case.

Still, whilst estimation is well-developed in the RCA context, tests for changepoint are more rare. [Lee \(1998\)](#), [Lee et al. \(2003\)](#) and [Aue \(2004\)](#) develop tests, based on the unweighted cumulative sum (CUSUM) process, but their results are limited to the stationary regime. To the best of our knowledge, the only paper which studies changepoint detection irrespective of stationarity or explosivity in an RCA model is [Horváth and Trapani \(2022\)](#), who use the WLS estimator. In this chapter we complement and refine the results in [Horváth and Trapani \(2022\)](#) by proposing a procedure based on the QML esti-

mator, and on using a sequence of Likelihood Ratio (LR) statistics, to test for the null of no changepoints (6.2). Similarly to Horváth and Trapani (2022), our test does not require any assumption, or prior knowledge, as to whether  $y_i$  is stationary or explosive. Technically, we derive the limiting distribution of the LR test, showing that this follows, asymptotically, a Darling-Erdős type (Darling and Erdős, 1956) distribution. As mentioned above, this holds true irrespective of whether  $y_i$  is stationary or not. Our tests are shown to have the correct size, and very good power, by simulations; we illustrate them through two empirical applications to economic and epidemiological data.

Our simulations show that using the sequence of LR statistics affords superior power compared to the CUSUM proposed in Horváth and Trapani (2022), and even bigger improvements are noted after implementing a finite sample correction for the critical values. Heuristically, this can be expected: the LR test is based on using the QML estimator, as opposed to the WLS estimator, which is employed in the case of the CUSUM. Although we are not aware of any theoretical result on efficiency in this context, Monte Carlo evidence reported in several papers (see e.g. Koul and Schick, 1996) indicates that QML should be efficient, hitting the Cramér-Rao lower bound, if the density of the innovations  $(\theta_{0,1}; \theta_{0,2})$  is correctly specified - whereas, if this is not the case, our simulations show that QML still performs very well, although more efficient estimators can be constructed (see e.g. a related contribution by Berkes and Horváth (2004), where the case of a GARCH(1,1) model is studied). Heuristically, the improvements obtained with QML with respect to the WLS estimator can be ascribed to the use of a different weighing scheme: the WLS weights each observation by  $1 = (1 + X_i^2)^{-1}$ ; conversely, the QML is based on a likelihood function where the more natural weights  $1 = (Var(\theta_{0,2}) + Var(\theta_{0,1}) X_i^2)^{-1}$  are used. As mentioned above, we are not aware of any theoretical results on estimation efficiency, nor are we aware on any results on testing efficiency (see, however, Davis et al., 1995); however, our simulations show that using the LR statistic (based on the QML constructed under the assumption of Gaussian errors) yields size control, and superior power, even when the density of the innovations is not correctly specified. The maximally selected LR statistic - if the innovations are independent and identically distributed (IID) standard normal - is equivalent to the standardised CUSUM process; given that our assumptions - modulo some requirements on the parameter space - are the same as Horváth and Trapani (2022), this suggests that the maximally selected LR statistic dominates CUSUM-based tests for changepoint.

As a final technical remark, we would like to point out that studying the LR statistic in our context is a highly non-trivial exercise. In order to derive the weak limit of the maximally selected LR statistic, we need to derive asymptotic approximations of the likelihood function that are uniform across  $i$  (in essence, our theorems need to control for two likelihoods at each point in time, one computed before and one after each  $i$ ). This is

further compounded by the fact that, in the nonstationary case  $E \log j + \rho_{0,1} j > 0$ , one of the parameters (the variance of  $\rho_{0,2}$ ) cannot be identified, which requires deriving bounds, for the asymptotic approximations, that are uniform also with respect to  $Var(\rho_{0,2})$  - we refer to section 6.2 for a more thorough discussion of this issue.

The remainder of the chapter is organised as follows. We present our main test statistic, and its asymptotics, in section 6.2. Monte Carlo experiments are reported in section 6.3, and the empirical illustrations are in section 6.4. Section 6.5 concludes. Technical lemmas and proofs are reported in the appendix.

## 6.2 The test statistics

In order to check for the possible presence of changepoints, we use the QML estimator studied in Aue et al. (2006) and Berkes et al. (2009), and propose a test based on the likelihood ratio (LR) statistic to test  $H_0$  against  $H_A$ . We begin by noting that, under the assumptions that the errors  $f(i;1; i;2); 1 < i < T$  are zero mean, IID Gaussian random vectors with  $E \rho_{0,1} \rho_{0,2} = 0$ , and that the parameter  $\rho_0$  changes once after the  $k$ -th observation, the conditional log likelihood function (up to a constant term) is given by

$$L_{n;k}(\mathbf{u}) = \sum_{i=1}^k \ell_i(u_1; u_2; u_3) + \sum_{i=k+1}^n \ell_i(u_4; u_2; u_3); \quad (6.4)$$

where  $\mathbf{u} = (u_1; u_2; u_3; u_4)^\top$  and

$$\ell_i(v_1; v_2; v_3) = \frac{1}{2} \left( \log(v_2 y_{i-1}^2 + v_3) + \frac{(y_i - v_1 y_{i-1})^2}{v_2 y_{i-1}^2 + v_3} \right); \quad (6.5)$$

Note that, in the derivation of the likelihood ratio, we assume - for now - that the variances of the errors remain the same.

Eq (6.4) is maximised at the (partial sample) maximum likelihood estimators  $\hat{\rho}_{k,1}, \hat{\rho}_{k,1}^2, \hat{\rho}_{k,2}^2$  and  $\hat{\rho}_{k,2}$ , defined as

$$\max_{\mathbf{u} \in \mathcal{U}} L_{n;k}(\mathbf{u}) = L_{n;k}(\hat{\rho}_{k,1}, \hat{\rho}_{k,1}^2, \hat{\rho}_{k,2}^2, \hat{\rho}_{k,2});$$

where

$$\mathcal{U} = \{\mathbf{u} : \underline{u}_1 \leq u_1 \leq u_4 \leq \bar{u}_1, 0 < \underline{u}_2 \leq u_2 \leq \bar{u}_2, 0 < \underline{u}_3 \leq u_3 \leq \bar{u}_3\};$$

Also, as far as the notation is concerned, if there is no change in the parameters (i.e., if  $H_0$  holds), we use the notation  $\hat{\alpha}_n = (\hat{\alpha}_{n;n}, \hat{\alpha}_{n,1}^2)$  and  $\hat{\alpha}_{n,2}^2$ , with

$$\max_{\mathbf{u} \in \mathbf{U}} L_{n;n}(\mathbf{u}) = L_{n;n}(\hat{\alpha}_{n;n}, \hat{\alpha}_{n,1}^2, \hat{\alpha}_{n,2}^2); \quad (6.6)$$

where  $L_{n;n}(\mathbf{u}) = L_{n;n}(\mathbf{u})$ ,  $\mathbf{u} = (u_1; u_2; u_3)$  and

$$\mathbf{U} = \{ \mathbf{u} : \underline{u}_1 \leq u_1 \leq \bar{u}_1; 0 < \underline{u}_2 \leq u_2 \leq \bar{u}_2; 0 < \underline{u}_3 \leq u_3 \leq \bar{u}_3 \};$$

The likelihood functions are different under the null and under the alternative since we have 3 parameters under the null and 4 under the alternative.

For a given  $k$ , a test for the null of no change occurring at  $k$  can be constructed based on (twice) the log of the ratio between the unrestricted likelihood in Eq (6.4), and the restricted one in Eq (6.6):

$$\hat{T}_n(k) = 2 \left( L_{n;n}(\hat{\alpha}_{n;n}, \hat{\alpha}_{n,1}^2, \hat{\alpha}_{n,2}^2) - L_{n;k}(\hat{\alpha}_{k,1}, \hat{\alpha}_{k,1}^2, \hat{\alpha}_{k,2}, \hat{\alpha}_{k,2}^2) \right); \quad (6.7)$$

Hence, changepoint detection can be carried out by considering the “worst case scenario”, represented by the maximally selected log likelihood ratio

$$T_n = \max_{1 \leq k < n} \hat{T}_n(k); \quad (6.8)$$

where the null hypothesis (6.2) is rejected for large values of  $T_n$ .

## 6.2.1 Assumptions

Before we state our main result, we list the necessary conditions. We use  $\alpha$  for the common value of the autoregressive parameter under the null hypothesis.

*Assumption 6.2.1. It holds that: (i)  $f_{i,1}, f_{i,2}; 1 < i < 1/g$  are independent and identically distributed random vectors; (ii)  $E f_{i,1} = E f_{i,2} = 0$ ,  $E f_{i,1}^2 = \frac{2}{1}$ ,  $E f_{i,2}^2 = \frac{2}{2}$ ,  $E f_{i,1} f_{i,2} = 0$ ,  $E j_{0,1} j < 1$  and  $E j_{0,2} j < 1$  with some  $\alpha > 2$ ; (iii)  $1 - E \log j + \alpha_{0,1} j < 1$ , and  $E \log j + \alpha_{0,1} j \notin 0$ ; (iv) if  $E \log j + \alpha_{0,1} j > 0$ , then  $Pf(\alpha + \alpha_{0,1})y_0 + \alpha_{0,2} = c g = 0$  for all  $c$ .*

Parts (i) and (ii) are standard (see e.g. [Aue and Horváth, 2011](#)); note that, in part (i), we do not assume that  $f_{i,1; 1 < i < 1 g}$  and  $f_{i,2; 1 < i < 1 g}$  are independent sequences, but only uncorrelated. Assumption 6.2.1 (iv) is taken from [Berkes et al. \(2009\)](#), where it is used to show that  $y_i$  does not converge to a degenerate sequence in the explosive case. As Theorem 6 in [Berkes et al. \(2009\)](#) shows, if the sequences  $f_{i,1; 1 < i < 1 g}$  and  $f_{i,2; 1 < i < 1 g}$  are independent, then Assumption 6.2.1 (iv) can be replaced with  $Pf_{0,2} = cg < 1$  for all  $c$ . Finally, part (iii) of the assumption allows for both stationary and nonstationary regimes. In particular, [Anděl \(1976\)](#) and [Quinn \(1982\)](#) showed that  $E \log j + \rho_{0,1} j < 0$  is the necessary and sufficient condition for stationarity; in that case,  $y_i$  converges exponentially fast to the unique, nonanticipative stationary solution of

$$y_i = (\rho_{0,1} + \rho_{i,1})y_{i-1} + \rho_{i,2} \quad (6.9)$$

Conversely, when  $E \log j + \rho_{0,1} j > 0$ ,  $y_i$  is non-stationary, and it holds that  $jy_j \rightarrow \infty$  a.s. at an exponential rate ([Berkes et al., 2009](#)).

As far as the parametric space is concerned, we make the following standard assumption.  
 Assumption 6.2.2. *It holds that  $(\rho_{0,1}; \rho_{1,1}; \rho_{1,2})$  is in the interior of  $\mathbf{U}$ .*

## 6.2.2 Asymptotics

We are now ready to present the asymptotics under the null. Let

$$a(x) = (2 \log x)^{1-2} \quad \text{and} \quad b(x) = 2 \log x + \frac{1}{2} \log \log x - \frac{1}{2} \log 2$$

Our main result is the asymptotic distribution of  $T_n^{1=2}$  under the null hypothesis.

Theorem 6.2.1. *Under  $H_0$ , if Assumptions 6.2.1 and 6.2.2 are satisfied, then we have*

$$\lim_{n \rightarrow \infty} P \{ a(\log n) T_n^{1=2} \leq x + b(\log n) \} = \exp(-2e^{-x});$$

for all  $-\infty < x < \infty$ .

Theorem 6.2.1 is our main theoretical result; in essence, it states that the asymptotic law of the suitably normalised  $T_n^{1=2}$  follows a Type I Extreme Value Distribution (EVT), whence asymptotic critical values can be computed as

$$c_{\alpha;n} = \frac{b(\log n) - \log(-\frac{1}{2} \log(1-\alpha))}{a(\log n)}; \quad (6.10)$$

such that, under  $H_0$ , it holds asymptotically that  $P(T_n^{1=2} \leq c; n) \rightarrow \alpha$ , where  $\alpha$  is the user-chosen nominal level of the test. The result holds irrespective of the value of  $E \log j + 0.1j$ , and therefore it can be employed with no prior knowledge as to whether the observations  $y_i$  are stationary or not. Further, due to the self-normalised nature of the likelihood ratio statistic, we do not need to estimate the variance of the weak limit of  $T_n$ . The result in Theorem 6.2.1 is not standard, and in Appendix 6.2.1 we report a full-blown derivation. Here, we would like to point out that - in order to obtain convergence to the Type I EVT - we show that, as  $n \rightarrow \infty$

$$P(T_n^{1=2} = \max_{1 \leq k \leq n} fT_a; T_b g) = 1;$$

where

$$T_a = \max_{1 \leq k \leq n} \frac{jW_1(k)j}{k^{1=2}} \text{ and } T_b = \max_{1 \leq k \leq n} \frac{jW_2(k)j}{k^{1=2}};$$

where  $fW_1(k); k \geq 1g$  and  $fW_2(k); k \geq 1g$  are two independent standard Wiener processes. Hence

$$P(\max_{1 \leq k \leq n} fT_a; T_b g \leq (x + b(\log n) = a(\log n))) = (P(T_a \leq (x + b(\log n) = a(\log n))))^2; \tag{6.11}$$

and the desired result follows from the application of Theorem 2 in Darling and Erdős (1956).

From a technical viewpoint, Theorem 6.2.1 holds under both the stationary case, i.e.  $E \log j + 0.1j < 0$ , and the explosive case,  $E \log j + 0.1j > 0$ . Whilst the final result is the same in both cases, the arguments to show the theorem are quite different. In the stationary case, we show that the likelihood function computed at the  $y_i$  is ‘‘very close’’ to the likelihood function computed at the stationary solution  $\bar{y}_i$ , so that, in our calculations, we can replace the former with the latter. We can subsequently follow the usual approach, based on Taylor’s expansion, to derive the consistency and the asymptotic law of the QML estimator; in particular, in order to derive our main result, we use a strong approximation for partial sums which we show can be employed for  $\bar{y}_i$  and transformations thereof. Conversely, the case of nonstationary observations is more delicate. In such a case, we note that, similarly to Berkes et al. (2009), the likelihood function inherits the explosive behaviour of the  $y_i$ , so that  $n^{-1}L_{n;k}(\mathbf{u}) \rightarrow 0$  in probability for all  $k$ . On the other hand, we show that, uniformly in  $k$

$$\sup_{\mathbf{u} \in \bar{\mathbf{U}}} \left| \frac{1}{k} \sum_{i=1}^k (f_i(u_1; u_2; u_3) - f_i(u_1; u_2)) \right| \rightarrow 0;$$

which entails that, whilst  $\beta_2$  is not estimable,  $\beta_1$  and  $\beta_0$  are. Thereafter, we are able to express the QML partial sample estimators of  $\beta_1$  and  $\beta_0$  as partial sum processes uniformly in  $u_3$ ; exploiting the exponential growth of  $y_i$ , we are then able to express the QML partial sample estimators of  $\beta_1$  and  $\beta_0$  as partial sum processes of the  $y_{i-1}$ s, thereby being able to use again strong approximations results. Naturally, this requires  $\text{Var}(y_{i-1}) > 0$  to avoid degeneracy (see also the discussion of the case  $\text{Var}(y_{i-1}) = 0$  in Horváth and Trapani, 2019); when  $\text{Var}(y_{i-1}) = 0$  and  $E \log |y_{i-1}| > 0$ , the LR statistic drifts to zero, thus rendering our asymptotics invalid. In this respect, the standard AR(1) model is not nested within the RCA framework,<sup>1</sup> and in the case of a non-random autoregression, the asymptotics differs between the stationary and the nonstationary cases.

As a final technical remark, we point out that, in order to derive our results, we need to rule out the boundary case  $E \log |y_{i-1}| = 0$ . This case is well-known to be problematic in all applications of nonlinear time series, and we refer, *inter alia*, to Horváth and Trapani (2019) for a discussion of this case, and to Francq and Zakočian (2012) and Horváth and Trapani (2022) for examples where the asymptotics breaks down. Technically, in our case, we require that  $\sum_{i=1}^n 1/(1 + y_{i-1}^2) = O_P(1)$ . This holds in the explosive case (heuristically, because in that case  $|y_i| \rightarrow \infty$  a.s. at an exponential rate as  $i \rightarrow \infty$ ), but it fails in the boundary case, because although  $|y_i|$  diverges (see Lemma 7.4 in Horváth and Trapani, 2016), it does not do it at a sufficiently fast rate. Indeed, it is possible to draw a parallel with the pure unit root case, i.e. the case  $\beta = 1$  and  $\beta_0 = 0$  a.s.: in that case, Eq (6.1) boils down to the random walk model, viz.

$$y_i = y_{i-1} + \varepsilon_i$$

By the Chung-Fuchs Theorem (Chung and Fuchs, 2008),  $y_i$  fluctuates arbitrarily close to zero infinitely often with probability one - hence,  $\sum_{i=1}^n 1/(1 + y_{i-1}^2) \rightarrow \infty$  a.s. as  $n \rightarrow \infty$ , which entails that our arguments could not be used in that case either. Thus, whilst we are able to propose a test whose limiting distribution can be used under both the stationary and the explosive regimes, our results ensure pointwise, rather than uniform, size control across the space of the nuisance parameter  $E \log |y_{i-1}|$ .

From a practical viewpoint, the critical values in Eq (6.10) may not be the best choice: indeed, it is well known (Hall, 1979) that convergence to the EVT is notoriously slow, and it is therefore important to find a better approximation for the critical values of  $T_n^{1=2}$ , especially for small and moderate sample sizes. Vostrikova (1981) provides the following approximation (see also the discussion in Chapter 1.3 in Csörgő and Horváth, 1997)

$$P \{ T_n^{1=2} \leq x \} \approx \frac{x \exp(-x^2/2)}{(2/x)^{1=2}} \left( \log n^2 - \frac{1}{x^2} \log n^2 + \frac{4}{x^2} + O\left(\frac{1}{x^4}\right) \right) : \quad (6.12)$$

---

<sup>1</sup>We are grateful to an anonymous Referee for his/her insight on this.

We note that, following our proof, it is possible to extend the main result in several directions. For example, and as is typical, in Theorem 6.2.1 we derive the limiting distribution of  $T_n^{1=2}$ . Davis et al. (1995) argue that using  $T_n$  may work better in simulations; in such a case, Theorem 6.2.1 can be readily extended to show that

$$\lim_{n \rightarrow \infty} P \{ a(\log n) T_n \leq x + b(\log n)g = \exp(-2e^{-x/2}) \}; \quad (6.13)$$

Similarly, in Theorem 6.2.1 we are interested only in the stability of  $\mu_i$ . Since the variance  $\sigma_{i,1}$  is also an estimable parameter irrespective of whether  $y_i$  is stationary or not, we can modify the original model to include changes in both  $\mu_i$  and  $\text{Var}(\sigma_{i,1})$ . As above, we consider the AMOC model for the sake of simplicity. Let

$$y_i = [(\sigma_{i,1}^{1/2} \epsilon_i | \mathcal{F}_{i-1}^k) + (\sigma_{i,1}^{1/2} \epsilon_i | \mathcal{F}_{i-1}^{k+1})] y_{i-1} + \sigma_{i,1}^{-1/2} \epsilon_i \quad i = 1, \dots, n; \quad (6.14)$$

with the normalisation  $E \sigma_{0,1}^2 = 1$ . According to Eq (6.14), the parameters change from  $(\mu_0, \sigma_{1,0}^2)$  to  $(\mu_A, \sigma_{1,A}^2)$ . We modify the likelihood function as

$$G_{n,k}(\mathbf{u}) = \sum_{i=1}^k \ell_i(u_1; u_2; u_3) + \sum_{i=k+1}^n \ell_i(u_4; u_5; u_3);$$

with  $\mathbf{u} = (u_1; u_2; \dots; u_5) \in \mathbb{R}^5$ , where  $\ell_i(v_1; v_2; v_3)$  is defined in Eq (6.5). Let  $Q_n$  denote  $2 \log$  of the maximally selected likelihood ratio. Following the proof of Theorem 6.2.1, one can show that if  $E \sigma_{0,1}^3 = E \sigma_{0,2}^3 = 0$ ,  $E j_{0,1}^{4+} < 1$  and  $E j_{0,2}^{4+} < 1$  for some  $\epsilon > 0$ , and  $E \sigma_{0,1}^4 = 3(E \sigma_{0,1}^2)^2$  and  $E \sigma_{0,2}^4 = 3(E \sigma_{0,1}^2)^2$ , it holds that

$$\lim_{n \rightarrow \infty} P \{ a(\log n) Q_n^{1=2} \leq x + b_2(\log n) \} = \exp(-2e^{-x});$$

for all  $x$ , where  $b_2(z) = 2 \log z + \log \log z$ .

### 6.2.2.1 Consistency

We briefly discuss the consistency of the test. We assume that there is exactly one change from  $\mu_0$  to  $\mu_A = \mu_{k+1}$  at time  $k$ , and we define the size of the change as  $\delta_n = \mu_A - \mu_0$ , where we emphasize that the break magnitude may be sample dependent - e.g., it may shrink as  $n \rightarrow \infty$ .

We require the following analogue of Assumption 6.2.2, before and after the change:

Assumption 6.2.3. *It holds that: (i)  $(\alpha; \frac{2}{1}; \frac{2}{2})$  is in the interior of  $\mathbf{U}$ ; (ii)  $(\alpha; \frac{2}{1}; \frac{2}{2})$  is in the interior of  $\mathbf{U}$ .*

Similarly, we need the analogue of Assumption 6.2.1 after the change:

Assumption 6.2.4. *It holds that: (i)  $1 - E \log j_{A+0,1} < 1$ , and  $E \log j_{A+0,1} \notin 0$ ; (ii) if  $E \log j_{A+0,1} > 0$ , then  $Pf_{(A+0,1)} y_0 + 0,2 = cg = 0$  for all  $c$ .*

Theorem 6.2.2. *Under  $H_A$ , if Assumptions 6.2.3-6.2.4 are satisfied, and if it holds that*

$$\lim_{n \rightarrow \infty} \frac{\frac{2}{n}k(n-k)}{n \log \log n} = 1; \tag{6.15}$$

then we have

$$a(\log n) T_n^{1-2} \rightarrow b(\log n) \neq 1;$$

Theorem 6.2.2 states that the maximally selected LR test is consistent versus the AMOC alternative. In particular, Eq (6.15) states that, when a break occurs mid-sample (i.e., with  $k = bn - c$  with some  $0 < c < 1$ ), then there is nontrivial power as long as the size of the break,  $c$ , is larger than  $O(n^{-1-2} (\log \log n)^{1-2})$ . On the other hand, when  $c$  is fixed, Eq (6.15) ensures that breaks are detected as long as  $k \sim (\log \log n)^{-1} \neq 1$ : breaks that are as close as  $O(\log \log n)$  periods to the sample endpoints can be detected. This result is similar to the results on the standardised CUSUM process, studied in Horváth and Trapani (2022).

## 6.3 Simulations

We begin by presenting the stationary region defined by the value of  $E \log j_{A+0,1}$ . This is depicted in Figure 6.1, where we have used Gaussian errors with 10,000,000 simulations for each estimate, with steps sizes of 0.01 for  $\alpha$  and  $E \frac{2}{0,1}$ . Outside of this region, the parameters give an explosive RCA(1) sequence.

Guided by the stationary region, we now turn to examining the empirical rejection frequencies, under both the null and the alternative, of our test. We consider using  $T_n^{1-2}$  with the asymptotic critical values Eq (6.10), based on Theorem 6.2.1 (this will be henceforth denoted as MLE), and we also use  $T_n^{1-2}$  with the critical values coming from the finite-sample approximation in Eq (6.12) (henceforth, we denote this as MLE\_Vost). Finally,

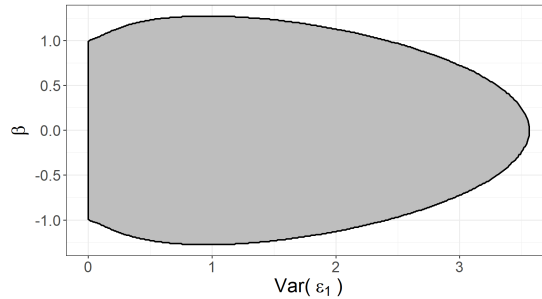


Figure 6.1: Stationary region. Simulated stationary region using Gaussian errors based on  $E \log j + \rho_{1j}$ .

by way of comparison, we use the maximally selected [CUSUM](#) process based on the WLS estimator studied in [Horváth and Trapani \(2022\)](#) (henceforth denoted as WLS).

We generate our data according to Eq (6.1), viz. using, under the null

$$y_i = (\alpha + \varepsilon_{i,1})y_{i-1} + \varepsilon_{i,2}; \quad 1 \leq i \leq n;$$

with  $n \in \{100; 200; 400; 800\}$ . We report results for  $\alpha \in \{0.5; 0.75; 1; 1.05\}$ ; we also experimented with different values of  $\alpha$  and found similar results. The errors  $\varepsilon_{i,1}$  and  $\varepsilon_{i,2}$  are generated as mean zero [IID](#) random variables simulated from centered Normal, Bernoulli, and exponential distributions with variance  $\frac{2}{1}$  and  $\frac{2}{2}$ , respectively - these scenarios are denoted as N, B, and E. We report results for  $\frac{2}{1} = \frac{2}{2} = 0.5$ ; however, in unreported simulations, different values for  $\frac{2}{1}$  and  $\frac{2}{2}$  resulted in similar results, except for the (expected) fact that the tests have better properties for smaller values of  $\frac{2}{2}$  and the empirical rejection frequencies are closer to their nominal values for larger values of  $\frac{2}{1}$ . Both effects vanish as  $n$  increases. Empirical rejection frequencies under the null hypothesis of no change are in [Table 6.1](#); simulations have been carried out with 500 iterations, and therefore, for a test at a nominal level of 0.05, the numbers in the table have a confidence band of  $[0.03; 0.07]$ .

[Table 6.1](#) clearly indicates that using the critical values coming from the finite sample correction in Eq (6.12), i.e. using `MLE_Vost`, systematically outperforms both MLE and WLS. In particular, empirical rejection frequencies using `MLE_Vost` are always either within or very close to the confidence interval even for  $n = 100$ , occasionally with a slight tendency to undersize (there are no instances of over-rejection). This can however be compared with the results obtained using asymptotic critical values (MLE), which indicate a massive undersize even for large sample sizes, likely to result in power reduction. Also, `MLE_Vost` seems to be remarkably stable across the various distributions employed,

Table 6.1: Size table. Empirical rejection frequencies under the null

| Error Type | 100   |       |       | 200   |       |       | 400   |       |        | 800   |       |       |
|------------|-------|-------|-------|-------|-------|-------|-------|-------|--------|-------|-------|-------|
|            | N     | B     | E     | N     | B     | E     | N     | B     | E      | N     | B     | E     |
| = 0.5      |       |       |       |       |       |       |       |       |        |       |       |       |
| MLE        | 0.002 | 0.002 | 0.006 | 0.012 | 0.006 | 0.010 | 0.004 | 0.006 | 0.028  | 0.010 | 0.004 | 0.016 |
| MLE_Vost   | 0.032 | 0.030 | 0.028 | 0.046 | 0.032 | 0.026 | 0.024 | 0.034 | 0.0342 | 0.032 | 0.028 | 0.052 |
| WLS        | 0.004 | 0.012 | 0.004 | 0.012 | 0.012 | 0.006 | 0.002 | 0.012 | 0.014  | 0.018 | 0.010 | 0.016 |
| = 0.75     |       |       |       |       |       |       |       |       |        |       |       |       |
| MLE        | 0.002 | 0.010 | 0.004 | 0.010 | 0.006 | 0.008 | 0.002 | 0.004 | 0.022  | 0.010 | 0.004 | 0.026 |
| MLE_Vost   | 0.030 | 0.044 | 0.026 | 0.028 | 0.024 | 0.024 | 0.028 | 0.032 | 0.048  | 0.036 | 0.026 | 0.052 |
| WLS        | 0.020 | 0.034 | 0.006 | 0.032 | 0.024 | 0.004 | 0.026 | 0.056 | 0.018  | 0.034 | 0.040 | 0.024 |
| = 1        |       |       |       |       |       |       |       |       |        |       |       |       |
| MLE        | 0.008 | 0.006 | 0.006 | 0.010 | 0.010 | 0.008 | 0.002 | 0.008 | 0.016  | 0.006 | 0.012 | 0.020 |
| MLE_Vost   | 0.022 | 0.048 | 0.030 | 0.040 | 0.032 | 0.036 | 0.032 | 0.036 | 0.054  | 0.036 | 0.038 | 0.040 |
| WLS        | 0.030 | 0.076 | 0.002 | 0.054 | 0.098 | 0.004 | 0.058 | 0.138 | 0.018  | 0.058 | 0.142 | 0.016 |
| = 1.05     |       |       |       |       |       |       |       |       |        |       |       |       |
| MLE        | 0.002 | 0.010 | 0.004 | 0.016 | 0.010 | 0.010 | 0.012 | 0.008 | 0.020  | 0.012 | 0.008 | 0.024 |
| MLE_Vost   | 0.024 | 0.046 | 0.028 | 0.052 | 0.038 | 0.036 | 0.032 | 0.038 | 0.058  | 0.044 | 0.036 | 0.048 |
| WLS        | 0.026 | 0.066 | 0.004 | 0.078 | 0.090 | 0.006 | 0.058 | 0.136 | 0.014  | 0.066 | 0.150 | 0.016 |

Proportion of rejections of  $H_0$  among the 500 independent simulations with nominal level = 0.05. DGP uses  $i_{;1} = i_{;2} = .5$ :

whereas WLS lacks such robustness and has a particularly pathological behaviour when considering Bernoulli innovations.

We also investigated the power of the proposed test statistics under the alternative. Power curves with nominal size = 0.05, as a function of the size of the change in the RCA coefficient, are shown in Figure 6.2, in which we only considered time series of length  $n = 400$  for brevity, although we saw similar results for other values of  $n$ . Using the errors previously defined, we consider both mid-sample and end-of-sample breaks defined as

$$y_i = ( \ + \ I(i \ 0.5n) + \ i_{;1})y_{i-1} + \ i_{;2} \quad 1 \ i \ n;$$

$$y_i = ( \ + \ I(i \ 0.9n) + \ i_{;1})y_{i-1} + \ i_{;2} \quad 1 \ i \ n;$$

We present the results where the parameter  $\beta$  is initially  $\beta = 2 \ f0;0.5g$  to show robustness over different values. We again see the MLE\_Vost method outperforms the MLE and WLS methods, which perform similarly.

## 6.4 Empirical applications

We illustrate our methodology through two applications to US house prices (section 4.1) and UK Covid hospitalisation data (section 4.2).

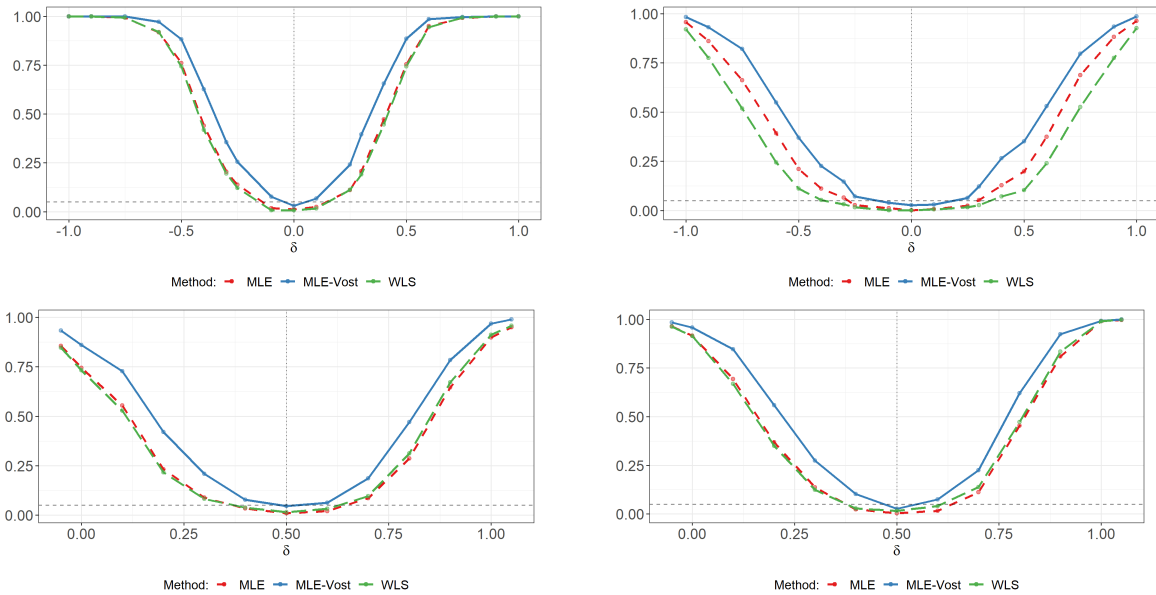


Figure 6.2: Power curves. Power curve as a function of the change in the RCA coefficient constructed from the percentage of rejections of  $H_0$  with nominal level  $\alpha = 0.05$ . (upper left panel)  $\delta = 0$ , change point is at  $0.5n$ , using Gaussian errors. (upper right panel)  $\delta = 0$ , change point is at  $0.9n$ , using Gaussian errors. (lower left panel)  $\delta = 0.5$ , change point is at  $0.5n$ , using Gaussian errors. (lower right panel)  $\delta = 0.5$ , change point is at  $0.5n$ , using Bernoulli errors.

In every exercise, we have used the LR based test as follows. For each of the series, we performed change point detection and estimation assuming they each follow an [RCA\(1\)](#) model. Since it is possible that more than one changepoint may exist, we have used [binary segmentation \(BS\)](#), with tests run at a nominal level  $\alpha = 0.05$  to identify the homogeneous regimes (we refer to the original paper by [Vostrikova, 1981](#), and also [Horváth and Trapani, 2022](#), for BS applied to the RCA model). In essence, when a break is detected, the sample is split around the estimated breakdate and our test is applied to each segment, and so on until no changes are detected. In order to rule out spurious detections, after conducting BS each segment containing a single estimated change point is re-tested, and if  $H_0$  is still rejected at level  $\alpha$  on this segment, then the change point is retained, otherwise it is discarded. Finally, on account of its superior performance in simulations, we only use the MLE\_Vost method to get critical values.

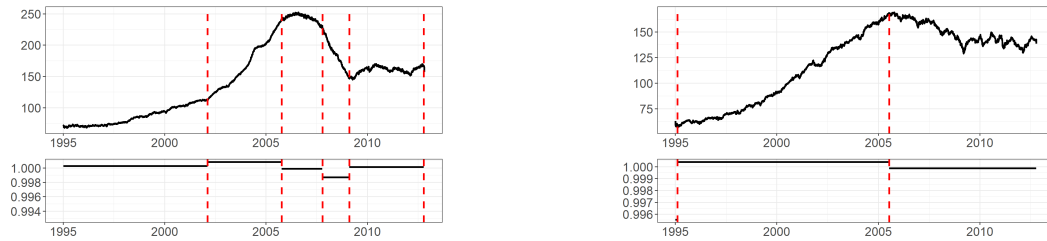


Figure 6.3: Housing changes. Los Angeles (left) and Boston (right) daily housing price indices (top plots) and the estimated beta values (bottom plots) with detected change points between January 4th, 1995, and 23rd October, 2012.

### 6.4.1 The US housing market

As mentioned in the introduction, using an [RCA\(1\)](#) model lends itself naturally to testing for the onset or collapse of bubbles, due to the possibility of carrying out inference irrespective of the stationarity of the observations or the lack thereof. Indeed, our approach can detect changes from a stationary or “nearly non stationary” regime to an explosive one (indicating the possible inception of a bubble), and also, with no modifications required, changes from explosive regimes to stationary, or less explosive, ones (thereby suggesting that a bubble has collapsed). In particular, testing for the presence of bubbles is particularly helpful in the context of the housing market, considering that housing is the most substantial investment for households; the presence of so-called “housing fever” can have pernicious consequences, as the 2006 US housing bubble demonstrated (see e.g. the recent survey by [Shi and Phillips, 2022a](#)).

Motivated by this, we investigate the presence of breaks in the house price indices of Boston and Los Angeles. Searching for breaks - and, more specifically, for bubble-type behaviour - in house prices is a “classical” exercise, and we refer to e.g. [Horváth et al. \(2021\)](#) who carry out a very similar analysis to ours, with monthly data. In our case, we use *daily* data, between January 4th, 1995, and 23rd October, 2012 (with  $n = 4645$ ); these data have been constructed by [Bollerslev et al. \(2016\)](#), to which we refer for a description of the datasets.<sup>2</sup>

Results are in [Figure 6.3](#), where we report the estimated break dates, and the values of on each approximately homogeneous segment.

<sup>2</sup>[Bollerslev et al. \(2016\)](#) construct daily prices for 10 major US cities; whilst we focus on Boston and L.A. only for the sake of a concise discussion, we have carried out the same analysis, obtaining broadly the same results, for the remaining 8 cities. Results are available upon request.

The break dates identified for the two series are highly suggestive. House prices in Los Angeles experienced five episodes of breaks. The first one is found in early 2002 (with an estimated date at 13 Feb 2002), with an increase in the autoregressive parameter which confirms the so-called “amplification mechanism” advocated by [Case and Shiller \(2003\)](#), and which indicates the presence and extent of the US housing bubble in the early 2000s. The second changepoint, with a decrease in the autoregressive coefficient, occurred in autumn 2005 (estimated date: 11 Oct 2005), which corresponds to the period of the US housing market correction, when prices began to stall. The third break, found at 12 Oct 2007, indicates an even sharper decline in the autoregressive coefficient, corresponding to a period of “hard landing”. This ended in early 2009 (estimated date: 10 Feb 2009), after which prices are found to have stabilised. We also found a break, corresponding to a sharp decline in the autoregressive coefficient, towards the end of the sample, on 10 Oct 2012, after a three-year high was recorded. The picture is very different for the housing market in Boston: we find evidence of a break at the very beginning of our sample, with the autoregressive coefficient sharply increasing (estimated date: 9 Feb 1995) and indicating evidence of the aforementioned “amplification mechanism”, followed by a correction starting on 12 July 2005.

#### 6.4.2 Covid-19 UK hospitalisation data

As we mentioned in the introduction, the [RCA\(1\)](#) model may be helpful when carrying out inference on the reduced form of an epidemiological model, where the co-existence of explosive (at the onset of the pandemic) and mean-reverting (when the pandemic subsides) behaviours is bound to be a feature of the data.

Motivated by this, and by the empirical application in [Magdalinos et al. \(2022\)](#), we apply the LR-based changepoint detection procedure to four time series of the daily number of Covid-19 hospitalisations in the four nations of the UK, viz. England, Northern Ireland, Scotland, and Wales - see [Figure 6.4](#) for a plot of the four time series. We use the number of hospitalisations, rather than infections, as the number of cases may be unreliable due to the changing number of tests administered;<sup>3</sup> the specific data we consider are available from the UK government website.<sup>4</sup>

The detected break points are visualized in the top panels of each series in [Figure 6.4](#), and the bottom panels show the estimated values of  $\alpha$  on each approximately homogeneous

---

<sup>3</sup>See e.g. <https://www.actuaries.org.uk/news-and-insights/news/coronavirus-statistics-uk-how-reliable-data>

<sup>4</sup><https://coronavirus.data.gov.uk/details/healthcare>

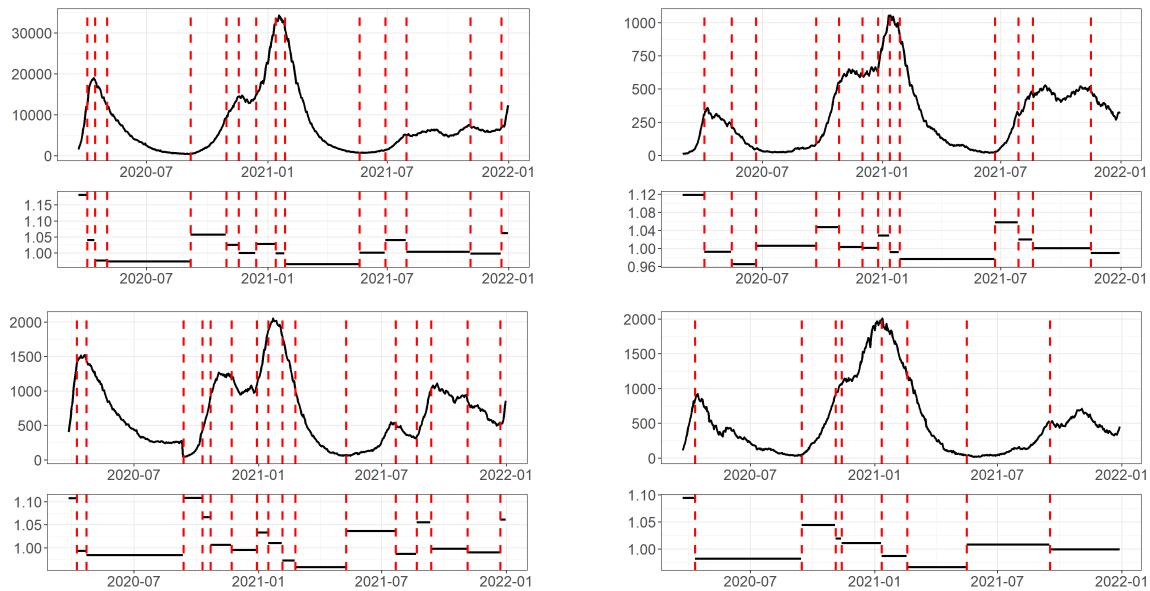


Figure 6.4: Covid-19 hospitalisation changes. Daily Covid-19 patients hospitalised for the four UK nations (top plot) and the estimated beta value (bottom plot). The England (upper left), Northern Ireland (upper right), Scotland (lower left), and Wales (lower right) data starts in March 2020 and ends on December 31, 2021.

segment. As expected, we note larger  $\beta$  values during times of explosive growth in hospitalisations, often exceeding 1, whereas lower values are estimated during more stationary periods.

While each nation experienced different lockdown and easing patterns, the estimated breakdates are highly suggestive. In particular, we detect a change point near the March 23, 2020 closures of education and hospitality sectors in all nations. Furthermore, change points are detected in late August of the same year, indicating the onset of the second wave, ascribed to increased travel that was formally acknowledged by the Prime Minister on September 18, 2020. Local and national lockdowns were implemented and eased throughout the holiday season. The series changed towards stationary again in January 2021, explained by national lockdown announced January 4, 2021 amid concerns of third wave due to the Christmas holidays. The series show another, smaller spike starting around July 2021 due to the emergence of the delta variant. The resulting lockdown limited the change seen in November 2021 with the emergence of the omicron variant. However, we do see a changepoint near the end of the series for England and Scotland (with the hospitalisation increasing in Northern Ireland and Wales at the last few entries in the data) in accordance

with holiday travels and increased prevalence of the omicron variant.

The local governments of the UK nations implemented slightly different policies, and therefore patterns are not the same across the whole UK. However, England, Northern Ireland and Scotland are essentially aligned, whereas Wales appears different; this is likely due to the fact that hospital admissions in Wales are counted in a different way than in the rest of the UK.<sup>5</sup> Finally, it is interesting to note that, especially for England, changepoints occur very closely to the start of each lockdown; factoring in some delay in detection, this seems to suggest that lockdowns tended to occur when a turning point in hospitalisations had already occurred, or was “in the making”. Such evidence can be read in conjunction with Wood (2021), who, albeit with a different methodology and focus, finds similar results for hospital deaths (see also Flaxman et al., 2020).

## 6.5 Concluding remarks

In this chapter, we extend the results in Horváth and Trapani (2022) on testing for changepoints in the deterministic part of a Random Coefficient AutoRegression. In particular, we study the maximally selected Likelihood Ratio test to check for the presence of a break. This is equivalent to the standardised CUSUM proposed in Horváth and Trapani (2022), and therefore, as in that case, we are able to detect breaks occurring very close to the beginning/end of the sample. Further, due to the fact that standard normal inference is always guaranteed in an RCA model, our test can be applied irrespective of whether the observations are stationary or not, also allowing to test for changes from e.g. an explosive to a non-explosive regime. Finally, the LR statistic has the added bonus of being self-normalising, and therefore no nuisance parameter appears (and needs to be estimated) in the asymptotic distribution. As our simulations show, the maximally selected LR statistic has very good power properties; further, our empirical illustrations show that our approach can be useful in a wide variety of datasets. We believe that this reinforces the claim in Aue and Horváth (2011) that “it would be worthwhile to give RCA models a closer look and that they are worthy of inclusion in the applied statistician’s toolbox”.

---

<sup>5</sup>Specifically, Wales reports also *suspected* Covid-19 cases, whereas all the other nations only report *confirmed* cases; see <https://www.cebm.net/covid-19/the-flaw-in-the-reporting-of-welsh-data-on-covid-hospital-admissions/>

# Chapter 7

## Conclusion

### 7.1 Discussions

In this thesis we have investigated several problems related to time series, functional data, and change point analysis. These problems have been approached from a theoretical and applied perspective.

In chapter 2 we proposed an empirical characteristic functional-based method for distributional change point detection in [functional time series \(FTS\)](#). We studied integrated- and supremum-type test statistics. Thresholds were considered for the statistics based on simulating Gaussian processes, permuting the data, and, in the case of the integrated test statistic, approximating the limiting behavior. Theoretical considerations were given. Empirically, the tests were well-sized and exhibited good power. The proposed methodology also exhibited strong performance when compared to existing functional change point detection methods. Practical applications of continuous electricity prices and high-frequency asset prices were explored.

Graph-based functional change point detection was discussed in chapter 3. Practical choices, and robustness to the choices, of several tuning parameters were explored. Comparative simulations for various change point methods were conducted. Graph-based methods were shown to exhibit relatively strong performance, particularly in the low sample scenarios. Graph-based methods were applied to multi-year pedestrian counts, high-dimensional stock returns, and continuous electricity prices.

Chapter 4 presented the R package `fChange`. The package provides a practical framework with numerous functions for functional change point analysis. Functionality of the

package such as visualizations, simulation, stationarity and white noise tests, and change point detection methods were discussed. The utility of the package was demonstrated on continuous electricity prices, cancer mortality, and long-term treasury rates.

Point patterns derived from cancer cell biopsy data were investigated in chapter 5. The data-rich images were summarized using cross-over  $\mathcal{K}$  functions and these functions were used to predict the outcome of interest in a functional random forest model. The model determined significant cell interactions via a variable importance plot and was robust to large-variable, low-sample settings. Extensive simulations showed that the model exhibited the ability to detect significant interactions and manage overfitting when there were no significant interactions. Strong predictive accuracy was also observed. Lung cancer and breast cancer data was explored. The package `funkycells` was presented.

Time series with stationary and nonstationary regimes were studied in chapter 6. Theoretical and empirical results for a maximum likelihood-based change point detection statistic were presented, along with a finite sample correction. Improved performance when compared to a weighted least squares approach was observed. The region which defines stationarity was simulated based the data parameters. The method was applied to UK hospitalisations and US housing market data.

## 7.2 Future work

Several avenues for future work are outlined in this section. We note, however, that the methodologies presented in this thesis are general and hence the extensions presented are not meant to be exhaustive.

Weighted characteristic functional-based change point detection: The characteristic functional-based change point detection method discussed in Chapter 2 was shown to be valuable in detecting distributional changes. Yet in many change point methods, adding a weight function can improve power. See discussions in section 1.4, [Horváth et al. \(2020\)](#), and [Horváth et al. \(2022\)](#). Adding weighting would be a valuable addition to the methodology.

Sequential characteristic functional-based change point detection: Chapter 2 showed that the characteristic functional-based change point detection method exhibited strong power when detecting distributional changes. Yet, the method was tuned to retrospective change point detection. Modifications of the method may improve its power in the sequential detection setting, e.g. see [Horváth et al. \(2021\)](#) and [Horváth and Rice \(2024\)](#).

Dependency in graph-based change point detection: Chapter 3 examines graph-based change point detection assuming independence. In many time series setting, this assumption does not hold. [Chen \(2019a\)](#) discuss using block-permutation to manage dependence for graph-based tests in the multivariate setting, but this has not yet been considered for [functional data \(FD\)](#). Investigations into how to define the blocks, and their size, would be a valuable contribution.

Sequential graph-based change point detection: Graph-based change point detection, as discussed in chapter 3, focused on retrospective analysis. However, many data are observed sequentially. Due to its power and computational efficiency, consideration of sequential change point detection with graph-based methods is of value. Further, many sequentially collected data are images, e.g. sensors on a car or security cameras. Applying graph-based detection to sequentially collected images would be of benefit in such fields.

Functional models: Chapter 4 presents an R package with many methods for functional change point detection. However, this package is limited in options for modeling and forecasting [FTS](#) beyond using the projection based approach by [Hyndman and Ullah \(2007\)](#). [FTS](#) models have been the subject of recent interest. [Aue et al. \(2015\)](#) consider functional autoregressive models, [Turbillon et al. \(2008\)](#) consider functional moving average models, and [Kuenzer \(2024\)](#) consider functional [autoregressive moving average \(ARMA\)](#) models; see also [Kühnert et al. \(2024\)](#). Nonlinear models have also been studied, e.g. see [Kühnert \(2020\)](#). Incorporation of these models into the package could improve analysis by providing a central location for change point detection and data modeling for practitioners.

Cell interaction relationships: The work in chapter 5 uses a functional random forest to describe how cell interactions are related to the outcome. However, the exact nature of the interaction-outcome relationships in a random forest model can be opaque. Incorporating partial dependence plots into the functional random forest model may improve this ([Friedman, 2001](#)). Yet, linear relationships may be better explained by a functional linear model. Developing such a model to fit the most important interactions may provide additional illumination into data.

Fully functional random forest: Chapter 5 considers a projection-based functional random forest model using [functional principal component analysis \(FPCA\)](#). While [total variance explained \(TVE\)](#) can influence how much information is retained by the projections, if the difference between outcome groups is orthogonal, or nearly orthogonal, to the projection direction, loss of power could be severe. Hence, a truly functional random forest may be of great benefit. Computing a separation metric between functions could be derived using some form of Monte Carlo integration, such as described in chapter 2.

Functional random survival forest: Chapter 5 considers response variables that

are encoded as categorical random variables, e.g. survival vs. death or type A vs. type B vs. type C. In many cases, considering the overall survival curve or some quality of life measure may be of more value to investigators. Random forest models which predict survival curves, i.e. random survival forests, in finite dimensions have seen interest in recent years; see [Ishwaran et al. \(2008\)](#), [Ishwaran and Kogalur \(2010\)](#), and [Ishwaran et al. \(2011\)](#). Developing a functional version would provide another valuable tool for investigators.

**$K$  function extensions:** The work in chapter 5 focused on simple crossover- $K$  functions. A clear extension is to consider inhomogeneous  $K$  functions to improve estimation when homogeneity cannot be assumed ([Baddeley et al., 2000](#)). Yet if clear tumors exist, cell location in relation to such tumors may be of more interest than described by traditional inhomogeneous  $K$  functions. Tumor-based  $K$  functions could be defined by considering the distance from a cell to the nearest edge of the tumor. Edges of the tumor could be defined manually or using a method such as  $\alpha$ -hulls ([Edelsbrunner et al., 1983](#); [Pateiro-López and Rodríguez-Casal, 2010](#)).

Sequential change point detection with maximum likelihood for an RCA model: Chapter 6 considers retrospective change point detection for [order one random coefficient additive \(RCA\(1\)\)](#) models. A recent preprint by [Horváth and Trapani \(2025\)](#) investigates sequential change point detection algorithm for [RCA\(1\)](#) models using a weighted least squares test statistic. As seen in the retrospective setting, a maximally selected likelihood statistic may exhibit improved power when compared to least squares in the sequential setting.

# References

- Abdi, H. and L. J. Williams (2010). Principal component analysis. *Wiley interdisciplinary reviews. Computational statistics* 2(4), 433–459.
- Akharif, A. and M. Hallin (2003). Efficient detection of random coefficients in autoregressive models. *The Annals of Statistics* 31(2), 675–704.
- Aminikhanghahi, S. and D. J. Cook (2017). A survey of methods for time series change point detection. *Knowledge and information systems* 51(2), 339–367.
- Anděl, J. (1976). Autoregressive series with random parameters. *Mathematische Operationsforschung und Statistik* 7(5), 735–741.
- Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59(3), 817–858.
- Andrews, D. W. K. and J. C. Monahan (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica* 60(4), 953–966.
- Angelo, M., S. C. Bendall, R. Finck, M. B. Hale, C. Hitzman, A. D. Borowsky, R. M. Levenson, J. B. Lowe, S. D. Liu, S. Zhao, et al. (2014). Multiplexed ion beam imaging of human breast tumors. *Nature medicine* 20(4), 436–442.
- Astill, S., D. I. Harvey, S. J. Leybourne, A. Taylor, and Y. Zu (2021, 05). CUSUM-based monitoring for explosive episodes in financial data in the presence of time-varying volatility. *Journal of Financial Econometrics* 21(1), 187–227. forthcoming.
- Aue, A. (2004). Strong approximation for RCA(1) time series with applications. *Statistics & Probability Letters* 68(4), 369–382.
- Aue, A., R. Gabrys, L. Horváth, and P. Kokoszka (2009). Estimation of a change-point in the mean function of functional data. *Journal of multivariate analysis* 100(10), 2254–2269.

- Aue, A., S. Hörmann, L. Horváth, and M. Hušková (2014). Dependent functional linear models with applications to monitoring structural change. *Statistica Sinica*, 1043–1073.
- Aue, A., S. Hörmann, L. Horváth, and M. Reimherr (2009). Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics* 37, 4046–4087.
- Aue, A. and L. Horváth (2011). Quasi-likelihood estimation in stationary and nonstationary autoregressive models with random coefficients. *Statistica Sinica*, 973–999.
- Aue, A. and L. Horváth (2013). Structural breaks in time series. *Journal of Time Series Analysis* 34, 1–16.
- Aue, A., L. Horváth, and J. Steinebach (2006). Estimation in random coefficient autoregressive models. *Journal of Time Series Analysis* 27(1), 61–76.
- Aue, A., D. D. Norinho, and S. Hörmann (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association* 110(509), 378–392.
- Aue, A., G. Rice, and O. Sönmez (2018). Detecting and dating structural breaks in functional data without dimension reduction. *Journal of the Royal Statistical Society. Series B, Statistical methodology* 80(3), 509–529.
- Aue, A., G. Rice, and O. Sönmez (2020). Structural break analysis for spectrum and trace of covariance operators. *Environmetrics* 31(1), e2617.
- Avanesov, V. and N. Buzum (2018). Change-point detection in high-dimensional covariance structure. *Electronic journal of statistics* 12(2).
- Baddeley, A., E. Rubak, and R. Turner (2015). *Spatial Point Patterns: Methodology and Applications with R*. London: Chapman and Hall/CRC Press.
- Baddeley, A. J., J. Møller, and R. Waagepetersen (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica* 54(3), 329–350.
- Barber, R. F. and E. J. Candés (2015). Controlling the false discovery rate via knockoffs. *The Annals of statistics* 43(5), 2055–2085.
- Barnard, G. A. (1959). Control charts and stochastic processes. *Journal of the Royal Statistical Society. Series B, Methodological* 21(2), 239–271.
- Bartlett, M. (1950). Periodogram analysis and continuous spectra. *Biometrika* 37, 1–16.

- Bartlett, M. S. (1964). The spectral analysis of two-dimensional point processes. *Biometrika* 51(3/4), 299–311.
- Barua, S., P. Fang, A. Sharma, J. Fujimoto, I. Wistuba, A. U. Rao, and S. H. Lin (2018). Spatial interaction of tumor cells and regulatory t cells correlates with survival in non-small cell lung cancer. *Lung Cancer* 117, 73–79.
- Bass, R. (1995). *Probabilistic Techniques in Analysis*. Springer.
- Bathia, N., Q. Yao, and F. Ziegelmann (2010). Identifying the finite dimensionality of curve time series. *The Annals of statistics* 38(6), 3352–3386.
- Benmoumen, M. and I. Salhi (2021). The strong consistency of quasi-maximum likelihood estimators for p-order random coefficient autoregressive (rca) models. *Sankhya A*, 1–16.
- Berkes, I., R. Gabrys, L. Horváth, and P. Kokoszka (2009). Detecting changes in the mean of functional observations. *Journal of the Royal Statistical Society. Series B, Statistical methodology* 71(5), 927–946.
- Berkes, I. and L. Horváth (2004). The efficiency of the estimators of the parameters in GARCH processes. *Annals of Statistics* 32(2), 633–655.
- Berkes, I., L. Horváth, and S. Ling (2009). Estimation in nonstationary random coefficient autoregressive models. *Journal of Time Series Analysis* 30(4), 395–416.
- Berkes, I., L. Horváth, and G. Rice (2013). Weak invariance principles for sums of dependent random functions. *Stochastic Processes and Their Applications* 123, 385–403.
- Besag, J. E. (1977). Discussion on dr ripley’s paper. *Journal of the Royal Statistical Society: Series B (Methodological)* 39(2), 193–195.
- Betancourt, J., F. Bachoc, T. Klein, D. Idier, J. Rohmer, and Y. Deville (2024). funGp: An R package for Gaussian process regression with scalar and functional inputs. *Journal of Statistical Software* 109(5), 1–51.
- Bezanson, J., A. Edelman, S. Karpinski, and V. B. Shah (2017). Julia: A fresh approach to numerical computing. *SIAM review* 59(1), 65–98.
- Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley.
- Black, S., D. Phillips, J. W. Hickey, J. Kennedy-Darling, V. G. Venkataraman, N. Samusik, Y. Goltsev, C. M. Schürch, and G. P. Nolan (2021). Codex multiplexed tissue imaging with dna-conjugated antibodies. *Nature protocols* 16(8), 3802–3835.

- Bleakley, K. and J.-P. Vert (2011). The group fused lasso for multiple change-point detection.
- Bogachev, V. (1998). *Gaussian Measures*. American Mathematical Society.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307–327.
- Bollerslev, T., A. J. Patton, and W. Wang (2016). Daily house price indices: Construction, modeling, and longer-run predictions. *Journal of Applied Econometrics* 31(6), 1005–1025.
- Boots, B. N. and A. Getis (1988). *Point pattern analysis*. Scientific geography series ; v. 8. Newbury Park, Calif: Sage Publications.
- Borkovec, M. and C. Klüppelberg (2001). The tail of the stationary distribution of an autoregressive process with ARCH(1) errors. *Annals of Applied Probability*, 1220–1241.
- Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*, Volume 149 of *Lecture Notes in Statistics*. New York, NY: Springer New York.
- Bottai, G., C. Raschioni, A. Losurdo, L. Di Tommaso, C. Tinterri, R. Torrisi, J. S. Reis-Filho, M. Roncalli, C. Sotiriou, A. Santoro, A. Mantovani, S. Loi, and L. Santarpia (2016). An immune stratification reveals a subset of pd-1/lag-3 double-positive triple-negative breast cancers. *Breast cancer research : BCR* 18(1), 121–121.
- Box, G. and G. Jenkins (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day series in time series analysis and digital processing. Holden-Day.
- Box, G. E. P., G. M. Jenkins, and G. C. Reinsel (2008). *Time series analysis : forecasting and control* (4th ed.). Wiley series in probability and statistics. Hoboken, NJ: John Wiley.
- Breiman, L. (1968). *Probability*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics.
- Breiman, L. (2001). Random forests. *Machine learning* 45(1), 5–32.
- Breiman, L., J. Friedman, C. J. Stone, and R. Olshen (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

- Brockwell, P. J. and R. A. Davis (1987). *Time Series : Theory and Methods* (First edition. ed.). Springer Series in Statistics Series. New York: Springer Science+Business Media New York.
- Brodsky, E. and B. Darkhovsky (2000). *Non-Parametric Statistical Diagnosis Problems and Methods* (1st 2000. ed.). Mathematics and Its Applications ; 509. Dordrecht: Springer Netherlands.
- Brown, R. G. (1963). *Smoothing, forecasting and prediction of discrete time series*. Prentice-Hall international series in management. Englewood Cliffs, N.J: Prentice-Hall.
- Bruckner, A., J. Bruckner, and B. Thomson (1997). *Real Analysis*. New Jersey: Prentice Hall.
- Bucchia, B. and M. Wendler (2017). Change-point detection and bootstrap for hilbert space valued random fields. *Journal of Multivariate Analysis* 155, 344–368.
- Bull, J. A., P. S. Macklin, T. Quaiser, F. Braun, S. L. Waters, C. W. Pugh, and H. M. Byrne (2020). Combining multiple spatial statistics enhances the description of immune cell localisation within tumours. *Scientific reports* 10(1), 18624.
- Canete, N. P., S. S. Iyengar, J. S. Wilmott, J. T. Ormerod, A. N. Harman, and E. Patrick (2021). spicyr: Spatial analysis of in situ cytometry data in r. *Health & Medicine Week n/a(n/a)*, 7776–.
- Canning, A. J., S. Viggiano, M. E. Fernandez-Zapico, and M. S. Cosgrove (2022). Parallel functional annotation of cancer-associated missense mutations in histone methyltransferases. *Scientific reports* 12(1), 18487–18487.
- Cardot, H., F. Ferraty, and P. Sarda (1999). Functional linear model. *Statistics & probability letters* 45(1), 11–22.
- Cardot, H., F. Ferraty, and P. Sarda (2003). Spline estimators for the functional linear model. *Statistica Sinica* 13(3), 571–591.
- Carlstein, E. (1986). The Use of Subseries Values for Estimating the Variance of a General Statistic from a Stationary Sequence. *The Annals of Statistics* 14(3), 1171 – 1179.
- Carstens, J. L., P. Correa de Sampaio, D. Yang, S. Barua, H. Wang, A. Rao, J. P. Allison, V. S. LeBleu, and R. Kalluri (2017). Spatial computation of intratumoral t cells correlates with survival of patients with pancreatic cancer. *Nature communications* 8(1), 15095.

- Case, K. E. and R. J. Shiller (2003). Is there a bubble in the housing market? *Brookings Papers on Economic Activity* 2003(2), 299–362.
- Cavaliere, G. and A. Rahbek (2021). A primer on bootstrap testing of hypotheses in time series models: With an application to double autoregressive models. *Econometric Theory* 37(1), 1–48.
- Characiejus, V. and G. Rice (2020). A general white noise test based on kernel lag-window estimates of the spectral density operator. *Econometrics and statistics* 13, 175–196.
- Chen, D., P. Hall, and H.-G. Müller (2011). Single and multiple index functional regression models with nonparametric link. *The Annals of statistics* 39(3), 1720–1747.
- Chen, H. (2019a). Change-point detection for multivariate and non-euclidean data with local dependency.
- Chen, H. (2019b). Sequential change-point detection based on nearest neighbors. *The Annals of statistics* 47(3), 1381–1407.
- Chen, H. and L. Chu (2023). Graph-based change-point analysis. *Annual Review of Statistics and Its Application* 10(1), 475–499.
- Chen, H. and J. H. Friedman (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association* 112(517), 397–409.
- Chen, H. and N. Zhang (2015). Graph-based change-point detection. *The Annals of statistics* 43(1), 139–176.
- Chen, H. and N. R. Zhang (2013). Graph-based tests for two-sample comparisons of categorical data. *Statistica Sinica* 23(4), 1479–1503.
- Chen, H., N. R. Zhang, L. Chu, and H. Song (2020). *gSeg: Graph-Based Change-Point Detection (g-Segmentation)*. R package version 1.0.
- Chen, H. Y., U. Palendira, and C. G. Feng (2022). Navigating the cellular landscape in tissue: Recent advances in defining the pathogenesis of human disease. *Computational and structural biotechnology journal* 20, 5256–5263.
- Chen, J. and A. K. Gupta (2012). *Parametric statistical change point analysis with applications to genetics, medicine, and finance* (2nd ed.). Boston: Birkhäuser.

- Chen, M., D. Li, and S. Ling (2014). Non-stationarity and quasi-maximum likelihood estimation on a double autoregressive model. *Journal of Time Series Analysis* 35(3), 189–202.
- Chen, Y. and C. S. Pun (2019). A bootstrap-based kpss test for functional time series. *Journal of multivariate analysis* 174, 104535–.
- Chenouri, S., A. Mozaffari, and G. Rice (2020). Robust multivariate change point analysis based on data depth. *Canadian journal of statistics* 48(3), 417–446.
- Chu, L. and H. Chen (2019). Asymptotic distribution-free change-point detection for multivariate and non-euclidean data. *The Annals of statistics* 47(1), 382–414.
- Chu, L. and H. Chen (2022). Sequential change-point detection for high-dimensional and non-euclidean data. *IEEE transactions on signal processing* 70, 4498–4511.
- Chui, C. K. (1992). *An introduction to wavelets*. Wavelet analysis and its applications ; v. 1. San Diego: Academic Press.
- Chung, K. L. and W. H. J. Fuchs (2008). On the distribution of values of sums of random variables. In *Selected Works Of Kai Lai Chung*, pp. 157–168. World Scientific.
- City of Melbourne (2024). Pedestrian counting system.
- Ciuperca, G. (2011). A general criterion to determine the number of change-points. *Statistics & probability letters* 81(8), 1267–1275.
- Ciuperca, G. (2012). Model selection by lasso methods in a change-point model. *Statistical papers (Berlin, Germany)* 55(2), 349–374.
- Cohn, D. (2015). *Measure Theory*. New York: Birkhäuser.
- Cressie, N. A. C. (1993). *Statistics for spatial data* (Rev. ed.). Wiley series in probability and mathematical statistics. Applied probability and statistics. New York: J. Wiley.
- Csörgő, M. and L. Horváth (1997). *Limit Theorems in Change-Point Analysis*. Wiley.
- Csörgő, M. and P. Révész (1981). *Strong approximations in probability and statistics*. Probability and mathematical statistics. New York: Academic Press.

- Damond, N., S. Engler, V. R. Zanutelli, D. Schapiro, C. H. Wasserfall, I. Kusmartseva, H. S. Nick, F. Thorel, P. L. Herrera, M. A. Atkinson, and B. Bodenmiller (2019). A map of human type 1 diabetes progression by imaging mass cytometry. *Cell Metabolism* 29(3), 755–768.e5.
- Darling, D. A. and P. Erdős (1956). A limit theorem for the maximum of normalized sum of independent random variables. *Duke Mathematical Journal* 23, 143–155.
- Dauxois, J., A. Pousse, and Y. Romain (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of multivariate analysis* 12(1), 136–154.
- Davis, R. A., D. Huang, and Y.-C. Yao (1995). Testing for a change in the parameter values and order of an autoregressive model. *The Annals of Statistics*, 282–304.
- Dawkins, B. (1990). Reply to comment on dawkins (1989) by w.f. kuhfeld. *American Statistician* 44, 58–60.
- De Boor, C. (2001). *A practical guide to splines* (Rev. ed.). Applied mathematical sciences ; v. 27. New York: Springer.
- Dehling, H., O. S. Sharipov, and M. Wendler (2015). Bootstrap for dependent Hilbert space-valued random variables with application to von mises statistics. *Journal of Multivariate Analysis* 133, 200–215.
- Dette, H. and J. Tang (2024). New energy distances for statistical inference on infinite dimensional hilbert spaces without moment conditions.
- Diba, B. T. and H. I. Grossman (1988). The theory of rational bubbles in stock prices. *The Economic Journal* 98(392), 746–754.
- Diggle, P. J. (1979). On parameter estimation and goodness-of-fit testing for spatial point patterns. *Biometrics* 35(1), 87–101.
- Diggle, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns*. CRC press.
- Edelsbrunner, H., D. Kirkpatrick, and R. Seidel (1983). On the shape of a set of points in the plane. *IEEE transactions on information theory* 29(4), 551–559.
- Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987–1007.

- Enikeeva, F. and Z. Harchaoui (2019). High-dimensional change-point detection under sparse alternatives. *The Annals of statistics* 47(4), 2051–2079.
- Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of probability and its applications* 14(1), 153–158.
- Fairbrother, J., L. Bardwell, K. Haynes, and D. Owens (2022). *Changepoints.jl*. Julia package version 0.4.1.
- Färkkilä, A., D. C. Gulhan, J. Casado, C. A. Jacobson, H. Nguyen, B. Kochupurakkal, Z. Maliga, C. Yapp, Y.-A. Chen, D. Schapiro, Y. Zhou, J. R. Graham, B. J. Dezube, P. Munster, S. Santagata, E. Garcia, S. Rodig, A. Lako, D. Chowdhury, G. I. Shapiro, U. A. Matulonis, P. J. Park, S. Hautaniemi, P. K. Sorger, E. M. Swisher, A. D. D’Andrea, and P. A. Konstantinopoulos (2020). Immunogenomic profiling determines responses to combined parp and pd-1 inhibition in ovarian cancer. *Nature communications* 11(1), 1459–13.
- Febrero-Bande, M. and M. Oviedo de la Fuente (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software* 51(4), 1–28.
- Ferraty, F. and P. Vieu (2006). *Nonparametric functional data analysis theory and practice*. Springer series in statistics. New York: Springer.
- Fisher, R. A. (1935). *The design of experiments* (1st ed ed.). Edinburgh, London, Oliver & Boyd: Oliver and Boyd.
- Flaxman, S., S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, C. Whitaker, H. Zhu, T. Berah, and J. W. Eaton (2020). Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature* 584(7820), 257–261.
- Francq, C. and J.-M. Zakoïan (2012). Strict stationarity testing and estimation of explosive and stationary generalized autoregressive conditional heteroscedasticity models. *Econometrica* 80(2), 821–861.
- Fremdt, S., J. G. Steinebach, L. Horváth, and P. Kokoszka (2013). Testing the equality of covariance operators in functional samples. *Scandinavian journal of statistics* 40(1), 138–152.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of statistics* 29(5), 1189–1232.

- Friedman, J. H. and L. C. Rafsky (1979). Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of statistics* 7(4), 697–717.
- Friedman, J. H. and B. W. Silverman (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics* 31(1), 3–21.
- Fryz, M. (2017). Conditional linear random process and random coefficient autoregressive model for EEG analysis. In *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, pp. 305–309. IEEE.
- Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of statistics* 42(6), 2243–2281.
- Gabrys, R., L. Horváth, and P. Kokoszka (2010). Tests for error correlation in the functional linear model. *Journal of the American Statistical Association* 105, 1113–1125.
- Gabrys, R. and P. Kokoszka (2007). Portmanteau test of independence for functional observations. *Journal of the American Statistical Association* 102(480), 1338–1348.
- Gaglia, G., M. L. Burger, C. C. Ritch, D. Rammos, Y. Dai, G. E. Crossland, S. Z. Tavana, S. Warchol, A. M. Jaeger, S. Naranjo, S. Coy, A. J. Nirmal, R. Krueger, J.-R. Lin, H. Pfister, P. K. Sorger, T. Jacks, and S. Santagata (2023). Lymphocyte networks are dynamic cellular communities in the immunoregulatory landscape of lung adenocarcinoma. *Cancer cell* 41(5), 871–886.e10.
- Galeano, P. and D. Peña (2007). Covariance changes detection in multivariate time series. *Journal of statistical planning and inference* 137(1), 194–211.
- Gardner, E. S. (2006). Exponential smoothing: The state of the art—part ii. *International journal of forecasting* 22(4), 637–666.
- Giesen, C., H. A. Wang, D. Schapiro, N. Zivanovic, A. Jacobs, B. Hattendorf, P. J. Schüffler, D. Grolimund, J. M. Buhmann, S. Brandt, et al. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods* 11(4), 417–422.
- Gnanadesikan, R. (1997). *Methods for statistical data analysis of multivariate observations / R. Gnanadesikan.* (2nd ed. ed.). Wiley series in probability and statistics. Applied probability and statistics. New York: Wiley.

- Goldsmith, J., F. Scheipl, L. Huang, J. Wrobel, C. Di, J. Gellar, J. Harezlak, M. W. McLean, B. Swihart, L. Xiao, C. Crainiceanu, P. T. Reiss, and E. Cui (2024). *refund: Regression with Functional Data*. R package version 0.1-35.
- Golovkine, S. (2021). Fdapy: a python package for functional data. *ArXiv abs/2101.11003*.
- Goltsev, Y., N. Samusik, J. Kennedy-Darling, S. Bhate, M. Hale, G. Vazquez, S. Black, and G. P. Nolan (2018). Deep profiling of mouse splenic architecture with codex multiplexed imaging. *Cell* 174(4), 968–981.
- Gombay, E., L. Horváth, and M. Husková (1996). Estimators and tests for change in variances. *Statistics & risk modeling* 14(2), 145–160.
- Gombay, E. and M. Hušková (1998). Rank based estimators of the change-point. *Journal of Statistical Planning and Inference* 67(1), 137–154.
- Gonzalez, J. P., A. Muñoz San Roque, and E. Alonso Perez (2018). Forecasting functional time series with a new Hilbertian ARMAX model: Application to electricity price forecasting. *IEEE transactions on power systems* 33(1), 545–556.
- Granger, C. W. and N. R. Swanson (1997). An introduction to stochastic unit-root processes. *Journal of Econometrics* 80(1), 35–62.
- Greenbaum, S., I. Averbukh, E. Soon, G. Rizzuto, A. Baranski, N. F. Greenwald, A. Kagel, M. Bosse, E. G. Jaswa, Z. Khair, S. Kwok, S. Warshawsky, H. Piyadasa, M. Goldston, A. Spence, G. Miller, M. Schwartz, W. Graf, D. Van Valen, V. D. Winn, T. Hollmann, L. Keren, M. van de Rijn, and M. Angelo (2023). A spatially resolved timeline of the human maternal-fetal interface. *Nature (London)* 619(7970), 595–3.
- Grenander, U. and M. Rosenblatt (1957). *Statistical analysis of stationary time series*. Wiley publication in mathematical statistics. New York: Wiley.
- Haase, P. (1995). Spatial pattern analysis in ecology based on ripley’s k-function: introduction and methods of edge correction. *Journal of vegetation science* 6(4), 575–582.
- Hall, P. (1979). On the rate of convergence of normal extremes. *Journal of Applied Probability* 16(2), 433–439.
- Hall, P. and C. C. Heyde (1980). *Martingale Limit Theory and its Applications*. New York: Academic Press.

- Hall, P. and I. Van Keilegom (2007). Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica* 17(4), 1511–1531.
- Hall, P. and C. Vial (2006). Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society. Series B, Statistical methodology* 68(4), 689–705.
- Hamilton, J. D. (1994). *Time series analysis*. Princeton, N.J: Princeton University Press.
- Happ-Kurz, C. (2020). Object-oriented software for functional data. *Journal of Statistical Software* 93(5), 1–38.
- Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B, Methodological* 55(4), 757–796.
- Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning data mining, inference, and prediction* (2nd ed.). Springer series in statistics. New York: Springer.
- Hein, L. (2016). FDA.jl. <https://github.com/LewisHein/FDA.jl>.
- Henze, N. and M. D. Jiménez-Gamero (2021). A test for Gaussianity in Hilbert spaces via the empirical characteristic functional. *Scandinavian Journal of Statistics* 48(2), 406–428.
- Herbel, C., V. Dittmer, M. Martínez-Osuna, L. N. Kuester, D. Schaefer, J. Drewes, J. Kollet, W. Mueller, M. Mallmann, P. Mallmann, et al. (2019). Evaluation of tumor-associated antigen expression with the macsimatm high-content imaging platform. *Cancer Research* 79(13\_Supplement), 4694–4694.
- Hill, J. and L. Peng (2014). Unified interval estimation for random coefficient autoregressive models. *Journal of Time Series Analysis* 35(3), 282–297.
- Hlávka, Z., M. Hušková, and S. G. Meintanis (2017). *Change Point Detection with Multivariate Observations Based on Characteristic Functions*, pp. 273–290. Cham: Springer International Publishing.
- Holt, C. C. (1957). Forecasting trends and seasonals by exponentially weighted moving averages. *ONR Memorandum* 52(52), 5–10.
- Homm, U. and J. Breitung (2012). Testing for speculative bubbles in stock markets: a comparison of alternative methods. *Journal of Financial Econometrics* 10(1), 198–231.

- Hörmann, S. and P. Kokoszka (2010). Weakly dependent functional data. *The Annals of Statistics* 38, 1845–1884.
- Hörmann, S. and P. Kokoszka (2012). Functional time series. In T. Subba Rao, S. Subba Rao, and C. Rao (Eds.), *Time Series Analysis: Methods and Applications*, Volume 30 of *Handbook of Statistics*, pp. 157–186. Elsevier.
- Horváth, L., M. Hušková, and G. Rice (2013). Test of independence for functional data. *Journal of multivariate analysis* 117, 100–119.
- Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications* (1. Aufl. ed.), Volume 200 of *Springer series in statistics*. New York, NY: Springer-Verlag.
- Horváth, L., P. Kokoszka, and G. Rice (2014). Testing stationarity of functional time series. *Journal of econometrics* 179(1), 66–82.
- Horváth, L., P. Kokoszka, and S. Wang (2021). Monitoring for a change point in a sequence of distributions. *The Annals of statistics* 49(4), 2271–2291.
- Horváth, L., Z. Liu, and S. Lu (2021). Sequential monitoring of changes in dynamic linear models, applied to the US housing market. *Econometric Theory*, 1–64.
- Horváth, L., C. Miller, and G. Rice (2020). A new class of change point test statistics of rényi type. *Journal of business & economic statistics* 38(3), 570–579.
- Horváth, L. and G. Rice (2014). Extensions of some classical methods in change point analysis (with dicussions). *TEST* 23.
- Horváth, L. and G. Rice (2024). *Change Point Analysis for Time Series* (1st ed. 2024. ed.). Springer Series in Statistics. Cham: Springer Nature Switzerland.
- Horváth, L., G. Rice, and J. VanderDoes (2025). Change point analysis for functional data using empirical characteristic functionals. *Journal Of Time Series Analysis*.
- Horváth, L., G. Rice, and S. Whipple (2016). Adaptive bandwidth selection in the estimation of the long run covariance of functional time series. *Computational Statistics and Data Analysis* 100, 676–693.
- Horváth, L., G. Rice, and Y. Zhao (2022). Change point analysis of covariance functions: A weighted cumulative sum approach. *Journal of multivariate analysis* 189, 104877–.

- Horváth, L. and L. Trapani (2016). Statistical inference in a random coefficient panel model. *Journal of Econometrics* 193(1), 54–75.
- Horváth, L. and L. Trapani (2019). Testing for randomness in a random coefficient autoregression model. *Journal of econometrics* 209(2), 338–352.
- Horváth, L. and L. Trapani (2022). Change point detection in heteroskedastic random coefficient autoregressive models. *Journal of Business and Economic Statistics*.
- Horváth, L. and L. Trapani (2025, March). Real-time monitoring with RCA models. *Econometric Theory*, 1–34.
- Horváth, L., L. Trapani, and J. VanderDoes (2024). The maximally selected likelihood ratio test in random coefficient models. *The Econometrics Journal* 27(3), 384–411.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* 24(6), 417–441.
- Hsing, T. and R. Eubank (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operations* (1st ed.). Wiley Series in Probability and Statistics. Chichester, England: Wiley.
- Hušková, M. and S. G. Meintanis (2006). Change point analysis based on empirical characteristic functions; empirical characteristic functions. *Metrika* 63(2), 145–168.
- Hušková, M. and S. G. Meintanis (2008). Tests for the multivariate k-sample problem based on the empirical characteristic function. *Journal of Nonparametric Statistics* 20(3), 263–277.
- Hwang, S. Y. and I. Basawa (2005). Explosive random-coefficient AR(1) processes and related asymptotics for least-squares estimation. *Journal of Time Series Analysis* 26(6), 807–824.
- Hyndman, R., A. Koehler, K. Ord, and R. Snyder (2008). *Forecasting with Exponential Smoothing: The State Space Approach* (1. Aufl. ed.). Springer Series in Statistics. Berlin, Heidelberg: Springer-Verlag.
- Hyndman, R. and H. L. Shang (2024). *ftsa: Functional Time Series Analysis*. R package version 6.4.
- Hyndman, R. and M. Ullah (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational statistics & data analysis* 51(10), 4942–4956.

- Hyndman, R. J. and Y. Khandakar (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software* 26(3), 1–22.
- Hyndman, R. J. and H. L. Shang (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of computational and graphical statistics* 19(1), 29–45.
- Hywood, J. D., M. N. Read, and G. Rice (2016). Statistical analysis of spatially homogeneous dynamic agent-based processes using functional time series analysis. *Spatial Statistics* 17, 199–219.
- Hywood, J. D., G. Rice, S. V. Pagoon, M. N. Read, and M. Biro (2021). Detection and characterization of chemotaxis without cell tracking. *Journal of the Royal Society Interface* 18(176), 20200879.
- Iacus, S. M. (2022). *sde: Simulation and Inference for Stochastic Differential Equations*. R package version 2.0.18.
- Illian, J., A. Penttinen, H. Stoyan, and D. Stoyan (2008). *Statistical analysis and modelling of spatial point patterns*. John Wiley & Sons.
- Ishwaran, H. and U. B. Kogalur (2010). Consistency of random survival forests. *Statistics & probability letters* 80(13), 1056–1064.
- Ishwaran, H., U. B. Kogalur, E. H. Blackstone, and M. S. Lauer (2008). Random survival forests. *The annals of applied statistics* 2(3), 841–860.
- Ishwaran, H., U. B. Kogalur, X. Chen, and A. J. Minn (2011). Random survival forests for high-dimensional data. *Statistical analysis and data mining* 4(1), 115–132.
- Jackson, H. W., J. R. Fischer, V. R. T. Zanutelli, H. R. Ali, R. Mechera, S. D. Soysal, H. Moch, S. Muenst, Z. Varga, W. P. Weber, and B. Bodenmiller (2020). The single-cell pathology landscape of breast cancer. *Nature (London)* 578(7796), 615–620.
- Jacques, J. and C. Preda (2014). Model-based clustering for multivariate functional data. *Computational statistics & data analysis* 71, 92–106.
- Jan, F., I. Shah, and S. Ali (2022). Short-term electricity prices forecasting using functional time series analysis. *Energies* 15(9).
- Jiang, F., R. Wang, and X. Shao (2023). Robust inference for change points in high dimension. *Journal of multivariate analysis* 193, 105114–.

- Jiang, S., C. N. Chan, X. Rovira-Clavé, H. Chen, Y. Bai, B. Zhu, E. McCaffrey, N. F. Greenwald, C. Liu, G. L. Barlow, J. L. Weirather, J. P. Oliveria, T. Nakayama, I. T. Lee, M. S. Matter, A. E. Carlisle, D. Philips, G. Vazquez, N. Mukherjee, K. Busman-Sahay, M. Nekorchuk, M. Terry, S. Younger, M. Bosse, J. Demeter, S. J. Rodig, A. Tzankov, Y. Goltsev, D. R. McIlwain, M. Angelo, J. D. Estes, and G. P. Nolan (2022). Combined protein and nucleic acid imaging reveals virus-dependent b cell and macrophage immunosuppression of tissue microenvironments. *Immunity (Cambridge, Mass.)* 55(6), 1118–1134.e8.
- Jolliffe, I. T. (1986). *Principal component analysis* (1st ed. 1986. ed.). Springer Series in Statistics. New York, New York: Springer.
- Jolliffe, I. T. and J. Cadima (2016). Principal component analysis: a review and recent developments. *Philosophical transactions of the Royal Society of London. Series A: Mathematical, physical, and engineering sciences* 374(2065), 20150202–20150202.
- Kao, C., L. Trapani, and G. Urga (2018). Testing for instability in covariance structures. *Bernoulli : official journal of the Bernoulli Society for Mathematical Statistics and Probability* 24(1), 740–771.
- Karhunen, K. (1946). Zur spektraltheorie stochastischer prozesse. *Ann. Acad. Sci. Fennicae, AI* 34.
- Keren, L., M. Bosse, D. Marquez, R. Angoshtari, S. Jain, S. Varma, S.-R. Yang, A. Kurian, D. Van Valen, R. West, S. C. Bendall, and M. Angelo (2018). A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell* 174(6), 1373–1387.e19.
- Keren, L., M. Bosse, S. Thompson, T. Risom, K. Vijayaragavan, E. McCaffrey, D. Marquez, R. Angoshtari, N. F. Greenwald, H. Fienberg, J. Wang, N. Kambham, D. Kirkwood, G. Nolan, T. J. Montine, S. J. Galli, R. West, S. C. Bendall, and M. Angelo (2019). Mibitof: A multiplexed imaging platform relates cellular phenotypes and tissue structure. *Science advances* 5(10), eaax5851–eaax5851.
- Kermack, W. O. and A. G. McKendrick (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character* 115(772), 700–721.
- Killick, R. and I. A. Eckley (2014). changepoint: An R package for changepoint analysis. *Journal of Statistical Software* 58(3), 1–19.

- Killick, R., P. Fearnhead, and I. A. Eckley (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association* 107(500), 1590–1598.
- Kim, M., P. Kokoszka, and G. Rice (2023). White noise testing for functional time series. *Statistics Surveys* 17(none), 119 – 168.
- Kim, M. and D. Petoukhov (2023). *wwntests: Hypothesis Tests for Functional Time Series*. R package version 1.1.0.
- Kim, M., C.-K. Yeh, and Y. Zhao (2024). *FTSgof: White Noise and Goodness-of-Fit Tests for Functional Time Series*. R package version 1.0.0.
- Kokoszka, P. and M. Reimherr (2017). *Introduction to Functional Data Analysis*. Chapman & Hall / CRC numerical analysis and scientific computing. CRC Press.
- Kokoszka, P., G. Rice, and H. L. Shang (2017). Inference for the autocovariance of a functional time series under conditional heteroscedasticity. *Journal of multivariate analysis* 162, 32–50.
- Kokoszka, P. and G. Young (2016). Kpss test for functional time series. *Statistics (Berlin, DDR)* 50(5), 957–973.
- Koner, S. and A.-M. Staicu (2023). Second-generation functional data. *Annual review of statistics and its application* 10(1), 547–572.
- Koul, H. L. and A. Schick (1996). Adaptive estimation in a random coefficient autoregressive model. *The Annals of Statistics* 24(3), 1025–1052.
- Kuenzer, T. (2024). Estimation of functional arma models. *Bernoulli : official journal of the Bernoulli Society for Mathematical Statistics and Probability* 30(1).
- Kuhfeld, W. F. (1990). Comment on dawkins (1989). *American Statistician* 44, 58–59.
- Kühnert, S. (2020). Functional arch and garch models: A yule-walker approach. *Electronic journal of statistics* 14(2).
- Kühnert, S., G. Rice, and A. Aue (2024). Estimating invertible processes in hilbert spaces, with applications to functional arma processes.
- Kwiatkowski, D., P. C. B. Phillips, P. Schmidt, and Y. Shin (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? *Journal of Econometrics* 54, 159–178.

- Lahiri, S. N. (1999). Theoretical comparisons of block bootstrap methods. *The Annals of Statistics* 27(1), 386 – 404.
- Lavielle, M. and G. Teyssi re (2006). Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal* 46, 287–306.
- Lee, S. (1998). Coefficient constancy test in a random coefficient autoregressive model. *Journal of Statistical Planning and Inference* 74(1), 93–101.
- Lee, S., J. Ha, O. Na, and S. Na (2003). The CUSUM test for parameter change in time series models. *Scandinavian Journal of Statistics* 30(4), 781–796.
- Lehmann, B. D., B. Jovanovi c, X. Chen, M. V. Estrada, K. N. Johnson, Y. Shyr, H. L. Moses, M. E. Sanders, and J. A. Pietenpol (2016). Refinement of triple-negative breast cancer molecular subtypes: Implications for neoadjuvant chemotherapy selection. *PLoS one* 11(6), e0157368–e0157368.
- Leng, X. and H.-G. M ller (2006). Classification using functional data analysis for temporal gene expression data. *Bioinformatics* 22(1), 68–76.
- Lewis, S. M., M.-L. Asselin-Labat, Q. Nguyen, J. Berthelet, X. Tan, V. C. Wimmer, D. Merino, K. L. Rogers, and S. H. Naik (2021). Spatial omics and multiplexed imaging to explore cancer biology. *Nature methods* 18(9), 997–1012.
- Leybourne, S. J., B. P. McCabe, and A. R. Tremayne (1996). Can economic time series be differenced to stationarity? *Journal of Business & Economic Statistics* 14(4), 435–446.
- Li, Q., Z. Ren, K. Cao, M. M. Li, K. Wang, and Y. Zhou (2022). Cancervar: An artificial intelligence-empowered platform for clinical interpretation of somatic mutations in cancer. *Science advances* 8(18), eabj1624–eabj1624.
- Lieberman, O. (2012). A similarity-based approach to time-varying coefficient non-stationary autoregression. *Journal of Time Series Analysis* 33(3), 484–502.
- Lieberman, O. and P. C. Phillips (2017). A multivariate stochastic unit root model with an application to derivative pricing. *Journal of Econometrics* 196(1), 99–110.
- Lieberman, O. and P. C. Phillips (2020). Hybrid stochastic local unit roots. *Journal of Econometrics* 215(1), 257–285.

- Lin, J.-R., M. Fallahi-Sichani, and P. K. Sorger (2015). Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. *Nature communications* 6(1), 8390.
- Lin, Y.-W. E., T. Shnitzer, R. Talmon, F. Villarroel-Espindola, S. Desai, K. Schalper, and Y. Kluger (2021). Graph of graphs analysis for multiplexed data with application to imaging mass cytometry. *PLoS Computational Biology* 17(3), e1008741.
- Ling, S. (2004). Estimation and testing stationarity for double-autoregressive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66(1), 63–78.
- Liu, Y.-W. and H. Chen (2022). A fast and efficient change-point detection framework based on approximate k-nearest neighbor graphs. *IEEE transactions on signal processing* 70, 1976–1986.
- Ljung, G. M. and G. E. P. Box (1978). On a measure of lack of fit in time series models. *Biometrika* 65(2), 297–303.
- Loève, M. (1946). Fonctions aléatoires à décomposition orthogonale exponentielle. *La Revue Scientifique* 84, 159–162.
- Magdalinos, T., K. Petrova, et al. (2022). Uniform and distribution-free inference with general autoregressive processes. Technical report.
- Mardia, K. V., J. T. J. T. Kent, and J. M. J. M. Bibby (1979). *Multivariate analysis*. Probability and mathematical statistics. London ;: Academic Press.
- Martínez-Álvarez, F., A. Schmutz, G. Asencio-Cortés, and J. Jacques (2019). A novel hybrid algorithm to forecast functional time series based on pattern sequence similarity with application to electricity demand. *Energies* 12(1).
- Matteson, D. S. and N. A. James (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association* 109, 334–345.
- Mercer, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character* 83(559), 69–70.

- Mestre, G., J. Portela, G. Rice, A. Muñoz San Roque, and E. Alonso (2021). Functional time series model identification and diagnosis by means of auto- and partial autocorrelation analysis. *Computational statistics & data analysis* 155, 107108–.
- NCI (2025, Feb). Milestones in cancer research and discovery.
- Newey, W. K. and K. D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55(3), 703–708.
- Neyman, J. (1939). On a new class of "contagious" distributions, applicable in entomology and bacteriology. *The Annals of mathematical statistics* 10(1), 35–57.
- Nicholls, D. F. and B. G. Quinn (2012). *Random Coefficient Autoregressive Models: An Introduction: An Introduction*, Volume 11. Springer Science & Business Media.
- Nie, L. and D. L. Nicolae (2021). Weighted-graph-based change point detection.
- Ohser, J. and D. Stoyan (1981). On the second-order and orientation analysis of planar stationary point processes. *Biometrical journal* 23(6), 523–533.
- Olshen, A. B., E. S. Venkatraman, R. Lucito, and M. Wigler (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics (Oxford, England)* 5(4), 557–572.
- Page, E. S. (1954). Continuous inspection schemes. *Biometrika* 41(1-2), 100–115.
- Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika* 42(3-4), 523–527.
- Page, E. S. (1957). On problems in which a change in a parameter occurs at an unknown point. *Biometrika* 44(1-2), 248–252.
- Pan, J. and J. Chen (2006). Application of modified information criterion to multiple change point problems. *Journal of multivariate analysis* 97(10), 2221–2241.
- Parra, E. R. (2021). Methods to determine and analyze the cellular spatial distribution extracted from multiplex immunofluorescence data to understand the tumor microenvironment. *Frontiers in Molecular Biosciences* 8, 668340.
- Parra, E. R., J. Zhai, A. Tamegnon, N. Zhou, R. K. Pandurengan, C. Barreto, M. Jiang, D. C. Rice, C. Creasy, A. A. Vaporciyan, et al. (2021). Identification of distinct immune landscapes using an automated nine-color multiplex immunofluorescence staining panel and image analysis in paraffin tumor tissues. *Scientific reports* 11(1), 1–14.

- Parzen, E. (1957). On consistent estimates of the spectrum of a stationary time series. *The Annals of mathematical statistics* 28(2), 329–348.
- Pateiro-López, B. and A. Rodríguez-Casal (2010). Generalizing the convex hull of a sample: The r package alphahull. *Journal of Statistical Software* 34(5), 1–28.
- Patwa, A., R. Yamashita, J. Long, T. Risom, M. Angelo, L. Keren, and D. L. Rubin (2021). Multiplexed imaging analysis of the tumor-immune microenvironment reveals predictors of outcome in triple-negative breast cancer. *Communications biology* 4(1), 852–852.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin philosophical magazine and journal of science* 2(11), 559–572.
- Peng, J. and D. Paul (2011). *fpca: Restricted MLE for Functional Principal Components Analysis*. R package version 0.2-1.
- Penttinen, A., D. Stoyan, and H. M. Henttonen (1992). Marked point processes in forest statistics. *Forest science* 38(4), 806–824.
- Phillips, P. C. and S. Shi (2019). Detecting financial collapse and ballooning sovereign risk. *Oxford Bulletin of Economics and Statistics* 81(6), 1336–1361.
- Phillips, P. C., S. Shi, and J. Yu (2015a). Testing for multiple bubbles: Historical episodes of exuberance and collapse in the S&P 500. *International Economic Review* 56(4), 1043–1078.
- Phillips, P. C., S. Shi, and J. Yu (2015b). Testing for multiple bubbles: Limit theory of real-time detectors. *International Economic Review* 56(4), 1079–1134.
- Phillips, P. C. and S.-P. Shi (2018). Financial bubble implosion and reverse regression. *Econometric Theory* 34(4), 705–753.
- Phillips, P. C., Y. Wu, and J. Yu (2011). Explosive behavior in the 1990s Nasdaq: When did exuberance escalate asset values? *International Economic Review* 52(1), 201–226.
- Picard, D. (1985). Testing and estimating change-points in time series. *Advances in applied probability* 17(4), 841–867.
- Pickett, K. L., K. Suresh, K. R. Campbell, S. Davis, and E. Juarez-Colunga (2021). Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. *BMC medical research methodology* 21(1), 1–216.

- Politis, D. N. (2003). Adaptive bandwidth choice. *Journal of nonparametric statistics* 15(4-5), 517–533.
- Politis, D. N. and J. P. Romano (1995). Bias-corrected nonparametric spectral estimation. *Journal of time series analysis* 16(1), 67–103.
- Quinn, B. (1982). A note on the existence of strictly stationary solutions to bilinear equations. *Journal of Time Series Analysis* 3(4), 249–252.
- R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Ramos-Carreño, C., J. L. Torrecilla, M. Carbajo Berrocal, P. Marcos Manchón, and A. Suárez (2024, May). scikit-fda: A Python Package for Functional Data Analysis. *Journal of Statistical Software* 109(2), 1–37.
- Ramsay, J. O. (1982). When the data are functions. *Psychometrika* 47(4), 379–396.
- Ramsay, J. O. and C. J. Dalzell (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B, Methodological* 53(3), 539–572.
- Ramsay, J. O., S. Graves, and G. Hooker (2022). *fda: Functional Data Analysis*. R package version 6.0.5.
- Ramsay, J. O. and B. W. Silverman (2002). *Applied Functional Data Analysis*. Springer, New York.
- Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis* (2nd ed. ed.). Springer series in statistics. New York: Springer.
- Ramsay, K. and S. Chenouri (2023). Robust nonparametric multiple changepoint detection for multivariate variability. *Econometrics and statistics*.
- Rao, A. R. and M. Reimherr (2021). Modern multiple imputation with functional data. *Stat (International Statistical Institute)* 10(1).
- Regis, M., P. Serra, and E. R. van den Heuvel (2022). Random autoregressive models: A structured overview. *Econometric Reviews* 41(2), 207–230.
- Rice, G. and H. L. Shang (2017). A plug-in bandwidth selection procedure for long-run covariance estimation with stationary functional time series. *Journal of time series analysis* 38(4), 591–609.

- Rice, G., T. Wirjanto, and Y. Zhao (2023). Exploring volatility of crude oil intraday return curves: A functional garch-x model. *Journal of Commodity Markets* 32, 100361.
- Rice, G. and C. Zhang (2022). Consistency of binary segmentation for multiple change-point estimation with functional data. *Statistics & Probability Letters* 180, 109228–.
- Rice, J. A. and B. W. Silverman (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B, Methodological* 53(1), 233–243.
- Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of applied probability* 13(2), 255–266.
- Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B, Methodological* 39(2), 172–212.
- Ripley, B. D. (1979). Tests of ‘randomness’ for spatial point patterns. *Journal of the Royal Statistical Society. Series B, Methodological* 41(3), 368–374.
- Ripley, B. D. (1988). *Statistical inference for spatial processes*. Cambridge: Cambridge University Press.
- Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society. Series B, Statistical methodology* 67(4), 515–530.
- Sadeghirad, H., N. Liu, J. Monkman, N. Ma, B. B. Cheikh, N. Jhaveri, C. W. Tan, M. E. Warkiani, M. N. Adams, Q. Nguyen, R. Ladwa, O. Braubach, K. O’Byrne, M. Davis, B. G. M. Hughes, and A. Kulasinghe (2023). Compartmentalized spatial profiling of the tumor microenvironment in head and neck squamous cell carcinoma identifies immune checkpoint molecules and tumor necrosis factor receptor superfamily members as biomarkers of response to immunotherapy. *Frontiers in immunology* 14, 1135489–1135489.
- Saka, S. K., Y. Wang, J. Y. Kishi, A. Zhu, Y. Zeng, W. Xie, K. Kirli, C. Yapp, M. Cicconet, B. J. Beliveau, et al. (2019). Immuno-saber enables highly multiplexed and amplified protein imaging in tissues. *Nature biotechnology* 37(9), 1080–1090.
- Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics* 2(6), 110–114.

- Scheipl, F., J. Goldsmith, and J. Wrobel (2024). *tidyfun: Tools for Tidy Functional Data*. R package version 0.0.98, <https://tidyfun.github.io/tidyfun/>.
- Schürch, C. M., S. S. Bhate, G. L. Barlow, D. J. Phillips, L. Noti, I. Zlobec, P. Chu, S. Black, J. Demeter, D. R. McIlwain, S. Kinoshita, N. Samusik, Y. Goltsev, and G. P. Nolan (2020). Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell* 183(3), 838–838.
- Scott, A. J. and M. Knott (1974). A cluster analysis method for grouping means in the analysis of variance. *Biometrics* 30(3), 507–512.
- Seal, S., T. Vu, T. Ghosh, J. Wrobel, and D. Ghosh (2022, 05). DenVar: density-based variation analysis of multiplex imaging data. *Bioinformatics Advances* 2(1), vbac039.
- Shang, H. L. (2017). Functional time series forecasting with dynamic updating: An application to intraday particulate matter concentration. *Econometrics and statistics* 1, 184–200.
- Shang, H. L. and R. Hyndman (2022). *rainbow: Bagplots, Boxplots and Rainbow Plots for Functional Data*. R package version 3.7.
- Sharipov, O., J. Tewes, and M. Wendler (2016). Sequential block bootstrap in a hilbert space with application to change point analysis. *Canadian Journal of Statistics* 44, 300–322.
- Sharipov, O. S. and M. Wendler (2020). Bootstrapping covariance operators of functional time series. *Journal of Nonparametric Statistics* 32(3), 648–666.
- Shi, S. and P. C. Phillips (2022a). Diagnosing housing fever with an econometric thermometer. *Journal of Economic Surveys*, 1–28.
- Shi, S. and P. C. Phillips (2022b). Econometric analysis of asset price bubbles.
- Shi, X., Y. Wu, and C. R. Rao (2017). Consistent and powerful graph-based change-point test for high-dimensional data. *Proceedings of the National Academy of Sciences - PNAS* 114(15), 3873–3878.
- Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* 68(1), 45–54.
- Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory of probability and its applications* 8(1), 22–46.

- Shumway, R. H. and D. S. Stoffer (2017). *Time Series Analysis and Its Applications: With R Examples* (4th ed.). Springer Texts in Statistics. Springer.
- Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of statistics* 24(1), 1–24.
- Silverman, B. W., J. C. Vassillious, and M. B. Priestley (2001). Wavelets: the key to intermittent information? *Journal of time series analysis* 22(1), 125–126.
- Ślęzak, J., K. Burnecki, and R. Metzler (2019). Random coefficient autoregressive processes describe brownian yet non-gaussian diffusion in heterogeneous systems. *New Journal of Physics* 21(7), 073056.
- Steland, A. and E. Rafałłowicz (2014). Decoupling change-point detection based on characteristic functions: Methodology, asymptotics, subsampling and application. *Journal of Statistical Planning and Inference* 145, 49–73.
- Stenseth, N. C., W. Falck, K.-S. Chan, O. N. Bjørnstad, M. O’Donoghue, H. Tong, R. Boonstra, S. Boutin, C. J. Krebs, and N. G. Yoccoz (1998). From patterns to processes: phase and density dependencies in the Canadian lynx cycle. *Proceedings of the National Academy of Sciences* 95(26), 15430–15435.
- Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC bioinformatics* 8(1), 25–25.
- Student (1907). On the error of counting with a haemocytometer. *Biometrika* 5(3), 351–360.
- Sugie, T., E. Sato, M. Miyashita, R. Yamaguchi, T. Sakatani, Y. Kozuka, S. Moritani, E. Suzuki, K. Kakimi, Y. Mikami, and T. Moriya (2020). Multispectral quantitative immunohistochemical analysis of tumor-infiltrating lymphocytes in relation to programmed death-ligand 1 expression in triple-negative breast cancer. *Breast cancer (Tokyo, Japan)* 27(4), 519–526.
- Sun, Y., P. C. B. Phillips, and S. Jin (2008). Optimal bandwidth selection in heteroskedasticity-autocorrelation robust testing. *Econometrica* 76(1), 175–194.
- Székely, G. J. and M. L. Rizzo (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference* 143(8), 1249–1272.

- The MathWorks Inc. (2024). *MATLAB version: 24.1.0 (R2024a)*. Natick, Massachusetts, United States.
- Therneau, T. and B. Atkinson (2022). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1.19.
- Thomas, M. (1949). A generalization of poisson’s binomial limit for use in ecology. *Biometrika* 36(1-2), 18–25.
- Toth, Z. E. and E. Mezey (2007). Simultaneous visualization of multiple antigens with tyramide signal amplification using antibodies from the same species. *Journal of Histochemistry & Cytochemistry* 55(6), 545–554.
- Truong, C., L. Oudre, and N. Vayatis (2018). ruptures: change point detection in python.
- Tsay, R. S. (1987). Conditional heteroscedastic time series models. *Journal of the American Statistical Association* 82(398), 590–604.
- Tucker, J. D. (2018). ElasticFDA.jl. <https://github.com/jdtuck/ElasticFDA.jl>.
- Tucker, J. D. (2024). *fdasrvf: Elastic Functional Data Analysis*. R package version 2.3.1.
- Turbillon, C., D. Bosq, J.-M. Marion, and B. Pumo (2008). Estimation du paramètre des moyennes mobiles hilbertiennes. *Comptes rendus. Mathématique* 346(5), 347–350.
- Ullah, S. and C. F. Finch (2013). Applications of functional data analysis: A systematic review. *BMC medical research methodology* 13(1), 43–43.
- Van Rossum, G. and F. L. Drake Jr (1995). *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.
- VanderDoes, J., C. Marceaux, K. Yokote, M.-L. Asselin-Labat, G. Rice, and J. D. Hywood (2024, 06). Using random forests to uncover the predictive power of distance-varying cell interactions in tumor microenvironments. *PLOS Computational Biology* 20(6), 1–28.
- Venkatraman, E. (1992). Consistency results in multiple change-point problems. *Ph.D. thesis, Stanford Univ., ProQuest LLC, Ann Arbor, MI. MR2687536*.
- Vostrikova, L. (1981). Detecting ‘disorder’ in multidimensional random processes. *Soviet Math. Dokl.* 24, 55–59.

- Vu, T., J. Wrobel, B. G. Bitler, E. L. Schenk, K. R. Jordan, and D. Ghosh (2022). Spf: a spatial and functional data analytic approach to cell imaging data. *PLoS Computational Biology* 18(6), e1009486.
- Wang, D., Y. Yu, and A. Rinaldo (2020). Univariate mean change point detection: Penalization, cusum and optimality. *Electronic journal of statistics* 14(1).
- Wang, D., Y. Yu, and A. Rinaldo (2021). Optimal covariance change point localization in high dimensions. *Bernoulli : official journal of the Bernoulli Society for Mathematical Statistics and Probability* 27(1).
- Wang, F., J. Flanagan, N. Su, L.-C. Wang, S. Bui, A. Nielson, X. Wu, H.-T. Vo, X.-J. Ma, and Y. Luo (2012). Rnascop: a novel in situ rna analysis platform for formalin-fixed, paraffin-embedded tissues. *The Journal of molecular diagnostics : JMD* 14(1), 22–29.
- Wang, G., C. Zou, and G. Yin (2018). Change-point detection in multinomial data with a large number of categories. *The Annals of statistics* 46(5), 2020–2044.
- Wang, J.-L., J.-M. Chiou, and H.-G. Müller (2016). Functional data analysis. *Annual review of statistics and its application* 3(1), 257–295.
- Wang, Y., Y. G. Wang, C. Hu, M. Li, Y. Fan, N. Otter, I. Sam, H. Gou, Y. Hu, T. Kwok, et al. (2021). Cell graph neural networks enable the digital staging of tumor microenvironment and precise prediction of patient survival in gastric cancer. *medRxiv*, 2021–09.
- Wegner, L. and M. Wendler (2024). Robust change-point detection for functional time series based on u-statistics and dependent wild bootstrap. *Statistical papers (Berlin, Germany)* 65(7), 4767–4810.
- Welch, B. L. (1947). The generalization of ‘student’s’ problem when several different population variances are involved. *Biometrika* 34(1/2), 28–35.
- White, H. and I. Domowitz (1984). Nonlinear regression with dependent observations. *Econometrica* 52(1), 152.
- Whittle, P. (1951). *Hypothesis Testing in Time Series Analysis*. Statistics / Uppsala universitet. Almqvist & Wiksells boktr.
- Whittle, P. (1954). *On Stationary Processes In The Plane*, Volume 41. Oxford University Press.

- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
- Wickham, H., M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani (2019). Welcome to the tidyverse. *Journal of Open Source Software* 4(43), 1686.
- Williams, C. G., H. J. Lee, T. Asatsuma, R. Vento-Tormo, and A. Haque (2022). An introduction to spatial transcriptomics for biomedical research. *Genome medicine* 14(1), 1–68.
- Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management science* 6(3), 324–342.
- Wold, S., K. Esbensen, and P. Geladi (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems* 2(1), 37–52.
- Wood, S. N. (2021). Inferring UK COVID-19 fatal infection trajectories from daily mortality data: were infections already in decline before the UK lockdowns? *Biometrics*.
- Wu, W. (2005). *Nonlinear System Theory: Another Look at Dependence*, Volume 102 of *Proceedings of The National Academy of Sciences of the United States*. National Academy of Sciences.
- Wu, Y. (2008). Simultaneous change point analysis and variable selection in a regression problem. *Journal of multivariate analysis* 99(9), 2154–2171.
- Yamada, I. and P. A. Rogerson (2003). An empirical comparison of edge effect correction methods applied to k-function analysis. *Geographical analysis* 35(2), 97–109.
- Yang, J. and P. Ren (2019). Bfda : A matlab toolbox for bayesian functional data analysis. *Journal of statistical software* 89(2), 1–21.
- Yao, F. (2007). Functional principal component analysis for longitudinal and survival data. *Statistica Sinica* 17(3), 965–983.
- Yao, F., H.-G. Müller, and W. Jane-Ling (2015). Pace package for functional data analysis and empirical dynamics (matlab).

- Yao, Y.-C. (1988). Estimating the number of change-points via schwartz's criterion. *Stat. Probab. Letters* 6, 173–177.
- Yi, M., T. Zhan, A. R. Peck, J. A. Hooke, A. J. Kovatich, C. D. Shriver, H. Hu, Y. Sun, H. Rui, and I. Chervoneva (2023). Quantile index biomarkers based on single-cell expression data. *Laboratory Investigation* 103(8), 100158.
- Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series?—a study in sampling and the nature of time-series. *Journal of the Royal Statistical Society* 89(1), 1–63.
- Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical transactions of the Royal Society of London. Series A, Containing papers of a mathematical or physical character* 226, 267–298.
- Zanotelli, V. R., M. Leutenegger, X.-K. Lun, F. Georgi, N. de Souza, and B. Bodenmiller (2020). A quantitative analysis of the interplay of environment, neighborhood, and cell state in 3d spheroids. *Molecular Systems Biology* 16(12), e9798.
- Zhang, J.-T. (2014). *Analysis of Variance for Functional Data*. CRC Press.
- Zhang, N. R. and D. O. Siegmund (2007). Modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* 63(1), 22–32.
- Zhang, N. R., D. O. Siegmund, H. Ji, and J. Z. Li (2010). Detecting simultaneous change-points in multiple sequences. *Biometrika* 97(3), 631–645.
- Zhang, X. (2016). White noise testing and model diagnostic checking for functional time series. *Journal of econometrics* 194(1), 76–95.
- Zhang, Y. and H. Chen (2021). Graph-based multiple change-point detection.
- Zhou, D. and H. Chen (2022). Ring-cpd: Asymptotic distribution-free change-point detection for multivariate and non-euclidean data.
- Zhou, Y., H. Chen, S. I. Iao, P. Kundu, H. Zhou, S. Bhattacharjee, C. Carroll, Y. Chen, X. Dai, J. Fan, A. Gajardo, P. Z. Hadjipantelis, K. Han, H. Ji, C. Zhu, H.-G. Müller, and J.-L. Wang (2024). *fdapace: Functional Data Analysis and Empirical Dynamics*. R package version 0.6.0.

- Zhu, X., J. Zhang, Y. Xu, J. Wang, X. Peng, and H.-D. Li (2020). Single-cell clustering based on shared nearest neighbor and graph partitioning. *Interdisciplinary sciences : computational life sciences* 12(2), 117–130.
- Zurrida, S. and U. Veronesi (2015). Milestones in breast cancer treatment. *The breast journal* 21(1), 3–12.

# APPENDICES

# Appendix A

## Chapter 2 Appendix

### A.1 Proofs

Throughout these proofs we let  $C_\gamma$  denote an unimportant positive constant that may change between uses.

Remark A.1.1 (Alternative method to compute  $T_n$  and  $M_n$ ). Let  $t \cdot = \cdot = D, \cdot \geq 1; \dots; Dg$ , and suppose that each  $X_j, j \geq 1$ , is Riemann-integrable. Then by the dominated convergence theorem, and the continuity of the complex exponential function, we have for any  $x \in [0; 1]$

$$\begin{aligned}
 \int_H jZ_n(v; x)^2 dQ(v) &= E \left[ jZ_n(W; x)^2 \middle| X_1; \dots; X_n \right] \\
 &= E \left[ \left| \frac{1}{n} \sum_{j=1}^{bnxc} \exp(ihX_j; Wi) - \frac{bnxc}{n^{3-2}} \sum_{j=1}^{bnxc} \exp(ihX_j; Wi) \right|^2 \middle| X_1; \dots; X_n \right] \\
 &= \lim_{D \rightarrow 1} E \left[ \left| \frac{1}{n} \sum_{j=1}^{bnxc} \exp(iR_{j;x;w}(D)) - \frac{bnxc}{n^{3-2}} \sum_{j=1}^{bnxc} \exp(iR_{j;x;w}(D)) \right|^2 \middle| X_1; \dots; X_n \right];
 \end{aligned} \tag{A.1}$$

where

$$R_{j;x;w}(D) = \sum_{\cdot=1}^D X_j(t)W(t)(t \cdot - t \cdot - 1);$$

The final line of Eq (A.1) may then be expressed as a limit of finite,  $D$ -dimensional integrals against the density of  $W(t_1); \dots; W(t_D)$ :

*Proof of Theorem 2.2.1.* We first begin by noting that for any probability measure  $Q$  on  $(L^2[0;1]; B(L^2[0;1]))$ ; where  $B(L^2[0;1])$  denotes the standard Borel  $\sigma$ -algebra on  $L^2[0;1]$ , the space  $L^2$  of measurable maps  $f : L^2[0;1] \rightarrow \mathbb{C}$  with

$$\int_H |f(v)|^2 dQ(v) < 1$$

may be made into a separable Hilbert space when equipped with the inner product

$$\langle f, g \rangle_L = \int_H f(v)g(v) dQ(v);$$

see Proposition 3.4.5 of [Cohn \(2015\)](#). Let  $k = k_L = \sqrt{h} / i_L$ . We write

$$\hat{r}_n(v; x) = \frac{1}{n} \sum_{i=1}^{bnxc} r_i(v); \quad v \in L^2[0;1]; \quad x \in [0;1];$$

where  $r_i(v) = \exp(ihX_i; vi)$ , so that

$$Z_n(v; x) = \frac{1}{n} \sum_{i=1}^{bnxc} r_i(v) = \frac{bnxc}{n} \frac{1}{bnxc} \sum_{i=1}^{bnxc} r_i(v);$$

Notice that  $r_i(v) = \cos(hv; X_i) + i \sin(hv; X_i)$ . Since the sine and cosine functions are bounded and Lipschitz, they are Hölder-continuous for any  $\alpha \in (0;1]$ , i.e. for all  $x, y \in \mathbb{R}$ ,  $|\sin(x) - \sin(y)| \leq C |x - y|^\alpha$ . Let  $2 < \alpha < 2 + \epsilon$ . We have with  $\alpha = p/\epsilon$ , where  $p$  is defined in Assumption 2.2.1, and by the Cauchy-Schwarz and Lyapounov inequalities, that

$$\begin{aligned} \int_H |\sin(hv; X_i) - \sin(hv; X_{i,m})|^2 dQ(v) &\leq C_{p=2}^2 \int_H |hv; X_i - hv; X_{i,m}|^{2p} dQ(v) \\ &\leq C_y k X_i - X_{i,m} k^{2p} \left[ \int_H |kv|^p dQ(v) \right]^{2/\epsilon}; \end{aligned}$$

with  $X_{i;m} = g(i_1, \dots, i_{m+1}; i_m, \dots, i_1)$  for  $f_i, g_{i2Z}$  being independent copies of  $\phi$ , independent of  $f_i, g_{i2Z}$ . As such

$$\left( E k \sin(h; X_{i;l}) \sin(h; X_{i;m}) k_L \right)^{1=(2+\epsilon)} = C_y (E k X_i X_{i;m} k^p)^{1=(2+\epsilon)};$$

and hence by Assumption 2.2.1

$$\sum_{m=0}^1 \left( E k \sin(h; X_{i;l}) \sin(h; X_{i;m}) k_L \right)^{1=(2+\epsilon)} < 1;$$

Using a similar argument for the cosine term we have that with  $\psi_{i;m}(v) = \exp(ihX_{i;m}v)$ ,

$$\sum_{m=0}^1 \left( E k \psi_{i;m} \psi_{i;m} k_L \right)^{1=(2+\epsilon)} < 1;$$

This shows that the sequence  $f_n \cdot g_{i2Z}$  satisfies the assumption of Theorem 1.1 of [Berkes et al. \(2013\)](#), which is valid for general elements of a separable Hilbert space. As such there exists a sequence of Gaussian processes  $\phi_n : L^2[0;1] \rightarrow \mathbb{C}$ , each with the common covariance structure as specified in the statement of the theorem, such that

$$\sup_{x \in [0;1]} \int_H |Z_n(v; x) - \phi_n(v; x)|^2 dQ(v) = o_P(1);$$

Moreover, as a result of Lemma 2.2 of [Berkes et al. \(2013\)](#), each of the covariance kernels in Eqs (2.11)–(2.13) are well defined elements of  $L^2(L^2[0;1] \times L^2[0;1])$ , and are trace class. The weak convergence of  $T_n$  and  $M_n$  follows straightforwardly from this.  $\square$

*Proof of Theorem 2.2.2.* We note that under the conditions of the theorem, we may write for  $0 < x < 1$

$$\rho_n \hat{f}_n(v; x) = \frac{1}{n} \sum_{i=1}^{bnxc} \psi_i(v);$$

and for  $x > 1$

$$\rho_n \hat{f}_n(v; x) = \frac{1}{n} \sum_{i=1}^{bn-c} \psi_i(v) + \frac{1}{n} \sum_{i=bn-c+1}^{bnxc} \psi_i(v);$$

As in the proof of Theorem 2.2.1, we may define sequences of mean zero Gaussian processes  $\overset{(1)}{w}_{n;C} : L^2[0;1] \rightarrow \mathbb{C}$  and  $\overset{(2)}{w}_{n;C} : L^2[0;1] \rightarrow \mathbb{C}$  with the following properties:  $\overset{(1)}{w}_{n;C}$  and  $\overset{(2)}{w}_{n;C}$  are independent,  $\overset{(1)}{w}_{n;C}$  are identically distributed for each  $n$ ,  $\overset{(2)}{w}_{n;C}$  are identically distributed for each  $n$ , the covariance kernels of  $\overset{(1)}{w}_{n;C}$  and  $\overset{(2)}{w}_{n;C}$  are well defined elements of  $L^2(L^2[0;1] \times L^2[0;1])$  that are trace class,

$$\sup_{x \in [0;1]} \int_H \left| \frac{1}{n} \sum_{i=1}^{bnxc} \overset{(1)}{w}_i(v) \cdot x f_{X_0}(v) \cdot \overset{(1)}{w}_{n;C}(v; x) \right|^2 dQ(v) = o_P(1);$$

and

$$\sup_{x \in [0;1]} \int_H \left| \frac{1}{n} \sum_{i=bn+1}^{bnxc} \overset{(1)}{w}_i(v) \cdot (1-x) f_{X_A}(v) \cdot \overset{(2)}{w}_{n;C}(v; x) \right|^2 dQ(v) = o_P(1);$$

As such, again with  $z(v; x) = x(1-x)(f_{X_0}(v) - f_{X_A}(v))1_{f < x} + (1-x)(f_{X_0}(v) - f_{X_A}(v))1_{f > x}$ ; and with

$$\overset{0}{w}_n(v; x) = \overset{(1)}{w}_{n;C}(v; x)1_{f < x} + [\overset{(1)}{w}_{n;C}(v; x) + \overset{(2)}{w}_{n;C}(v; x)]1_{f > x};$$

we have that

$$\sup_{x \in [0;1]} \int_H |Z_n(v; x) - \frac{1}{n} \int_H z(v; x) \cdot [\overset{0}{w}_n(v; x) - x \overset{0}{w}_n(v; 1)]|^2 dQ(v) = o_P(1);$$

Since

$$\sup_{x \in [0;1]} \int_H \int_H \overset{0}{w}_n(v; x) \cdot x \overset{0}{w}_n(v; 1) dQ(v) = O_P(1);$$

we obtain that

$$\frac{T_n}{n} \stackrel{P}{\rightarrow} \int_0^1 \int_H \int_H z(v; x) f^2 dQ(v); \quad (\text{A.2})$$

and

$$\frac{M_n}{n} \stackrel{P}{\rightarrow} \sup_{x \in [0;1]} \int_H \int_H z(v; x) f^2 dQ(v); \quad (\text{A.3})$$

Under Eq (2.15), the right hand side of both Eqs (A.2) and (A.3) are strictly positive, which implies that  $T_n \stackrel{P}{\rightarrow} 1$  and  $M_n \stackrel{P}{\rightarrow} 1$ , as needed. The proof of the second half of the theorem that  $\hat{k}_n - k = O_P(1)$  follows from Theorem 8.4.5 of Horváth and Rice (2024).  $\square$

*Proof of Corollary 2.2.1.* Suppose  $Q$  is the probability measure of standard Brownian motion on  $(L^2[0;1]; B(L^2[0;1]))$ . When  $H_A$  of Eq (2.2) holds, there must exist a  $v_0$  so that  $f_{X_0}(v_0) \notin f_{X_A}(v_0)$ . Due to the continuity of  $v \mapsto \exp(\mathbf{i}hX; vi)$ ; there exists an open ball in  $L^2[0;1]$

$$B_{v_0}(\cdot) = \{f \in L^2[0;1] : \|f - v_0\| < \delta\}$$

and an  $\delta > 0$  so that for all  $g \in B_{v_0}(\cdot)$ ,  $|f_{X_0}(g) - f_{X_A}(g)| > \epsilon$ . Let  $C[0;1]$  denote the space of continuous functions on  $[0;1]$ . By the support theorem for Brownian motion, see e.g. pg. 59 of Bass (1995), for any  $h \in C[0;1]$  such that  $h(0) = 0$ , and  $\delta > 0$ , the  $C[0;1]$ -ball about  $h$  of radius  $\delta$ ,

$$B_h^1(\cdot) = \left\{ g \in C[0;1] : \sup_{x \in [0;1]} |g(x) - h(x)| < \delta \right\};$$

satisfies  $Q(B_h^1(\cdot)) > 0$ : Since continuous functions taking value zero at the origin are dense in  $L^2[0;1]$ , there exists a  $C[0;1]$ -ball  $B_{h_0}^1(\cdot)$  so that  $B_{h_0}^1(\cdot) \subset B_{v_0}(\cdot)$ . As such,

$$\int_H |f_{X_0}(v) - f_{X_A}(v)|^2 dQ(v) \Big|_{Q(B_{h_0}^1(\cdot))} > 0:$$

Therefore in this case Eq (2.15) is always satisfied. □

*Proof of Theorem 2.3.1.* We again write

$$\begin{aligned} Z_n(v; X) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{bnxc} \mathbf{i} \cdot (v) \quad \frac{bnxc}{n} \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{i} \cdot (v) \\ &:= \frac{1}{\sqrt{n}} \sum_{i=1}^{bnxc} (\mathbf{i} \cdot (v) - \mathbf{i} \cdot (v)) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{bnxc} (\cos(hv; X \cdot i) - \overline{\text{Re}}_n(v)) + \frac{\mathbf{i}}{\sqrt{n}} \sum_{i=1}^{bnxc} (\sin(hv; X \cdot i) - \overline{\text{Im}}_n(v)) \\ &=: Z_{n;Re}(v; X) + Z_{n;Im}(v; X): \end{aligned}$$

We write  $Z_n(v; X)$ ,  $Z_{n;Re}(v; X)$ ; and  $Z_{n;Im}(v; X)$  to denote the same processes based on the randomly permuted sample, i.e. with  $X$  replaced by  $X_{(\cdot)}$ . We note that, without conditioning on  $\mathbf{X}$ ,  $X_{(\cdot)}$  is distributed for each  $i$  as  $Y$  in the statement of the Theorem. For any fixed  $g \in L^2$ , and conditionally on  $\mathbf{X}$ ,

$$S_{k;n} = \frac{1}{\sqrt{n}} \sum_{i=1}^k \cos(h; X_{(\cdot)} i) \overline{\text{Re}}_n(\cdot); g \Big|_{L^2}$$

is a martingale array with respect to the filtration  $F_k = (\mathcal{F}_1; \dots; \mathcal{F}_k)$ ; see pg 53 of [Hall and Heyde \(1980\)](#). The variables  $Y_{i;n} = h \cos(h; X_{(\cdot)} i) \overline{\text{Re}}_n(\cdot); g i_{L^2}$  satisfy the conditions (3.12)–(3.14) of Lemma 3.1 in on pg. 57 of [Hall and Heyde \(1980\)](#), from which it follows as on pg. 168 of [Hušková and Meintanis \(2006\)](#) that

$$hZ_{n;\text{Re}}(\cdot; X_{(\cdot)}); g i_{L^2} \stackrel{D^{[0,1]}}{\rightarrow} g Br(x);$$

where  $Br$  is a standard Brownian bridge, and

$$\frac{2}{g} = \left\langle \int_H C_{\text{Re}}^A(v; \cdot) g(v) dQ(v); g \right\rangle_{L^2} :$$

Moreover, for any  $g_1; g_2 \in L^2$  we have that,

$$\begin{pmatrix} hZ_{n;\text{Re}}(\cdot; X_{(\cdot)}); g_1 i_{L^2} \\ hZ_{n;\text{Im}}(\cdot; X_{(\cdot)}); g_2 i_{L^2} \end{pmatrix} \stackrel{D^{[0,1]}}{\rightarrow} \begin{matrix} 1=2 \\ g_1; g_2 \end{matrix} \begin{pmatrix} Br_1(x) \\ Br_2(x) \end{pmatrix} ;$$

where  $Br_i, i = 1; 2$ , are independent standard Brownian bridges, and

$$= \left( \begin{matrix} \left\langle \int_H C_{\text{Re}}^A(v; \cdot) g(v) dQ(v); g \right\rangle_{L^2} & \left\langle \int_H C_{\text{ReIm}}^A(v; \cdot) g_1(v) dQ(v); g_2 \right\rangle_{L^2} \\ \left\langle \int_H C_{\text{ReIm}}^A(\cdot; v) g_2(v) dQ(v); g_1 \right\rangle_{L^2} & \left\langle \int_H C_{\text{Im}}^A(v; \cdot) g_2(v) dQ(v); g_2 \right\rangle_{L^2} \end{matrix} \right) \in \mathbb{R}^{2 \times 2};$$

This implies that the process  $Z_n(v; X_{(\cdot)})$  satisfies the conclusion of Lemma 2.5 of [Berkes et al. \(2013\)](#). The result now follows from Lemma 2.8 of [Berkes et al. \(2013\)](#), which also holds for general elements of a separable Hilbert space.  $\square$

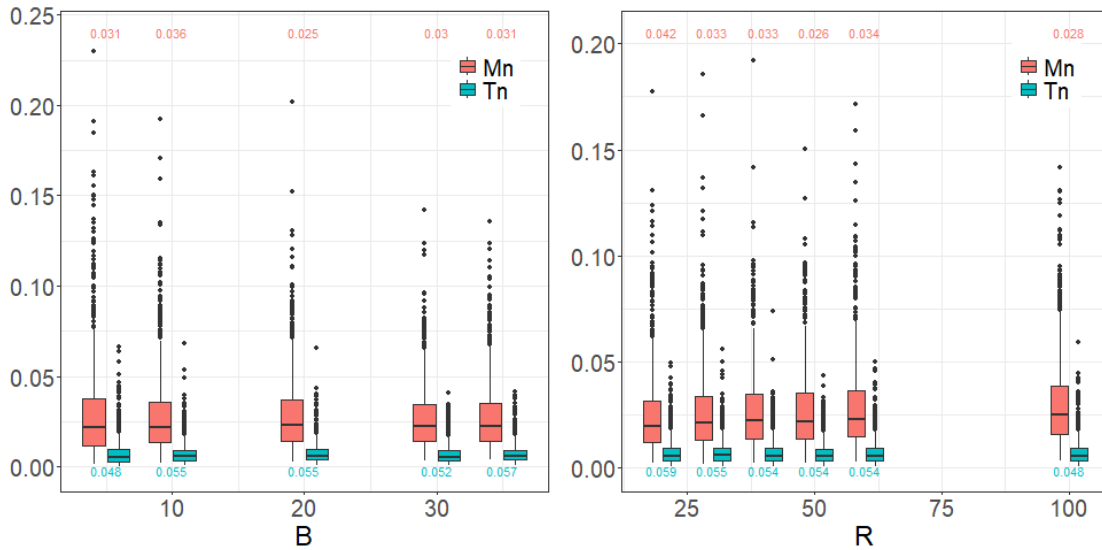
## A.2 Selection of $B$ and $R$

We present results in this section investigating the choice of  $B$  and  $R$ , while providing rationale behind the default choices,  $B = 20$  and  $R = 50$ , given in the chapter. On length  $n = 100$  data generated via ARKL DGP described in section 3.3, several scenarios based on a different number of  $B$  or  $R$  is run for 1,000 iterations. The aggregated values and log values are given in Figure A.1. When  $B$  or  $R$  are being investigated, we respectively set  $R = 50$  and  $B = 20$ . Figure A.1 shows the effect of  $B$  and  $R$  on the test statistics  $T_n$  and  $M_n$ .

Figures A.1(a) and A.1(c) show that increasing  $B$  beyond even a very small amount has marginal impact on the value of the test statistics. Thus, choosing a sufficient number of Brownian motions to properly approximate the integral seems to be governed more by the computational time. In this chapter, we assume  $B = 20$ ; however, an even smaller choice may also be reasonable.

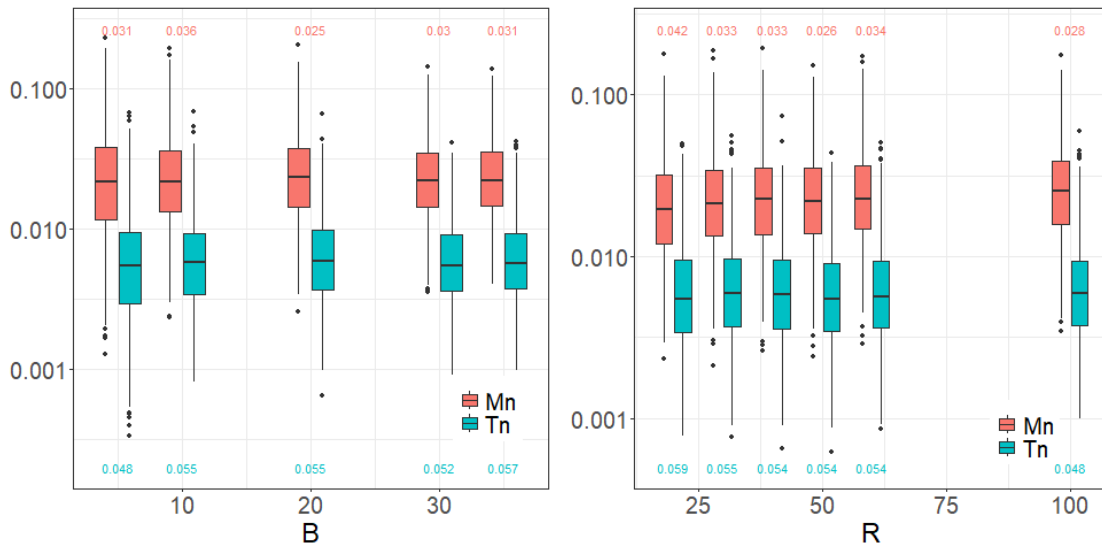
Figures A.1(b) and A.1(d) show that increasing  $R$  also does not seem to substantially increase the value of the test statistics. Close inspection may indicate a marginal increase however, the empirical sizes show that the test statistics remain properly sized indicating any potential value increases may be due to randomness or are at the very least offset by increases in the statistic. Hence even low dimensional approximations of the functions are possible to use with the characteristic functional-based approaches. We use  $R = 50$  throughout the chapter.

The value of  $B$  is also of interest in terms of size and power. Figure A.2 shows the power curve when  $R = 50$ ,  $B \in \{1; 5; 10; 20\}$ , and  $n = 100$ . As expected, there is some power loss when  $B = 1$ , but most values above this do not exhibit significant differences and indeed  $B = 20$  is sufficient for our needs.



(a) Test statistics based on  $B$ .

(b) Test statistics based on  $R$ .



(c) Log test statistics based on  $B$ .

(d) Log test statistics based on  $R$ .

Figure A.1: Parameter choices. Investigations into the selection and effects of  $B$  and  $R$ . Box-plots, labeled with the respective sizes, are shown for the test statistics  $T_n$  and  $M_n$ —Figures (a) and (b)—and their logged values—Figures (c) and (d). All data is length 100 generated from ARKL DGP given in section 3.3. Each setting is run for 1,000 iterations to generate the given box-plots. When  $B$  or  $R$  are being investigated, we respectively set  $R = 50$  and  $B = 20$ .

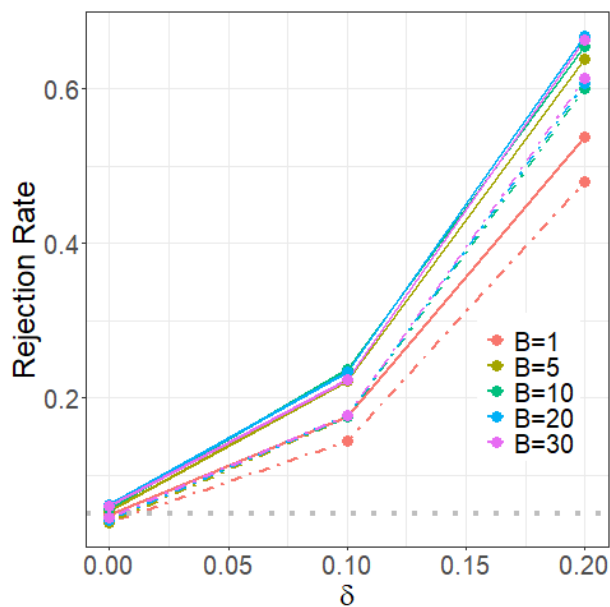


Figure A.2:  $B$  power curves. Power curves based on mid-sample mean changes for varying choices of  $B$ . All data is length 100 generated from ARKL DGP given in section 3.3.  $T_n$  simulation (solid) and  $M_n$  simulation (dot-dash) are compared for moderate changes. 1;000 simulations are computed at each point. Beyond  $B = 1$ , there seems to be negligible power gain for a given test statistic, with  $T_n$  performing better than  $M_n$ . Each setting displayed approximately nominal empirical size.

### A.3 Null tables

This section contains null tables for of a ARKL, AROU and FAR(1) time series. In all tables, conclusions akin to those given in section 3.3 are seen.

Tables A.1 and A.2 investigate the size of a ARKL DGP for  $\alpha = 0.01$ .  $T_n$  seems to have a better size than  $M_n$ . As before, Gaussian simulation is more robust to dependence than permutation or Welch approximations. However, each can be reasonably tuned with sufficiently large bandwidth, perhaps selected via a data-driven approach; see Rice and Shang (2017) and Horváth et al. (2016).

Tables A.3 and A.4 investigate the size of a AROU DGP for  $\alpha = 0.05$ .  $T_n$  seems to have a better size than  $M_n$ . All approaches perform similarly under dependence. Each can be reasonably tuned with sufficiently large bandwidth, perhaps selected via a data-driven approach; see Rice and Shang (2017) and Horváth et al. (2016).

Tables A.5 and A.6 investigate the size of a FAR(1) DGP for  $\alpha = 0.05$ .  $T_n$  seems to have a better size than  $M_n$ . All approaches perform similarly under dependence. Each can be reasonably tuned with sufficiently large bandwidth, perhaps selected via a data-driven approach; see Rice and Shang (2017) and Horváth et al. (2016).

Table A.1:  $T_n$  size table at **0:01** significance. Empirical size for  $T_n$  test statistic using simulation, permutation, and approximation based on 1,000 simulations from a ARKL DGP with Gaussian errors. Data lengths of 100, 250, and 500 are considered under varying degrees of dependence. Several bandwidths used to compute the long-run variance estimators and permutation block sizes, both denoted  $h$ , are also considered. The nominal level is set at 0.01.

| $T_n$ TEST STATISTIC |               |       |       |                     |       |       |                   |       |       |
|----------------------|---------------|-------|-------|---------------------|-------|-------|-------------------|-------|-------|
| $n =$                | No Dependence |       |       | Moderate Dependence |       |       | Strong Dependence |       |       |
|                      | 100           | 250   | 500   | 100                 | 250   | 500   | 100               | 250   | 500   |
| Gaussian Simulation  |               |       |       |                     |       |       |                   |       |       |
| $h = 1$              | 0.004         | 0.007 | 0.008 | 0.038               | 0.039 | 0.040 | 0.223             | 0.262 | 0.244 |
| $h = n^{1=3}$        | 0.005         | 0.005 | 0.008 | 0.012               | 0.010 | 0.007 | 0.050             | 0.039 | 0.031 |
| $h = 2n^{1=3}$       | 0.000         | 0.006 | 0.004 | 0.000               | 0.008 | 0.006 | 0.009             | 0.016 | 0.017 |
| Permutation          |               |       |       |                     |       |       |                   |       |       |
| $h = 1$              | 0.007         | 0.007 | 0.010 | 0.112               | 0.117 | 0.137 | 0.451             | 0.515 | 0.524 |
| $h = n^{1=3}$        | 0.011         | 0.006 | 0.011 | 0.027               | 0.023 | 0.020 | 0.173             | 0.142 | 0.100 |
| $h = 2n^{1=3}$       | 0.009         | 0.010 | 0.008 | 0.002               | 0.015 | 0.015 | 0.066             | 0.063 | 0.040 |
| Welch Approximation  |               |       |       |                     |       |       |                   |       |       |
| $h = 1$              | 0.018         | 0.020 | 0.020 | 0.151               | 0.156 | 0.161 | 0.493             | 0.548 | 0.566 |
| $h = n^{1=3}$        | 0.012         | 0.022 | 0.018 | 0.040               | 0.039 | 0.036 | 0.208             | 0.194 | 0.137 |
| $h = 2n^{1=3}$       | 0.008         | 0.011 | 0.018 | 0.015               | 0.020 | 0.021 | 0.080             | 0.077 | 0.053 |

Table A.2:  $M_n$  size table at **0.01** significance. Empirical size for  $M_n$  test statistic using simulation and permutation based on 1,000 simulations from a ARKL DGP with Gaussian errors. Data lengths of 100, 250, and 500 are considered under varying degrees of dependence. Several bandwidths used to compute the long-run variance estimators and permutation block sizes, both denoted  $h$ , are also considered. The nominal level is set at 0.01.

| $M_n$ TEST STATISTIC |               |       |       |                     |       |       |                   |       |       |
|----------------------|---------------|-------|-------|---------------------|-------|-------|-------------------|-------|-------|
|                      | No Dependence |       |       | Moderate Dependence |       |       | Strong Dependence |       |       |
|                      | 100           | 250   | 500   | 100                 | 250   | 500   | 100               | 250   | 500   |
| Gaussian Simulation  |               |       |       |                     |       |       |                   |       |       |
| IID                  | 0.003         | 0.002 | 0.005 | 0.035               | 0.027 | 0.037 | 0.235             | 0.272 | 0.277 |
| $n^{1=3}$            | 0.001         | 0.000 | 0.004 | 0.001               | 0.001 | 0.008 | 0.025             | 0.022 | 0.025 |
| $2n^{1=3}$           | 0.000         | 0.001 | 0.003 | 0.000               | 0.001 | 0.004 | 0.000             | 0.004 | 0.009 |
| Permutation          |               |       |       |                     |       |       |                   |       |       |
| IID                  | 0.005         | 0.008 | 0.012 | 0.146               | 0.161 | 0.169 | 0.475             | 0.553 | 0.582 |
| $n^{1=3}$            | 0.008         | 0.009 | 0.012 | 0.031               | 0.026 | 0.024 | 0.208             | 0.166 | 0.144 |
| $2n^{1=3}$           | 0.008         | 0.009 | 0.008 | 0.014               | 0.014 | 0.013 | 0.078             | 0.059 | 0.042 |

Table A.3:  $T_n$  AROU size table. Empirical size for  $T_n$  test statistic using simulation, permutation, and approximation based on 1,000 simulations from a AROU DGP. Data lengths of 100, 250, and 500 are considered under varying degrees of dependence. Several bandwidths used to compute the long-run variance estimators and permutation block sizes, both denoted  $h$ , are also considered. The nominal level is set at 0.05.

| $T_n$ TEST STATISTIC |               |       |       |                     |       |       |                   |       |       |
|----------------------|---------------|-------|-------|---------------------|-------|-------|-------------------|-------|-------|
|                      | No Dependence |       |       | Moderate Dependence |       |       | Strong Dependence |       |       |
|                      | 100           | 250   | 500   | 100                 | 250   | 500   | 100               | 250   | 500   |
| Gaussian Simulation  |               |       |       |                     |       |       |                   |       |       |
| IID                  | 0.059         | 0.050 | 0.052 | 0.390               | 0.399 | 0.444 | 0.939             | 0.399 | 0.957 |
| $n^{1=3}$            | 0.057         | 0.051 | 0.059 | 0.117               | 0.097 | 0.093 | 0.343             | 0.097 | 0.240 |
| $2n^{1=3}$           | 0.054         | 0.053 | 0.058 | 0.080               | 0.070 | 0.078 | 0.175             | 0.070 | 0.120 |
| Permutation          |               |       |       |                     |       |       |                   |       |       |
| IID                  | 0.054         | 0.046 | 0.056 | 0.384               | 0.387 | 0.440 | 0.929             | 0.936 | 0.950 |
| $n^{1=3}$            | 0.058         | 0.052 | 0.061 | 0.114               | 0.091 | 0.084 | 0.331             | 0.248 | 0.227 |
| $2n^{1=3}$           | 0.053         | 0.045 | 0.061 | 0.076               | 0.072 | 0.074 | 0.172             | 0.125 | 0.122 |
| Welch Approximation  |               |       |       |                     |       |       |                   |       |       |
| IID                  | 0.058         | 0.055 | 0.061 | 0.399               | 0.393 | 0.446 | 0.933             | 0.945 | 0.959 |
| $n^{1=3}$            | 0.054         | 0.056 | 0.062 | 0.118               | 0.091 | 0.093 | 0.355             | 0.258 | 0.240 |
| $2n^{1=3}$           | 0.052         | 0.054 | 0.069 | 0.079               | 0.062 | 0.077 | 0.179             | 0.143 | 0.133 |

Table A.4:  $M_n$  AROU size table. Empirical size for  $M_n$  test statistic using simulation and permutation based on 1,000 simulations from a AROU DGP. Data lengths of 100, 250, and 500 are considered under varying degrees of dependence. Several bandwidths used to compute the long-run variance estimators and permutation block sizes, both denoted  $h$ , are also considered. The nominal level is set at 0.05.

| $M_n$ TEST STATISTIC |               |       |       |                     |       |       |                   |       |       |
|----------------------|---------------|-------|-------|---------------------|-------|-------|-------------------|-------|-------|
|                      | No Dependence |       |       | Moderate Dependence |       |       | Strong Dependence |       |       |
|                      | 100           | 250   | 500   | 100                 | 250   | 500   | 100               | 250   | 500   |
| Gaussian Simulation  |               |       |       |                     |       |       |                   |       |       |
| IID                  | 0.040         | 0.034 | 0.052 | 0.444               | 0.465 | 0.492 | 0.943             | 0.962 | 0.970 |
| $n^{1=3}$            | 0.034         | 0.034 | 0.042 | 0.103               | 0.080 | 0.081 | 0.378             | 0.274 | 0.246 |
| $2n^{1=3}$           | 0.036         | 0.033 | 0.043 | 0.061               | 0.053 | 0.060 | 0.170             | 0.132 | 0.119 |
| Permutation          |               |       |       |                     |       |       |                   |       |       |
| IID                  | 0.050         | 0.045 | 0.056 | 0.471               | 0.486 | 0.523 | 0.949             | 0.967 | 0.972 |
| $n^{1=3}$            | 0.052         | 0.044 | 0.063 | 0.143               | 0.100 | 0.101 | 0.392             | 0.298 | 0.273 |
| $2n^{1=3}$           | 0.048         | 0.044 | 0.071 | 0.081               | 0.072 | 0.080 | 0.203             | 0.156 | 0.148 |

Table A.5:  $T_n$  FAR(1) size table. Empirical size for  $T_n$  test statistic using simulation, permutation, and approximation based on 1,000 simulations from a FAR(1) DGP. Data lengths of 100, 250, and 500 are considered under varying degrees of dependence. Several bandwidths used to compute the long-run variance estimators and permutation block sizes, both denoted  $h$ , are also considered. The nominal level is set at 0.05.

| $T_n$ TEST STATISTIC |               |       |       |                     |       |       |                   |       |       |
|----------------------|---------------|-------|-------|---------------------|-------|-------|-------------------|-------|-------|
|                      | No Dependence |       |       | Moderate Dependence |       |       | Strong Dependence |       |       |
|                      | 100           | 250   | 500   | 100                 | 250   | 500   | 100               | 250   | 500   |
| Gaussian Simulation  |               |       |       |                     |       |       |                   |       |       |
| IID                  | 0.048         | 0.045 | 0.062 | 0.351               | 0.322 | 0.328 | 0.817             | 0.852 | 0.883 |
| $n^{1=3}$            | 0.044         | 0.046 | 0.057 | 0.114               | 0.073 | 0.080 | 0.278             | 0.232 | 0.196 |
| $2n^{1=3}$           | 0.044         | 0.039 | 0.048 | 0.065               | 0.053 | 0.062 | 0.145             | 0.130 | 0.112 |
| Permutation          |               |       |       |                     |       |       |                   |       |       |
| IID                  | 0.047         | 0.044 | 0.051 | 0.336               | 0.314 | 0.328 | 0.808             | 0.845 | 0.869 |
| $n^{1=3}$            | 0.042         | 0.041 | 0.057 | 0.100               | 0.067 | 0.077 | 0.268             | 0.228 | 0.199 |
| $2n^{1=3}$           | 0.048         | 0.046 | 0.053 | 0.063               | 0.057 | 0.068 | 0.149             | 0.120 | 0.108 |
| Welch Approximation  |               |       |       |                     |       |       |                   |       |       |
| IID                  | 0.061         | 0.058 | 0.060 | 0.350               | 0.345 | 0.338 | 0.825             | 0.856 | 0.888 |
| $n^{1=3}$            | 0.051         | 0.050 | 0.063 | 0.111               | 0.077 | 0.091 | 0.269             | 0.228 | 0.198 |
| $2n^{1=3}$           | 0.048         | 0.048 | 0.059 | 0.065               | 0.062 | 0.072 | 0.154             | 0.126 | 0.114 |

Table A.6:  $M_n$  FAR(1) size table. Empirical size for  $M_n$  test statistic using simulation and permutation based on 1,000 simulations from a FAR(1) DGP. Data lengths of 100, 250, and 500 are considered under varying degrees of dependence. Several bandwidths used to compute the long-run variance estimators and permutation block sizes, both denoted  $h$ , are also considered. The nominal level is set at 0.05.

| $M_n$ TEST STATISTIC |               |       |       |                     |       |       |                   |       |       |
|----------------------|---------------|-------|-------|---------------------|-------|-------|-------------------|-------|-------|
|                      | No Dependence |       |       | Moderate Dependence |       |       | Strong Dependence |       |       |
|                      | 100           | 250   | 500   | 100                 | 250   | 500   | 100               | 250   | 500   |
| Gaussian Simulation  |               |       |       |                     |       |       |                   |       |       |
| IID                  | 0.035         | 0.036 | 0.369 | 0.356               | 0.357 | 0.369 | 0.821             | 0.880 | 0.907 |
| $n^{1=3}$            | 0.030         | 0.028 | 0.063 | 0.066               | 0.054 | 0.063 | 0.258             | 0.218 | 0.189 |
| $2n^{1=3}$           | 0.027         | 0.023 | 0.038 | 0.031               | 0.024 | 0.038 | 0.088             | 0.097 | 0.096 |
| Permutation          |               |       |       |                     |       |       |                   |       |       |
| IID                  | 0.052         | 0.044 | 0.056 | 0.381               | 0.392 | 0.414 | 0.831             | 0.889 | 0.922 |
| $n^{1=3}$            | 0.048         | 0.044 | 0.056 | 0.103               | 0.090 | 0.090 | 0.306             | 0.266 | 0.227 |
| $2n^{1=3}$           | 0.047         | 0.058 | 0.049 | 0.079               | 0.063 | 0.066 | 0.161             | 0.133 | 0.138 |

# Appendix B

## Chapter 4 Appendix

### B.1 Data summary table

| Patient (Images) | Cancer | B    | CD4T | CD8T | Dendritic | Granulocytes | Macrophages | Mast | NK | Treg | ydT |
|------------------|--------|------|------|------|-----------|--------------|-------------|------|----|------|-----|
| 1 (2)            | LUAD   | 152  | 148  | 93   | 199       | 123          | 178         | 137  | 17 | 48   | 7   |
| 2 (1)            | LUAD   | 391  | 248  | 286  | 57        | 3            | 144         | 410  | 17 | 32   | 4   |
| 3 (2)            | LUAD   | 1752 | 699  | 981  | 75        | 68           | 278         | 14   | 51 | 69   | 46  |
| 4 (3)            | LUSC   | 18   | 89   | 48   | 30        | 13           | 37          | 61   | 3  | 66   | 12  |
| 5 (1)            | LUSC   | 6    | 30   | 32   | 5         | 19           | 6           | 22   | 2  | 19   | 2   |
| 6 (1)            | LUAD   | 290  | 330  | 206  | 22        | 442          | 99          | 12   | 8  | 35   | 11  |
| 7 (1)            | LUAD   | 669  | 509  | 791  | 122       | 27           | 129         | 139  | 9  | 25   | 18  |
| 8 (1)            | LUAD   | 14   | 33   | 62   | 6         | 9            | 27          | 13   | 2  | 2    | 4   |
| 9 (1)            | LUAD   | 9    | 47   | 83   | 3         | 2            | 10          | 16   | 1  | 3    | 1   |
| 10 (1)           | LUAD   | 35   | 188  | 143  | 48        | 13           | 117         | 11   | 66 | 32   | 48  |
| 11 (3)           | LUSC   | 340  | 394  | 115  | 77        | 134          | 186         | 141  | 19 | 57   | 21  |
| 12 (1)           | LUSC   | 27   | 78   | 89   | 0         | 6            | 31          | 78   | 3  | 0    | 2   |
| 13 (2)           | LUAD   | 16   | 52   | 96   | 11        | 27           | 91          | 43   | 26 | 64   | 11  |
| 14 (1)           | LUSC   | 623  | 271  | 224  | 247       | 490          | 164         | 154  | 16 | 64   | 30  |
| 15 (4)           | LUAD   | 320  | 215  | 232  | 2         | 49           | 89          | 22   | 13 | 2    | 34  |
| 16 (4)           | LUAD   | 303  | 376  | 211  | 22        | 206          | 100         | 5    | 30 | 75   | 45  |
| 17 (1)           | LUAD   | 443  | 106  | 45   | 16        | 33           | 8           | 12   | 0  | 8    | 0   |
| 18 (2)           | LUAD   | 28   | 166  | 211  | 79        | 39           | 61          | 488  | 7  | 8    | 12  |
| 19 (1)           | LUAD   | 30   | 57   | 49   | 12        | 9            | 43          | 38   | 40 | 15   | 2   |
| 20 (1)           | LUSC   | 24   | 14   | 30   | 11        | 122          | 72          | 13   | 0  | 1    | 0   |
| 21 (1)           | LUAD   | 474  | 11   | 24   | 36        | 13           | 4           | 0    | 0  | 1    | 4   |
| 22 (2)           | LUAD   | 533  | 186  | 160  | 87        | 79           | 69          | 3    | 17 | 40   | 44  |
| 23 (1)           | LUAD   | 488  | 282  | 124  | 81        | 6            | 172         | 7    | 9  | 25   | 13  |
| 24 (1)           | LUAD   | 16   | 172  | 54   | 4         | 32           | 170         | 45   | 5  | 36   | 4   |
| 25 (3)           | LUAD   | 216  | 253  | 169  | 78        | 1404         | 97          | 17   | 23 | 4    | 74  |
| 26 (5)           | LUSC   | 7    | 251  | 190  | 112       | 748          | 207         | 68   | 66 | 113  | 20  |
| 27 (3)           | LUSC   | 15   | 271  | 429  | 9         | 382          | 63          | 10   | 13 | 10   | 103 |
| 28 (1)           | LUAD   | 12   | 72   | 35   | 1         | 0            | 9           | 3    | 0  | 5    | 3   |
| 29 (1)           | LUAD   | 238  | 314  | 265  | 78        | 23           | 171         | 65   | 18 | 83   | 11  |
| 30 (1)           | LUAD   | 11   | 12   | 15   | 0         | 2            | 3           | 24   | 0  | 2    | 4   |
| 31 (4)           | LUAD   | 482  | 312  | 536  | 190       | 25           | 101         | 30   | 10 | 19   | 55  |
| 32 (5)           | LUAD   | 148  | 690  | 806  | 623       | 60           | 842         | 62   | 19 | 104  | 62  |
| 33 (3)           | LUAD   | 110  | 469  | 199  | 51        | 279          | 177         | 89   | 10 | 42   | 15  |
| 34 (7)           | LUSC   | 694  | 1521 | 386  | 460       | 463          | 158         | 350  | 94 | 135  | 87  |
| 35 (4)           | LUAD   | 1023 | 280  | 262  | 141       | 2            | 329         | 448  | 21 | 34   | 33  |

| Patient (Images) | Cancer | B    | CD4T | CD8T | Dendritic | Granulocytes | Macrophages | Mast | NK  | Treg | ydT |
|------------------|--------|------|------|------|-----------|--------------|-------------|------|-----|------|-----|
| 36 (1)           | LUAD   | 12   | 26   | 103  | 0         | 97           | 57          | 15   | 9   | 0    | 3   |
| 37 (1)           | LUSC   | 5    | 4    | 8    | 0         | 7            | 4           | 1    | 0   | 1    | 0   |
| 38 (1)           | LUAD   | 275  | 73   | 85   | 22        | 1206         | 110         | 7    | 2   | 3    | 11  |
| 39 (2)           | LUAD   | 184  | 466  | 543  | 163       | 12           | 145         | 18   | 19  | 37   | 20  |
| 40 (4)           | LUAD   | 9    | 11   | 29   | 0         | 197          | 45          | 5    | 3   | 8    | 4   |
| 41 (2)           | LUSC   | 9    | 21   | 30   | 2         | 21           | 39          | 10   | 21  | 13   | 3   |
| 42 (3)           | LUAD   | 67   | 300  | 398  | 41        | 4            | 369         | 25   | 5   | 52   | 39  |
| 43 (1)           | LUSC   | 20   | 12   | 13   | 3         | 6            | 59          | 2    | 1   | 0    | 0   |
| 44 (1)           | LUSC   | 19   | 22   | 15   | 4         | 3            | 39          | 5    | 3   | 5    | 4   |
| 45 (2)           | LUAD   | 10   | 90   | 43   | 8         | 2            | 43          | 9    | 6   | 3    | 1   |
| 46 (1)           | LUSC   | 3    | 21   | 76   | 0         | 6            | 44          | 10   | 0   | 48   | 3   |
| 47 (1)           | LUAD   | 17   | 0    | 1    | 4         | 0            | 11          | 0    | 0   | 0    | 0   |
| 48 (2)           | LUAD   | 699  | 200  | 165  | 23        | 1            | 181         | 35   | 5   | 25   | 25  |
| 49 (9)           | LUAD   | 69   | 324  | 387  | 18        | 8            | 520         | 350  | 5   | 18   | 26  |
| 50 (7)           | LUAD   | 1715 | 1058 | 869  | 360       | 765          | 402         | 58   | 142 | 103  | 36  |
| 51 (1)           | LUSC   | 76   | 69   | 95   | 5         | 771          | 144         | 26   | 4   | 35   | 3   |
| 52 (7)           | LUSC   | 538  | 745  | 885  | 23        | 243          | 518         | 257  | 20  | 622  | 64  |
| 53 (5)           | LUSC   | 240  | 340  | 458  | 104       | 623          | 728         | 100  | 35  | 117  | 64  |
| 54 (6)           | LUAD   | 70   | 90   | 292  | 23        | 4            | 247         | 122  | 168 | 78   | 38  |
| 55 (6)           | LUAD   | 59   | 313  | 151  | 140       | 324          | 265         | 181  | 14  | 24   | 31  |
| 56 (2)           | LUSC   | 62   | 176  | 113  | 17        | 30           | 248         | 21   | 7   | 27   | 17  |
| 57 (1)           | LUSC   | 6    | 112  | 274  | 2         | 37           | 129         | 0    | 3   | 8    | 8   |
| 58 (1)           | LUAD   | 62   | 0    | 0    | 0         | 9            | 2           | 0    | 0   | 0    | 1   |
| 59 (6)           | LUAD   | 5    | 200  | 80   | 184       | 98           | 545         | 290  | 104 | 166  | 45  |
| 60 (1)           | LUAD   | 122  | 33   | 57   | 14        | 0            | 90          | 17   | 4   | 20   | 2   |
| 61 (1)           | LUAD   | 408  | 74   | 44   | 6         | 101          | 44          | 33   | 0   | 8    | 4   |
| 62 (1)           | LUAD   | 2    | 0    | 0    | 0         | 0            | 0           | 0    | 6   | 0    | 0   |
| 63 (1)           | LUAD   | 767  | 976  | 1409 | 358       | 37           | 817         | 43   | 38  | 245  | 37  |
| 64 (1)           | LUSC   | 13   | 108  | 62   | 2         | 0            | 22          | 9    | 2   | 1    | 0   |

## B.2 Edge corrections

It is common when computing such  $K$  functions to correct them for what are referred to as “edge effects”. Edge effects describe the issue that cells near the edge of an image appear to have fewer cells around them when  $r$  extends beyond the nearest boundary. One option is toroidal edge corrections as described in Ripley (1979), which replicates and reflects the image data occurring near the boundary. Another is isotropic edge corrections which weights edge cells, see Ohser and Stoyan (1981); Ripley (1988); Baddeley et al. (2015). We examined several methods and found similar results. We use a standard isotropic edge correction in this chapter. For a single point  $x$  in some image  $I$ , then the isotropic edge correction weight is

$$e(u; r) = \frac{2 r}{c(u; r) \setminus I} \quad (\text{B.1})$$

where the denominator is the intersection of the circle  $c(u; r)$  with radius  $r$  and center  $u$ . For discussions on approaches to edge corrections see Haase (1995); Yamada and Rogerson (2003).

## B.3 Surrogate splitting

The principle behind surrogate splits is that at each split a selection for the splitting variable must be made, even if the selection is between nearly identical variables. Traditional splitting would select only one, giving it a measure of variable importance, and ignore any other options. In this way, traditional splitting can result in masked variables and inaccurate representations of variable importances. A model using surrogate splits revisits each split after selecting the “primary” splitting variable and considers alternative, “secondary” splitting variables, mitigating the masking behavior.

Surrogate splits also take into account that the variables used after the primary split are, to some degree, lower quality and weight the splits accordingly. Weighting is performed according to a split’s impurity reduction, and varies slightly based on classification or regression problems Breiman et al. (1984). Moreover, these surrogates allow for use of another split if an observation is missing data for the primary split.

Consider that a naïve choice for any variable missing data used in the primary split would be to go with the majority split. In order to evaluate potential splits, they should perform better than this naïve approach. In particular secondary splits should outperform this approach on missing data, and perhaps even split nearly as good as the primary split on

all data. Consequently, one potential weighting scheme for node  $A$  and potential surrogate  $v$  is

$$w_v(A) = \frac{c_v}{n_A^{(\text{node})}} \frac{c_{maj}}{c_{maj}} \quad (\text{B.2})$$

where  $c_v$  is the number of observations correctly classified due to splitting by variable  $v$ ,  $c_{maj}$  is the number of observations correct by naively selecting the majority class, and  $n_A^{(\text{node})}$  is the total number of observations in the node.

A typical variable may appear many times in a tree, both as a primary and surrogate splitting variable. Therefore, the overall measure of variable importance for a given variable is a combination of the splitting importances

$$\text{VI}(v) = \frac{1}{\sqrt{\text{Number of Trees}_j}} \left( \sum_{i \in P_v} I_i + \sum_{s \in S_v} w_s I_s \right) \quad (\text{B.3})$$

where for variable  $v$ ,  $P_v$  and  $S_v$  respectively indicate the set of primary and surrogate splits.

## B.4 Choosing $B$ and $H$

The parameters  $B$  and  $H$  are user selected and are used in computing the thresholds for evaluating variable importance. The parameter  $B$  defines the number of synthetic spatial variables simulated independently of the response of interest used in the proposed procedure while parameter  $H$  defines the number of iterations for quantile estimation. This section demonstrates the effect of these choices via simulations similar to those in the section Simulation study in the main manuscript, which mimic the TNBC dataset.

Throughout these simulations, we set  $H = B$ . We considered both the 4 cell type and 16 cell type experiments as defined in the section Simulation study for varying choices of  $B$ . The data were simulated so that only one interaction ( $c2\_c2$ ) was related to the response, and all other interactions were simulated independently of the response. A mild interaction effect was imposed, with underlying standard deviation in placement of the cells changing from 1=25 to 1=60 between the binary outcomes for the modified Thomas process. The percentage of variable importance values exceeding the noise and interpolation thresholds, calibrated to the 95% level were recorded for 50 independent simulation trials. Figure B.1 shows these percentages as a function of  $B \in \{50; 100; 250; 500\}$  for the interaction that has a difference between groups ( $c2\_c2$ ) and for the interactions that do not have a difference (shown as the mean of all other interactions). We observed that the two thresholds (i.e. the

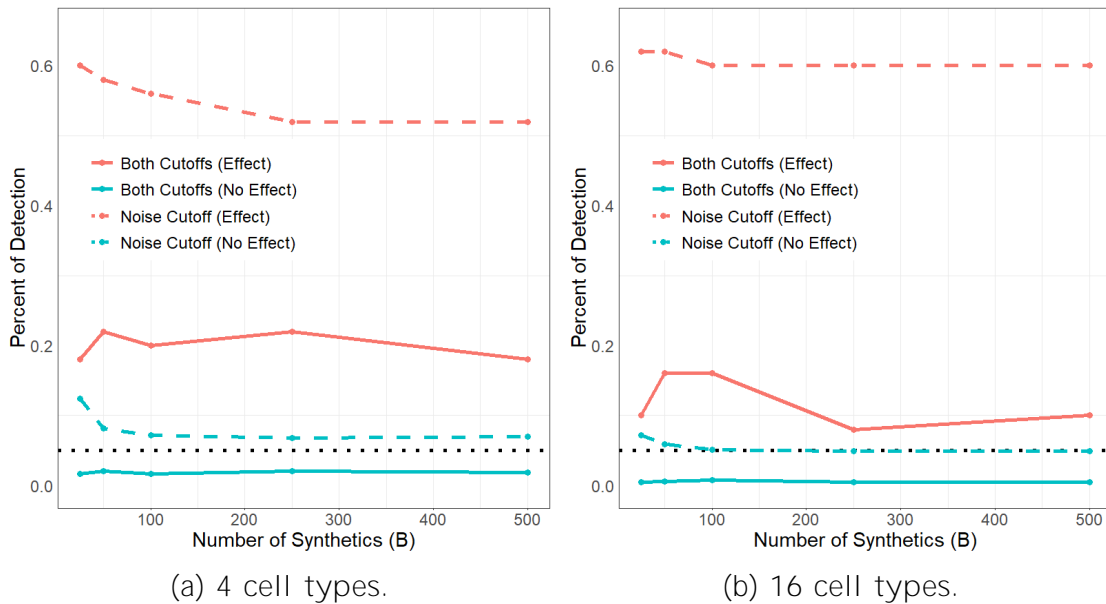


Figure B.1: Selection of  $B$ . Simulations of into the number of synthetics,  $B$ , for (a) 4 and (b) 16 cell scenarios. In both, only one interaction,  $c2\_c2$ , is significant, and all other interactions are non-significant. Per Figure 8 in the main manuscript, the expected effect is mild (standard deviation in placement of the cells from the Thomas process changing from  $1=25$  to  $1=60$ ). Running 50 trials for significance of 0.05, the power is reasonably level across selections of  $B$ , excluding (expected) low power when the number of synthetics is far too low to accurately approximate the desired quantile.

95% threshold using noise only and that using both noise and the interpolation threshold) showed relative stability in terms of their false positive and false negative rates for all values of  $B$  exceeding 50. In terms of estimating the two thresholds at the 95% level, there was no apparent advantage to choosing a larger value of  $B$  for the number of cell interactions considered. Since increasing  $B$  increases the computational burden without obvious benefits we recommend  $B = H = 100$  as default choices.

## B.5 OOB accuracy and naïve model comparisons

We also investigate the summary metrics OOB, BIAS, and GUESS. We conduct simulations similar to those described in the Simulation study section – we consider the case of 16 cell types in two scenarios: one in which the spatial interactions were independent of the

Table B.2: Null summary metrics table. In 100 simulations without differences between the outcome groups (i.e., the null scenario), we compare the out-of-bag (OOB) accuracy of the true model to the naïve approaches of always selecting the most common outcome (BIAS) and selecting a random outcome based on frequency (GUESS)– see Predictive accuracy estimates and Simulation study sections in the main manuscript for more details. The table shows the methods are unskewed and all randomly report the largest values about an equal number of times (about one third for each metric). Compare this table to Table B.3.

|                | OOB   | BIAS  | GUESS |
|----------------|-------|-------|-------|
| Mean           | 0.431 | 0.500 | 0.500 |
| Median         | 0.438 | 0.500 | 0.500 |
| Std. Deviation | 0.134 | 0.000 | 0.000 |
| Largest (%)    | 0.360 | 0.640 |       |

response (null) and when they were dependent on the response (alternative). Each scenario was simulated 100 times, and we recorded the OOB value. The summary results of these simulations are given in Tables B.2 and B.3. In the null case, the OOB was largest among the three metrics at approximately the same rate. However, in the alternative case, OOB far exceed the other estimates in terms of value and rate being the largest.

## B.6 Performance in the low sample, high interaction count scenario

It is of value to know the performance, i.e. the detection and false discovery rates, of models in sub-optimal situations. For our model, sub-optimal can be taken to indicate a small sample size or a large number of interactions.

In this section, we consider simulations for small samples of 20 patients, 10 “positive” and 10 “negative” with an increasing number of interactions. For each patient, a single image is recorded, with no meta-variables. The cells in the images are placed according the process described in Simulation study. No interactions between cell types are present and all images contain only a single, weak interaction between one cell type that differs between the positive and negative groups. The number of each cell type is random, selected via a Binomial distribution with 50% inclusion and 100 possible cells. The number of cell types, and therefore the number of interactions, is varied and each scenario is iterated 100

Table B.3: Alternative summary metrics table. In 100 simulations with a difference between the outcome groups (i.e., the “alternative” scenario), we compare the out-of-bag (OOB) accuracy of the true model to the naïve approaches of always selecting the most common outcome (BIAS) and selecting a random outcome based on frequency (GUESS)—see Predictive accuracy estimates and Simulation study sections in the main manuscript for more details. The table shows the methods are unskewed and OOB consistently reports the largest values, showing the model does capture the existing information—that is the model does seem to recognize informative interactions and/or meta-variables. Compare this table to Table B.2.

|                | OOB   | BIAS  | GUESS |
|----------------|-------|-------|-------|
| Mean           | 0.845 | 0.500 | 0.500 |
| Median         | 0.850 | 0.500 | 0.500 |
| Std. Deviation | 0.072 | 0.000 | 0.000 |
| Largest (%)    | 1.000 | 0.000 |       |

times. In each iteration, the variable importance, and its relation to the thresholds, is recorded to investigate the ability of funkycells to detect the significant interaction and mitigate false discovery of non-significant interactions.

The results of the simulation, showing the detection of the significant interaction (red) and the mean of all non-significant interactions (blue), is given in Figure B.1. In the 4 cell type case, the significant interaction is detected only slightly over 50% of the time. Yet, the power loss is small as the number of unique cell types increases. Yet while 4 unique types corresponds to only 10 interactions, 24 unique types corresponds to 300 interactions. Even under the high interaction count scenario, the method detects the true interaction far more than any other non-significant interaction. In fact, using both thresholds the model becomes more conservative as the number of interactions is increased, managing the false discovery rate. Unreported figures considering only the noise threshold or only the interpolation threshold shows each holds detection of non-significant interactions to the specified significance level .

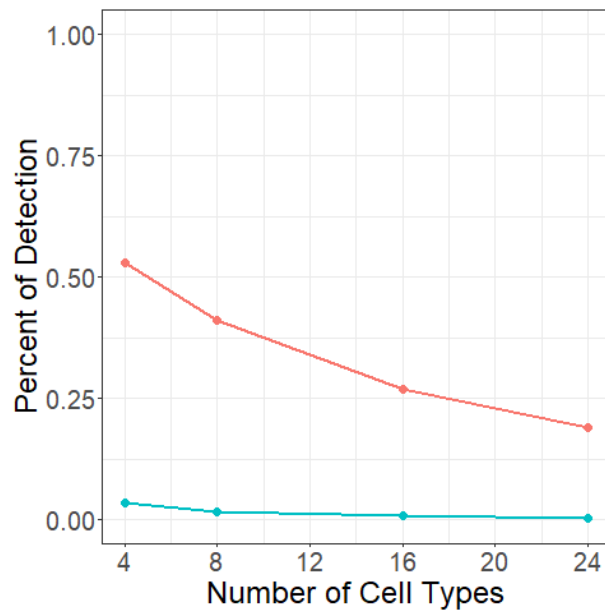


Figure B.1: Detection rate. Mean detection rate of the significant interaction (red) and mean detection rate of all non-significant interactions (blue) in a small sample scenario with an increasing number of cell types. In each scenario, 100 iterations per scenario, all interactions were considered ( $T(T + 1)=2$  interactions). As the number of interactions increase, the method suffers only minor power loss while controlling for false discovery rate.

# Appendix C

## Chapter 5 Appendix

### C.1 Technical lemmas

We report all the technical lemmas used in the derivation of our main results; all proofs are relegated to the Supplement to save space.

We start with some notation and formulas. Our proofs require the computation of the first and second derivatives of  $\ell_k(\mathbf{u})$ ;  $\mathbf{u} = (u_1; u_2; u_3) \in \mathbb{R}^3$ , which are given by:

$$\begin{aligned} \frac{\partial}{\partial u_1} \ell_i(\mathbf{u}) &= \frac{(y_i - u_1 y_{i-1}) y_{i-1}}{u_2 y_{i-1}^2 + u_3}; & \frac{\partial}{\partial u_2} \ell_i(\mathbf{u}) &= \frac{1}{2} \left[ \frac{y_{i-1}^2}{u_2 y_{i-1}^2 + u_3} - \frac{(y_i - u_1 y_{i-1})^2 y_{i-1}^2}{(u_2 y_{i-1}^2 + u_3)^2} \right] \\ \frac{\partial}{\partial u_3} \ell_i(\mathbf{u}) &= \frac{1}{2} \left[ \frac{1}{u_2 y_{i-1}^2 + u_3} - \frac{(y_i - u_1 y_{i-1})^2}{(u_2 y_{i-1}^2 + u_3)^2} \right] \end{aligned} \tag{C.1}$$

and

$$\begin{aligned}
\frac{\partial^2}{\partial u_1 \partial u_1} \ell_i(\mathbf{u}) &= \frac{y_{i-1}^2}{u_2 y_{i-1}^2 + u_3}; & \frac{\partial^2}{\partial u_1 \partial u_2} \ell_i(\mathbf{u}) &= \frac{(y_i - u_1 y_{i-1}) y_{i-1}^3}{(u_2 y_{i-1}^2 + u_3)^2}; \\
\frac{\partial^2}{\partial u_1 \partial u_3} \ell_i(\mathbf{u}) &= \frac{(y_i - u_1 y_{i-1}) y_{i-1}}{(u_2 y_{i-1}^2 + u_3)^2}; \\
\frac{\partial^2}{\partial u_2 \partial u_2} \ell_i(\mathbf{u}) &= \frac{1}{2} \frac{y_{i-1}^4}{(u_2 y_{i-1}^2 + u_3)^2} - \frac{(y_i - u_1 y_{i-1})^2 y_{i-1}^4}{(u_2 y_{i-1}^2 + u_3)^3}; \\
\frac{\partial^2}{\partial u_2 \partial u_3} \ell_i(\mathbf{u}) &= \frac{1}{2} \frac{y_{i-1}^2}{(u_2 y_{i-1}^2 + u_3)^2} - \frac{(y_i - u_1 y_{i-1})^2 y_{i-1}^2}{(u_2 y_{i-1}^2 + u_3)^3}; \\
\frac{\partial^2}{\partial u_3 \partial u_3} \ell_i(\mathbf{u}) &= \frac{1}{2} \frac{1}{(u_2 y_{i-1}^2 + u_3)^2} - \frac{(y_i - u_1 y_{i-1})^2}{(u_2 y_{i-1}^2 + u_3)^3};
\end{aligned} \tag{C.2}$$

We begin with a series of Lemmas (Lemmas C.1.1-C.1.7) which cover the stationary case. Recall that the necessary and sufficient condition for stationarity is

$$1 - E \log j + \rho_{i,j} < 0; \tag{C.3}$$

and in this case [Aue et al. \(2006\)](#) showed that  $y_i$  is exponentially close to  $\bar{y}_i$ , the unique stationary sequence satisfying

$$\bar{y}_i = (\rho_{i,1} + \rho_{i,2}) \bar{y}_{i-1} + \rho_{i,3}; \quad 1 < i < \infty; \tag{C.4}$$

We begin by showing that it is enough to investigate the likelihood function which is based on the stationary sequence  $y_i$ . Hence we replace  $y_i$  with  $\bar{y}_i$  in the definition of the likelihood function and work with

$$\bar{\ell}_i(v_1; v_2; v_3) = \frac{1}{2} \left( \log(v_2 \bar{y}_{i-1}^2 + v_3) + \frac{(v_1 - v_2 \bar{y}_{i-1})^2}{v_2 \bar{y}_{i-1}^2 + v_3} \right);$$

Lemma C.1.1. *Let  $g(\cdot)$  be a function, and denote its derivative as  $g'(\cdot)$ , both bounded on the real line. Under Assumption 6.2.1, it holds that*

$$E j g(y_i) - g(\bar{y}_i) j \leq c_0 \rho_{i,j};$$

where  $0 < c_0 < 1$  and  $0 < \rho_{i,j} < 1$ .

Lemma C.1.2. *If the conditions of Theorem 6.2.1 are satisfied, and Eq (C.3) holds, then we have*

$$\lim_{C \downarrow 1} \limsup_{n \uparrow} P \left\{ \max_{1 \leq k \leq n} \sup_{\mathbf{u} \in \mathcal{U}} \left| \sum_{i=1}^k (\dot{\ell}_i(u_1; u_2; u_3) - \dot{\ell}_i(u_1; u_2; u_3)) \right| > C \right\} = 0; \quad (\text{C.5})$$

$$\lim_{C \downarrow 1} \limsup_{n \uparrow} P \left\{ \max_{1 \leq k \leq n} \sup_{\mathbf{u} \in \mathcal{U}} \left| \sum_{i=k+1}^n (\dot{\ell}_i(u_4; u_2; u_3) - \dot{\ell}_i(u_4; u_2; u_3)) \right| > C \right\} = 0; \quad (\text{C.6})$$

$$\lim_{M \downarrow 1} \limsup_{n \uparrow} P \left\{ \max_{M \leq k \leq n} \sup_{\mathbf{u} \in \mathcal{U}} \frac{1}{k} \left| \sum_{i=1}^k (\dot{\ell}_i(u_1; u_2; u_3) - E \dot{\ell}_0(u_1; u_2; u_3)) \right| > \right\} = 0; \quad (\text{C.7})$$

and

$$\lim_{M \downarrow 1} \limsup_{n \uparrow} P \left\{ \max_{1 \leq k \leq n} \sup_{\mathbf{u} \in \mathcal{U}} \frac{1}{n-k} \left| \sum_{i=k+1}^n (\dot{\ell}_i(u_4; u_2; u_3) - E \dot{\ell}_0(u_4; u_2; u_3)) \right| > \right\} = 0; \quad (\text{C.8})$$

for all  $\epsilon > 0$ .

Lemma C.1.3. *If the conditions of Theorem 6.2.1 are satisfied, and Eq (C.3) holds, then we have*

$$\lim_{M \downarrow 1} \limsup_{n \uparrow} P \left\{ \max_{M \leq k \leq n} \left\| \left( \hat{\Lambda}_{k;1}, \hat{\Lambda}_{k;2}, \hat{\Lambda}_{k;1}^2, \hat{\Lambda}_{k;2}^2 \right) \left( \dot{\ell}_1; \dot{\ell}_2; \frac{\partial}{\partial u_1}; \frac{\partial}{\partial u_2} \right) \right\| > \right\} = 0$$

for all  $\epsilon > 0$ .

We introduce the notation  $\hat{L}_{n;k}(u_1; u_2; u_3; u_4) = L_{n;k}(u_1; u_4; u_2; u_3)$ . Let

$$r \hat{L}_{n;k}(\mathbf{u}) = \left( \frac{\partial}{\partial u_1} L_{n;k}(\mathbf{u}); \frac{\partial}{\partial u_4} L_{n;k}(\mathbf{u}); \frac{\partial}{\partial u_2} L_{n;k}(\mathbf{u}); \frac{\partial}{\partial u_3} L_{n;k}(\mathbf{u}) \right)^T \quad (\text{C.9})$$

and define the matrix of the second derivatives  $\mathbf{G}_{n;k}(\mathbf{u}) = (fg_{n;k}^{(i;j)}; 1 \leq i, j \leq 4) g = r^2 L_{n;k}(\mathbf{u})$ .

We introduce the stationary version of  $L_{n;k}(\mathbf{u})$

$$L_{n;k}(\mathbf{u}) = \sum_{i=1}^k \dot{\ell}_i(u_1; u_2; u_3) + \sum_{i=k+1}^n \dot{\ell}_i(u_4; u_2; u_3)$$

and the corresponding first and second derivatives:

$$r L_{n;k}(\mathbf{u}) = \left( \frac{\partial}{\partial U_1} L_{n;k}(\mathbf{u}); \frac{\partial}{\partial U_4} L_{n;k}(\mathbf{u}); \frac{\partial}{\partial U_2} L_{n;k}(\mathbf{u}); \frac{\partial}{\partial U_3} L_{n;k}(\mathbf{u}) \right)^>; \quad (\text{C.10})$$

$$\mathbf{G}_{n;k}(\mathbf{u}) = f \mathbf{g}_{n;k}^{(ij)}; 1 \quad i; j \quad 4g = r^2 L_{n;k}(\mathbf{u})$$

and  $\mathfrak{G}_{n;k}(\mathbf{u}) = f \mathfrak{g}_{n;k}^{(ij)}; 1 \quad i; j \quad 4g = E \mathbf{G}_{n;k}(\mathbf{u})$ .

Lemma C.1.4. *If the conditions of Theorem 6.2.1 are satisfied, and Eq (C.3) holds, then we have*

$$\lim_{M! \rightarrow 1} \limsup_{n! \rightarrow 1} P \left\{ \max_{M \leq k \leq n} \sup_{\mathbf{u} \in \mathcal{U}} \frac{1}{k} \left| g_{n;k}^{(1;j)}(\mathbf{u}) - \mathfrak{g}_{n;k}^{(1;j)}(\mathbf{u}) \right| > \right\} = 0; \quad 1 \leq j \leq 4 \quad (\text{C.11})$$

$$\lim_{M! \rightarrow 1} \limsup_{n! \rightarrow 1} P \left\{ \max_{M \leq k \leq n} \sup_{\mathbf{u} \in \mathcal{U}} \frac{1}{n-k} \left| g_{n;k}^{(2;j)}(\mathbf{u}) - \mathfrak{g}_{n;k}^{(2;j)}(\mathbf{u}) \right| > \right\} = 0; \quad 1 \leq j \leq 4 \quad (\text{C.12})$$

and

$$\lim_{M! \rightarrow 1} \limsup_{n! \rightarrow 1} P \left\{ \max_{M \leq k \leq n} \sup_{\mathbf{u} \in \mathcal{U}} \frac{1}{n} \left| g_{n;k}^{(ij)}(\mathbf{u}) - \mathfrak{g}_{n;k}^{(ij)}(\mathbf{u}) \right| > \right\} = 0; \quad 3 \leq i; j \leq 4 \quad (\text{C.13})$$

for all  $\epsilon > 0$ .

We now turn to considering the inverse of the matrix of the second derivatives, defining  $\mathbf{V}_{n;k}(\mathbf{u}) = \mathbf{G}_{n;k}^{-1}(\mathbf{u}) = f \mathbf{V}_{n;k}^{(ij)}; 1 \quad i; j \quad 4g$ , and  $\mathfrak{V}_{n;k}(\mathbf{u}) = \mathfrak{G}_{n;k}^{-1}(\mathbf{u}) = f \mathfrak{V}_{n;k}^{(ij)}; 1 \quad i; j \quad 4g$ .

Lemma C.1.5. *If the conditions of Theorem 6.2.1 are satisfied and Eq (C.3) holds, then we have for all  $\epsilon > 0$*

$$\lim_{M! \rightarrow 1} \limsup_{n! \rightarrow 1} P \left\{ \max_{M \leq k \leq n} \sup_{\mathbf{u} \in \mathcal{U}} k \left| v_{n;k}^{(1;1)}(\mathbf{u}) - \mathfrak{v}_{n;k}^{(1;1)}(\mathbf{u}) \right| > \right\} = 0; \quad (\text{C.14})$$

$$\lim_{M! \rightarrow 1} \limsup_{n! \rightarrow 1} P \left\{ \max_{M \leq k \leq n} \sup_{\mathbf{u} \in \mathcal{U}} (n-k) \left| v_{n;k}^{(2;2)}(\mathbf{u}) - \mathfrak{v}_{n;k}^{(2;2)}(\mathbf{u}) \right| > \right\} = 0; \quad (\text{C.15})$$

and

$$\lim_{M! \rightarrow 1} \limsup_{n! \rightarrow 1} P \left\{ \max_{M \leq k \leq n} \sup_{\mathbf{u} \in \mathcal{U}} n \left| v_{n;k}^{(ij)}(\mathbf{u}) - \mathfrak{v}_{n;k}^{(ij)}(\mathbf{u}) \right| > \right\} = 0; \quad (\text{C.16})$$

for all  $(i; j); 1 \leq i; j \leq 4$ , except  $(1; 1)$  and  $(2; 2)$ .

Lemma C.1.6. *If the conditions of Theorem 6.2.1 are satisfied and Eq (C.3) holds, then we have*

$$\lim_{C \downarrow 1} \limsup_{M \downarrow 1} \limsup_{n \downarrow 1} P \left\{ \max_{M \leq k \leq n} \left( \frac{k}{\log \log k} \right)^{1=2} j^{\wedge}_{k,1} \quad j > C \right\} = 0; \quad (\text{C.17})$$

$$\lim_{C \downarrow 1} \limsup_{M \downarrow 1} \limsup_{n \downarrow 1} P \left\{ \max_{M \leq k \leq n} \left( \frac{n-k}{\log \log(n-k)} \right)^{1=2} j^{\wedge}_{k,2} \quad j > C \right\} = 0; \quad (\text{C.18})$$

$$\lim_{C \downarrow 1} \limsup_{M \downarrow 1} \limsup_{n \downarrow 1} P \left\{ n^{1=2} \max_{M \leq k \leq n} j^{\wedge 2}_{k,1} \quad \frac{2}{1}j > C \right\} = 0; \quad (\text{C.19})$$

and

$$\lim_{C \downarrow 1} \limsup_{M \downarrow 1} \limsup_{n \downarrow 1} P \left\{ n^{1=2} \max_{M \leq k \leq n} j^{\wedge 2}_{k,2} \quad \frac{2}{2}j > C \right\} = 0; \quad (\text{C.20})$$

Let

$$\mathbf{G}_0 = \begin{pmatrix} ; & 0; & 0; & 0 \\ 0; & ; & 0; & 0 \\ 0; & 0; & 1;1; & 1;2 \\ 0; & 0; & 2;1; & 2;2 \end{pmatrix}; \quad \mathbf{\Lambda} = \begin{pmatrix} 1;1; & 1;2 \\ 2;1; & 2;2 \end{pmatrix}; \quad (\text{C.21})$$

$$= \frac{1}{2} E \left[ \frac{y_0^2}{\frac{2}{1}y_0^2 + \frac{2}{2}} \right]; \quad 1;1 = \frac{1}{2} E \left[ \frac{y_0^4}{(\frac{2}{1}y_0^2 + \frac{2}{2})^2} \right];$$

$$1;2 = 2;1 = \frac{1}{2} E \left[ \frac{y_0^2}{(\frac{2}{1}y_0^2 + \frac{2}{2})^2} \right]; \quad 2;2 = \frac{1}{2} \left[ E \frac{1}{\frac{2}{1}y_0^2 + \frac{2}{2}} \right];$$

and

$$\mathbf{z}_k = \left( \sum_{i=1}^k i;1; \sum_{i=k+1}^n i;1; \sum_{i=1}^n i;2; \sum_{i=1}^n i;3 \right)^{\wedge} = (Z_{k,1}; Z_{k,2}; Z_{n,3}; Z_{n,4})^{\wedge};$$

where we define  $\mathbf{u}_0 = ( ; \frac{2}{1}; \frac{2}{2})^{\wedge}$ , and

$$i;1 = \frac{\partial}{\partial U_1} i(\mathbf{u}_0); \quad i;2 = \frac{\partial}{\partial U_2} i(\mathbf{u}_0); \quad \text{and} \quad i;3 = \frac{\partial}{\partial U_3} i(\mathbf{u}_0);$$

and the corresponding stationary versions

$$i;1 = \frac{\partial}{\partial U_1} i(\mathbf{u}_0); \quad i;2 = \frac{\partial}{\partial U_2} i(\mathbf{u}_0); \quad \text{and} \quad i;3 = \frac{\partial}{\partial U_3} i(\mathbf{u}_0);$$

Lemma C.1.7. *If the conditions of Theorem 6.2.1 are satisfied and Eq (C.3) holds, then we have*

$$\left( k(\hat{\alpha}_{k,1}) ; (n-k)(\hat{\alpha}_{k,2}) ; n(\hat{\alpha}_{k,1}^2) ; n(\hat{\alpha}_{k,2}^2) \right)^> = \mathbf{G}_0^{-1} \mathbf{z}_k + \mathbf{R}_k; \quad (\text{C.22})$$

with  $\mathbf{R}_k = (R_{k,1}; \dots; R_{k,4})^>$  satisfying

$$\lim_{C \downarrow 1} \limsup_{M \downarrow 1} \limsup_{n \downarrow 1} P \left\{ \max_{M \leq k \leq n} \max_{M \leq j \leq n} j R_{k,1j} = \log \log k > C \right\} = 0;$$

$$\lim_{C \downarrow 1} \limsup_{M \downarrow 1} \limsup_{n \downarrow 1} P \left\{ \max_{M \leq k \leq n} \max_{M \leq j \leq n} j R_{k,2j} = \log \log(n-k) > C \right\} = 0;$$

$$\lim_{C \downarrow 1} \limsup_{M \downarrow 1} \limsup_{n \downarrow 1} P \left\{ \max_{M \leq k \leq n} \max_{M \leq j \leq n} j R_{k,3j} > C \right\} = 0;$$

and

$$\lim_{C \downarrow 1} \limsup_{M \downarrow 1} \limsup_{n \downarrow 1} P \left\{ \max_{M \leq k \leq n} \max_{M \leq j \leq n} j R_{k,4j} > C \right\} = 0;$$

We now report a series of Lemmas (Lemmas C.1.8-C.1.13) which are used to study the nonstationary case

$$E \log j + o_p(j) > 0; \quad (\text{C.23})$$

In such a case, our proofs are based on the observation that  $j y_j \rightarrow 1$  in probability: [Berkes et al. \(2009\)](#) showed that

$$\exp(-S(i)) y_i \rightarrow Y_i; \quad (\text{C.24})$$

where  $S(i) = \sum_{k=1}^i \log j + o_p(j)$  a.s., and  $P f Y = 0 g = 0$ .

We begin by showing that we can approximate the log likelihood function with

$$g_i(u_1; u_2) = \frac{1}{2} \left( \log(u_2 y_i^2 + 1) + \frac{(y_i - u_1 y_i)^2}{u_2 y_i^2 + 1} \right);$$

Lemma C.1.8. *If the conditions of Theorem 6.2.1 are satisfied and Eq (C.23) holds, then we have*

$$\max_{1 \leq k \leq n} \sup_{\mathbf{u} \in \mathbf{U}} \left| \sum_{i=1}^k \log \ell_i(\mathbf{u}) - \sum_{i=1}^k g_i(u_1; u_2) \right| = O_P(1); \quad (\text{C.25})$$

and

$$\max_{1 \leq k < n} \sup_{\mathbf{u} \in \mathbf{U}} \left| \sum_{i=k+1}^n \log \ell_i(\mathbf{u}) - \sum_{i=k+1}^n g_i(u_1; u_2) \right| = O_P(1); \quad (\text{C.26})$$

Let

$$g(u_1; u_2) = \frac{1}{2} \log \left( \frac{u_2}{\frac{1}{2}} \right) - \frac{1}{2} \left( \frac{\frac{1}{2} u_2}{u_2} \right) + \frac{1}{2} \frac{u_1^2}{u_2}.$$

Lemma C.1.9. *If the conditions of Theorem 6.2.1 are satisfied and Eq (C.23) holds, then we have*

$$\max_{1 \leq k \leq n} \sup_{\mathbf{u} \in \mathcal{U}} \left| \sum_{i=1}^k [g_i(u_1; u_2) - g_i(\cdot; \frac{1}{2}) - g(u_1; u_2)] \right| = O_P(1); \quad (\text{C.27})$$

and

$$\max_{1 \leq k \leq n} \sup_{\mathbf{u} \in \mathcal{U}} \left| \sum_{i=k+1}^n [g_i(u_1; u_2) - g_i(\cdot; \frac{1}{2}) - g(u_1; u_2)] \right| = O_P(1); \quad (\text{C.28})$$

Let

$$h_1(u_1; u_2) = l_{i,1}(\mathbf{u}) = \frac{u_1 + u_{i,1}}{u_2}; \quad l_{i,2}(u_1; u_2) = l_{i,2}(\mathbf{u}) = \frac{1}{2} \left( \frac{1}{u_2} - \frac{(u_1 + u_{i,1})^2}{u_2^2} \right);$$

$$l_2(\mathbf{u}) = 0; \quad l_{i,1,1}(u_1; u_2) = l_{i,1,1}(\mathbf{u}) = \frac{1}{u_2}; \quad l_{i,1,2}(u_1; u_2) = l_{i,1,2}(\mathbf{u}) = \frac{u_1 + u_{i,1}}{u_2^2};$$

$$l_{2,2}(u_1; u_2) = l_{2,2}(\mathbf{u}) = \frac{1}{2u_2^2} - \frac{(u_1 + u_{i,1})^2}{u_2^3}; \quad l_{i,3}(\mathbf{u}) = l_{2,3}(\mathbf{u}) = l_{3,3}(\mathbf{u}) = 0;$$

$$l_{k,\cdot}(\mathbf{u}) = l_{\cdot,k}(\mathbf{u}); \quad 1 \leq k \leq 3;$$

Lemma C.1.10. *If the conditions of Theorem 6.2.1 are satisfied and Eq (C.23) holds, then we have*

$$\max_{1 \leq k \leq n} \sup_{\mathbf{u} \in \mathcal{U}} \left| \sum_{i=1}^k \left( \frac{\partial}{\partial u_j} l_{i,j}(\mathbf{u}) - l_{i,j}(\mathbf{u}) \right) \right| = O_P(1); \quad j = 1; 2; 3;$$

$$\max_{1 \leq k \leq n} \sup_{\mathbf{u} \in \mathcal{U}} \left| \sum_{i=k+1}^n \left( \frac{\partial}{\partial u_j} l_{i,j}(\mathbf{u}) - l_{i,j}(\mathbf{u}) \right) \right| = O_P(1); \quad j = 1; 2; 3;$$

$$\max_{1 \leq k \leq n} \sup_{\mathbf{u} \in \mathcal{U}} \left| \sum_{i=1}^k \left( \frac{\partial^2}{\partial u_j \partial u_n} l_{i,j,n}(\mathbf{u}) - l_{i,j,n}(\mathbf{u}) \right) \right| = O_P(1); \quad j; n = 1; 2; 3;$$

$$\max_{1 \leq k \leq n-1} \sup_{\mathbf{u} \in \mathcal{U}} \left| \sum_{i=k+1}^n \left( \frac{\partial^2}{\partial u_j \partial u_n} \ell_i(\mathbf{u}) - l_{i,j;n}(\mathbf{u}) \right) \right| = O_P(1); \quad j; n = 1; 2; 3;$$

The result implies immediately an approximation with independent identically distributed random variables, viz.

$$\max_{1 \leq k \leq n} \sup_{u_3 \geq 1} \left| \sum_{i=1}^k \left( \frac{\partial}{\partial u_j} \ell_i(\cdot; \frac{2}{1}; u_3) - l_{i,j}(\cdot; \frac{2}{1}) \right) \right| = O_P(1); \quad j = 1; 2; 3; \quad (\text{C.29})$$

and

$$\max_{1 \leq k \leq n-1} \sup_{u_3 \geq 1} \left| \sum_{i=k+1}^n \left( \frac{\partial}{\partial u_j} \ell_i(\cdot; \frac{2}{1}; u_3) - l_{i,j}(\cdot; \frac{2}{1}) \right) \right| = O_P(1); \quad j = 1; 2; 3; \quad (\text{C.30})$$

with any  $0 < \epsilon < 1$ . This means that the approximating sequence does not depend on  $u_3$  and it is uniform in  $u_3$ .

The estimates  $\hat{\lambda}_{k,1}$ ,  $\hat{\lambda}_{k,2}$  and  $\hat{\lambda}_{k,1}^2$  depend on  $u_4$  of the likelihood function  $\hat{L}_{n,k}(\mathbf{u})$ , and we stress this dependence in the next lemmas to show that in the limit  $u_4$  disappears.

Lemma C.1.11. *If the conditions of Theorem 6.2.1 are satisfied and Eq (C.23) holds, then we have*

$$\lim_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} P \left\{ \max_{M \leq k \leq n} \sup_{\underline{u}_4 \leq u_4 \leq \bar{u}_4} \left\| \left( \hat{\lambda}_{k,1}(u_4); \hat{\lambda}_{k,2}(u_4); \hat{\lambda}_{k,1}^2(u_4) \right) - \left( \cdot; \cdot; \frac{2}{1} \right) \right\| > \epsilon \right\} = 0;$$

for all  $\epsilon > 0$

Lemma C.1.12. *If the conditions of Theorem 6.2.1 are satisfied and Eq (C.23) holds, then we have*

$$\lim_{C \rightarrow \infty} \limsup_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} P \left\{ \max_{M \leq k \leq n} \sup_{\underline{u}_4 \leq u_4 \leq \bar{u}_4} \left( \frac{k}{\log \log k} \right)^{1-2} j^{\hat{\lambda}_{k,1}(u_4)} - j > C \right\} = 0; \quad (\text{C.31})$$

$$\lim_{C \downarrow 1} \limsup_{M \downarrow 1} \limsup_{n \downarrow 1} P \left\{ \max_{M \leq k \leq n} \sup_{\underline{u}_4 \leq u_4 \leq \bar{u}_4} \left( \frac{n-k}{\log \log(n-k)} \right)^{1=2} j^{\wedge}_{k;2}(u_4) \quad j > C \right\} = 0; \quad (\text{C.32})$$

$$\lim_{C \downarrow 1} \limsup_{M \downarrow 1} \limsup_{n \downarrow 1} P \left\{ \max_{M \leq k \leq n} \sup_{\underline{u}_4 \leq u_4 \leq \bar{u}_4} n^{1=2} j^{\wedge}_{k;1}(u_4) \quad j > C \right\} = 0; \quad (\text{C.33})$$

Let

$$\mathbf{H} = \begin{pmatrix} \frac{1}{2}; & 0; & 0 \\ 0; & \frac{1}{2}; & 0 \\ 0; & 0; & \frac{1}{2} \end{pmatrix}$$

and

$$\mathbf{w}(k) = \left( \sum_{i=1}^k \frac{i;1}{2}; \sum_{i=k+1}^n \frac{i;1}{2}; \sum_{i=1}^n \frac{2}{2} \frac{2}{4} \right)^>$$

Lemma C.1.13. *If the conditions of Theorem 6.2.1 are satisfied and Eq (C.23) holds, then we have*

$$(k^{\wedge}_{k;1}(u_4); (n-k)^{\wedge}_{k;2}(u_4); n^{\wedge}_{k;1}(u_4))> = \mathbf{H}\mathbf{w}(k) + \mathbf{R}_k$$

with  $\mathbf{R}_k(u_4) = (R_{k;1}(u_4); R_{k;2}(u_4); R_{k;3}(u_4))>$  satisfying

$$\lim_{C \downarrow 1} \limsup_{n \downarrow 1} P \left\{ \max_{1 \leq k \leq n} \sup_{\underline{u}_4 \leq u_4 \leq \bar{u}_4} j R_{k;1}(u_4) j = \log \log k > C \right\} = 0;$$

$$\lim_{C \downarrow 1} \limsup_{n \downarrow 1} P \left\{ \max_{1 \leq k \leq n} \sup_{\underline{u}_4 \leq u_4 \leq \bar{u}_4} j R_{k;2}(u_4) j = \log \log(n-k) > C \right\} = 0;$$

and

$$\lim_{C \downarrow 1} \limsup_{n \downarrow 1} P \left\{ \max_{1 \leq k \leq n} \sup_{\underline{u}_4 \leq u_4 \leq \bar{u}_4} j R_{k;3}(u_4) j > C \right\} = 0;$$

## C.2 Proofs

Proof of Theorem 6.2.1

We divide the proof into the stationary case,  $E \log j + \rho_1 j < 0$ , and the explosive case,  $E \log j + \rho_1 j > 0$ .

We begin by considering the former case. Using a two term Taylor expansion we get

$$\begin{aligned} \hat{L}_{n;k}(\mathbf{u}_0) - \hat{L}_{n;k}(\hat{\mathbf{u}}_{n;k}) &= r \hat{L}_{n;k}(\hat{\mathbf{u}}_{n;k})^{\wedge_{n;k}} + \frac{1}{2} (\hat{\mathbf{u}}_{n;k})^{\triangleright} r^2 \hat{L}_{n;k}(\mathbf{u})^{\wedge_{n;k}} \\ &= \frac{1}{2} (\hat{\mathbf{u}}_{n;k})^{\triangleright} r^2 \hat{L}_{n;k}(\mathbf{u})^{\wedge_{n;k}}, \end{aligned} \quad (\text{B.1})$$

where

$$\hat{\mathbf{u}}_{n;k}^{\triangleright} = \left( \hat{u}_{k,1} \quad ; \quad \hat{u}_{k,2} \quad ; \quad \hat{u}_{k,1}^2 \quad \frac{2}{1} \hat{u}_{k,2}^2 \quad \frac{2}{2} \right)^{\triangleright}; \quad (\text{B.2})$$

and with some  $\mathbf{u} \geq \mathcal{R}^4$ , where  $\hat{\mathbf{u}}_{n;k} = (\hat{u}_{k,1}; \hat{u}_{k,2}; \hat{u}_{k,1}^2; \hat{u}_{k,2}^2)$ . Following the proof of Lemma C.1.6, we get

$$(\hat{\mathbf{u}}_{n;k})^{\triangleright} r^2 \hat{L}_{n;k}(\mathbf{u})^{\wedge_{n;k}} = (\hat{\mathbf{u}}_{n;k})^{\triangleright} \mathbf{G}_{n;k}(\mathbf{u}_0)^{\wedge_{n;k}} + A_{k,1};$$

where the term

$$A_{k,1} = (\hat{\mathbf{u}}_{n;k})^{\triangleright} [r^2 \hat{L}_{n;k}(\mathbf{u}) - \mathbf{G}_{n;k}(\mathbf{u}_0)]^{\wedge_{n;k}}$$

satisfies

$$\lim_{C \downarrow 1} \limsup_{M \uparrow 1} \limsup_{n \uparrow 1} P \left\{ \max_{M \leq k \leq n} j A_{k,1} \left( \frac{(\log \log k)^{3=2}}{k^{1=2}} + \frac{(\log \log(n-k))^{3=2}}{(n-k)^{1=2}} \right)^1 > C \right\} = 0;$$

Finally, we get

$$\begin{aligned} (\hat{\mathbf{u}}_{n;k})^{\triangleright} \mathbf{G}_{n;k}(\mathbf{u}_0)^{\wedge_{n;k}} &= k \left( \hat{u}_{k,1} \right)^2 + (n-k) \left( \hat{u}_{k,2} \right)^2 \\ &\quad + n \left( \hat{u}_{k,1}^2 \quad \frac{2}{1} \hat{u}_{k,2}^2 \quad \frac{2}{2} \right) \mathbf{\Lambda} \left( \hat{u}_{k,1}^2 \quad \frac{2}{1} \hat{u}_{k,2}^2 \quad \frac{2}{2} \right)^{\triangleright}; \end{aligned}$$

where recall that  $\mathbf{\Lambda}$  is defined in Eq (C.21). By Lemma C.1.7 we have

$$n \left( \hat{u}_{k,1}^2 \quad \frac{2}{1} \hat{u}_{k,2}^2 \quad \frac{2}{2} \right) \mathbf{\Lambda} \left( \hat{u}_{k,1}^2 \quad \frac{2}{1} \hat{u}_{k,2}^2 \quad \frac{2}{2} \right)^{\triangleright} = \frac{1}{n} (Z_{n,3}; Z_{n,4}) \mathbf{\Lambda} (Z_{n,3}; Z_{n,4})^{\triangleright} + A_{k,2}; \quad (\text{B.3})$$

and

$$\lim_{C \downarrow 1} \limsup_{M \uparrow 1} \limsup_{n \uparrow 1} P \left\{ n^{1-2} \max_{M \leq k \leq n} j A_{k,2j} > C \right\} = 0;$$

Next we observe

$$\hat{L}_{n;n}(\mathbf{u}_0) = \hat{L}_{n;k}(\mathbf{u}_0) \quad \text{and} \quad \hat{L}_{n;n}(\hat{\mathbf{u}}_{n;n}) = \hat{L}_{n;k}(\hat{\mathbf{u}}_{n;n})$$

Using again Lemma C.1.7 with Eqs (B.1) and (B.3) we conclude

$$\begin{aligned} & 2[\hat{L}_{n;k}(\hat{\mathbf{u}}_{n;k}) - \hat{L}_{n;n}(\hat{\mathbf{u}}_{n;n})] & (B.4) \\ & = 2[\hat{L}_{n;k}(\hat{\mathbf{u}}_{n;k}) - \hat{L}_{n;k}(\hat{\mathbf{u}}_{n;n})] \\ & = n \binom{n}{k}^{-2} k \binom{k}{k;1}^{-2} (n-k) \binom{n-k}{k;2}^{-2} + A_{k;3} \\ & = \bar{n} \left( \sum_{i=1}^n i;1 \right)^{-2} \bar{k} \left( \sum_{i=1}^k i;1 \right)^{-2} \frac{n-k}{n-k} \left( \sum_{i=k+1}^n i;1 \right)^{-2} + A_{k;4}; \end{aligned}$$

and

$$\lim_{C \downarrow 1} \limsup_{M \uparrow 1} \limsup_{n \uparrow 1} P \left\{ \max_{M \leq k \leq n} j A_{k,j} \left( \frac{(\log \log k)^{3-2}}{k^{1-2}} + \frac{(\log \log(n-k))^{3-2}}{(n-k)^{1-2}} \right)^1 > C \right\} = 0; \quad (B.5)$$

$j = 3$  and  $4$ . It is easy to check that

$$\frac{1}{k} \left( \sum_{i=1}^k i;1 \right)^2 + \frac{1}{n-k} \left( \sum_{i=k+1}^n i;1 \right)^2 = \frac{1}{n} \left( \sum_{i=1}^n i;1 \right)^2 = \frac{n}{k(n-k)} \left( \sum_{i=1}^k i;1 + \frac{k}{n} \sum_{j=1}^n j;1 \right)^2;$$

We study the maximum of  $T_n$  on several subsets. Let  $r(n) = (\log n)^4$ ,  $q(n) = n - (\log n)^4$ ,

and  $M < 1$ , and define

$$\begin{aligned}
T_{n,1} &= \max_{1 \leq k \leq M} 2[\hat{L}_{n;k}(\mathcal{U}_{n;k}) - \hat{L}_{n;n}(\mathcal{U}_{n;n})]; \\
T_{n,2} &= \max_{M \leq k \leq r(n)} 2[\hat{L}_{n;k}(\mathcal{U}_{n;k}) - \hat{L}_{n;n}(\mathcal{U}_{n;n})]; \\
T_{n,3} &= \max_{r(n) \leq k \leq q(n)} 2[\hat{L}_{n;k}(\mathcal{U}_{n;k}) - \hat{L}_{n;n}(\mathcal{U}_{n;n})]; \\
T_{n,4} &= \max_{q(n) \leq k \leq n} 2[\hat{L}_{n;k}(\mathcal{U}_{n;k}) - \hat{L}_{n;n}(\mathcal{U}_{n;n})]; \\
T_{n,5} &= \max_{n \leq q(n) \leq k \leq n} 2[\hat{L}_{n;k}(\mathcal{U}_{n;k}) - \hat{L}_{n;n}(\mathcal{U}_{n;n})]; \\
T_{n,6} &= \max_{n \leq r(n) \leq k \leq n} 2[\hat{L}_{n;k}(\mathcal{U}_{n;k}) - \hat{L}_{n;n}(\mathcal{U}_{n;n})]
\end{aligned}$$

and

$$T_{n,7} = \max_{n \leq M \leq k \leq n} 2[\hat{L}_{n;k}(\mathcal{U}_{n;k}) - \hat{L}_{n;n}(\mathcal{U}_{n;n})].$$

First we note that, for every fixed  $M$ , it holds that  $T_{n,1} = O_P(1)$ , and  $T_{n,7} = O_P(1)$ , so we have

$$a(\log n) T_{n,1} \leq b(\log n) \mathbb{P} \leq 1 \quad \text{and} \quad a(\log n) T_{n,7} \leq b(\log n) \mathbb{P} \leq 1 :$$

Using Lemma C.1.7 we get

$$T_{n,2} = O_P(1) + \max_{M \leq k \leq r(n)} \frac{n}{k(n-k)} \left( \sum_{i=1}^k \mathbb{1}_{i,1} - \frac{k}{n} \sum_{j=1}^n \mathbb{1}_{j,1} \right)^2 :$$

We showed in the proof of Lemma C.1.6 (see also Horváth and Trapani, 2022) that  $\mathbb{1}_{i,1}$  is a decomposable Bernoulli shift; hence, we can use the results in Aue et al. (2014) to construct independent Wiener processes  $\mathbb{1}W_{n,1}(t); 0 \leq t \leq n=2g$  and  $\mathbb{1}W_{n,2}(t); 0 \leq t \leq n=2g$  such that

$$\max_{1 \leq k \leq n=2} \frac{1}{k} \left| \sum_{i=1}^k \mathbb{1}_{i,1} - \mathbb{1}W_{n,1}(k) \right| = O_P(1); \tag{B.6}$$

and

$$\max_{1 \leq k \leq n=2} \frac{1}{(n-k)} \left| \sum_{i=k+1}^n \mathbb{1}_{i,1} - \mathbb{1}W_{n,2}(n-k) \right| = O_P(1); \tag{B.7}$$

with some  $\epsilon < 1/2$ . The approximations in Eqs (B.6) and (B.7) with the law of the iterated logarithm for the Wiener process yield

$$\begin{aligned} T_{n,2} &= O_P(1) + \max_{k \leq r(n)} \frac{n}{k(n-k)} \left( W_{n,1}(k) - \frac{k}{n} (W_{n,1}(n-2) + W_{n,2}(n-2)) \right)^2 \\ &= O_P(\log \log \log n) : \end{aligned}$$

Hence it follows that

$$a(\log n) T_{n,2} \stackrel{P}{\sim} b(\log n) : 1 :$$

and similarly  $a(\log n) T_{n,6} \stackrel{P}{\sim} b(\log n) : 1$ . The approximations in Eqs (B.6) and (B.7) yield on the interval  $q(n) \leq k \leq n - q(n)$

$$T_{n,4} = O_P(1) + \max_{q(n) \leq k \leq n - q(n)} \frac{n}{k(n-k)} \left( W_{n,1}(k) - \frac{k}{n} (W_{n,1}(n-2) + W_{n,2}(n-2)) \right)^2 :$$

The computation of the covariance function shows that

$$B_n(t) = \begin{cases} n^{-1/2} (W_{n,1}(Nt) - t(W_{n,1}(n-2) + W_{n,2}(n-2))) & 0 \leq t \leq 1/2 \\ n^{-1/2} (W_{n,1}(n(1-t)) + (1-t)(W_{n,1}(n-2) + W_{n,2}(n-2))) & 1/2 \leq t \leq 1 \end{cases} ;$$

is a Brownian bridge; further, note that the distribution of  $B_n(t)$  does not depend on  $n$ , so that we can write  $fB_n(t) : 0 \leq t \leq 1 \stackrel{D}{=} fB(t) : 0 \leq t \leq 1$ . Thus, applying the Darling–Erdős theorem (cf. Theorem A.4.2 in [Csörgő and Horváth, 1997](#)), it can be shown that

$$\max_{q(n) \leq k \leq n - q(n)} \frac{n}{k(n-k)} \left( W_{n,1}(k) - \frac{k}{n} (W_{n,1}(n-2) + W_{n,2}(n-2)) \right)^2 = O_P(\log \log \log n) ;$$

and therefore

$$a(\log n) T_{n,4} \stackrel{P}{\sim} b(\log n) : 1 :$$

Using again Eqs (B.6) and (B.7) we get  $T_{n,3} = O_P((\log n)^{4(1-\epsilon)}) + T_{n,3}$ , and  $T_{n,5} = O_P((\log n)^{4(1-\epsilon)}) + T_{n,5}$ , with

$$T_{n,3} = \max_{r(n) \leq k \leq q(n)} \frac{n}{k(n-k)} \left( W_{n,1}(k) - \frac{k}{n} (W_{n,1}(n-2) + W_{n,2}(n-2)) \right)^2$$

and

$$T_{n,5} = \max_{n - q(n) \leq k \leq r(n)} \frac{n}{k(n-k)} \left( W_{n,1}(n) + W_{n,1}(k) + \frac{n-k}{n} (W_{n,1}(n-2) + W_{n,2}(n-2)) \right)^2 :$$

Thus

$$\lim_{n \uparrow} Pfa(\log n) T_n^{1=2} \quad x+b(\log n)g = \lim_{n \uparrow} Pfa(\log n) \max(T_{n,3}; T_{n,5})^{1=2} \quad x+b(\log n)g+o(1) :$$

Using Theorem 2 in [Darling and Erdős \(1956\)](#),<sup>1</sup> we finally obtain

$$\lim_{n \uparrow} Pfa(\log n) \max(T_{n,3}; T_{n,5})^{1=2} \quad x + b(\log n)g = \exp(-2e^{-x}); \quad 1 < x < 1 ;$$

computing the proof of Theorem [6.2.1](#) when  $E \log j + o(1) < 0$ .

We now turn to the nonstationary case,  $E \log j + o(1) > 0$ . In this case, however, we can repeat the previous proof for the stationary case, since we established exactly the same results in the explosive case. We even obtained better approximations for the likelihood function and its first and second derivatives in Lemmas [C.1.8–C.1.10](#). We also proved approximation with partial sums of independent and identically distributed random variables for the maximum likelihood estimators in the explosive case. Hence the proof used when Eq [\(C.3\)](#) was assumed can be repeated verbatim.

### Proof of Theorem [6.2.2](#)

We report the details of the proof under the case where  $y_i$  changes from a stationary regime to another stationary regime after the changepoint  $k$ . Exactly the same arguments can be used to study the other cases, e.g. changes from stationarity to explosivity, from explosivity to stationary and from an explosive to another explosive regime.

After the change at  $k$ , the recursion restarts from  $y_k$  and converges to a stationary sequence, denoted as  $z_i$ . Following the proof of Lemmas [C.1.2](#) and [C.1.3](#) we get

$$\frac{1}{n} \hat{L}_{n;n}(\mathbf{u}) \stackrel{P}{\rightarrow} \frac{1}{2} E \left( \log(u_2 y_0^2 + u_3) + \frac{(y_1 - u_1 y_0)^2}{u_2 y_0^2 + u_3} \right) \quad \frac{1}{2} E \left( \log(u_2 z_0^2 + u_3) + \frac{(z_1 - u_1 z_0)^2}{u_2 z_0^2 + u_3} \right)$$

where  $\mathbf{u} = (u_1; u_2; u_3)$  uniformly on  $\mathbf{U}$ . Standard arguments give that  $(\hat{y}_n; \hat{y}_{n,1}^2; \hat{y}_{n,2}^2) \stackrel{P}{\rightarrow} (y; y_1^2; y_2^2)$ . Now

$$T_n(k) = 2L_{n;k}(\hat{y}_k; \hat{y}_{k,1}^2; \hat{y}_{k,2}^2) - 2L_{n;n}(\hat{y}_n; \hat{y}_{n,1}^2; \hat{y}_{n,2}^2);$$

According to the proof of Theorem [6.2.1](#), it holds that

$$\left| L_{n;k}(\hat{y}_k; \hat{y}_{k,1}^2; \hat{y}_{k,2}^2) - L_{n;k}(y; y_1^2; y_2^2) \right| = O_P(1);$$

<sup>1</sup>See the Supplement for further details.

Next we write

$$\begin{aligned}
& 2 \left( L_{n;k} \left( \hat{\gamma}_1^2, \hat{\gamma}_2^2 \right) - L_{n;n} \left( \hat{\gamma}_1^2, \hat{\gamma}_2^2 \right) \right) \\
&= \sum_{i=1}^k \left\{ \left( \log \left( \hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,2}^2 \right) + \frac{(\gamma_i - \hat{\gamma}_n \gamma_{i-1})^2}{\hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,2}^2} \right) \left( \log \left( \frac{\gamma_i^2}{\gamma_{i-1}^2} + \frac{\gamma_i - \gamma_{i-1}}{2} \right) + \frac{(\gamma_i - \gamma_{i-1})^2}{\frac{\gamma_i^2}{\gamma_{i-1}^2} + \frac{\gamma_i - \gamma_{i-1}}{2}} \right) \right\} \\
&+ \sum_{i=k+1}^n \left\{ \left( \log \left( \hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,2}^2 \right) + \frac{(\gamma_i - \hat{\gamma}_n \gamma_{i-1})^2}{\hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,2}^2} \right) \left( \log \left( \frac{\gamma_i^2}{\gamma_{i-1}^2} + \frac{\gamma_i - \gamma_{i-1}}{2} \right) + \frac{(\gamma_i - \gamma_{i-1})^2}{\frac{\gamma_i^2}{\gamma_{i-1}^2} + \frac{\gamma_i - \gamma_{i-1}}{2}} \right) \right\};
\end{aligned}$$

Following the proof of Lemma C.1.2 we get

$$\begin{aligned}
& \sum_{i=1}^k \left| \left( \log \left( \hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,2}^2 \right) + \frac{(\gamma_i - \hat{\gamma}_n \gamma_{i-1})^2}{\hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,2}^2} \right) \left( \log \left( \hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,2}^2 \right) + \frac{(\gamma_i - \hat{\gamma}_n \gamma_{i-1})^2}{\hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,1}^2} \right) \right| \\
&= O_P(1)
\end{aligned}$$

and

$$\begin{aligned}
& \sum_{i=k+1}^n \left| \left( \log \left( \hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,2}^2 \right) + \frac{(\gamma_i - \hat{\gamma}_n \gamma_{i-1})^2}{\hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,2}^2} \right) \left( \log \left( \hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,2}^2 \right) + \frac{(\gamma_i - \hat{\gamma}_n \gamma_{i-1})^2}{\hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,2}^2} \right) \right| \\
&= O_P(1);
\end{aligned}$$

By standard arguments, it can be shown that  $\hat{\gamma}_{n,1}^2 = \frac{\gamma_1^2}{n} + o_P(1)$ ,  $\hat{\gamma}_{n,2}^2 = \frac{\gamma_2^2}{n} + o_P(1)$  and, under the alternative

$$\hat{\gamma}_n = \frac{k}{n} + \frac{n-k}{n} + o_P(1);$$

Then, exploiting the continuity of the likelihood function we get

$$\begin{aligned}
& \sum_{i=1}^k \left\{ \left( \log \left( \hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,2}^2 \right) + \frac{(\gamma_i - \hat{\gamma}_n \gamma_{i-1})^2}{\hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,1}^2} \right) \left( \log \left( \frac{\gamma_i^2}{\gamma_{i-1}^2} + \frac{\gamma_i - \gamma_{i-1}}{2} \right) + \frac{(\gamma_i - \gamma_{i-1})^2}{\frac{\gamma_i^2}{\gamma_{i-1}^2} + \frac{\gamma_i - \gamma_{i-1}}{2}} \right) \right\} \\
&+ \sum_{i=k+1}^n \left\{ \left( \log \left( \hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,2}^2 \right) + \frac{(\gamma_i - \hat{\gamma}_n \gamma_{i-1})^2}{\hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,2}^2} \right) \left( \log \left( \frac{\gamma_i^2}{\gamma_{i-1}^2} + \frac{\gamma_i - \gamma_{i-1}}{2} \right) + \frac{(\gamma_i - \gamma_{i-1})^2}{\frac{\gamma_i^2}{\gamma_{i-1}^2} + \frac{\gamma_i - \gamma_{i-1}}{2}} \right) \right\} \\
&= \sum_{i=1}^k \left\{ \left( \frac{(\gamma_i - \hat{\gamma}_n \gamma_{i-1})^2}{\hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,1}^2} \right) \left( \frac{(\gamma_i - \gamma_{i-1})^2}{\frac{\gamma_i^2}{\gamma_{i-1}^2} + \frac{\gamma_i - \gamma_{i-1}}{2}} \right) \right\} + \sum_{i=k+1}^n \left\{ \left( \frac{(\gamma_i - \hat{\gamma}_n \gamma_{i-1})^2}{\hat{\gamma}_{n,1}^2 \gamma_{i-1}^2 + \hat{\gamma}_{n,2}^2} \right) \left( \frac{(\gamma_i - \gamma_{i-1})^2}{\frac{\gamma_i^2}{\gamma_{i-1}^2} + \frac{\gamma_i - \gamma_{i-1}}{2}} \right) \right\};
\end{aligned}$$

and after some algebra, it can be shown that this equals

$$E \left( \frac{\gamma_0^2}{\frac{\gamma_0^2}{n} + \frac{\gamma_1^2}{n}} \right) \left( \frac{\gamma_1^2}{\gamma_0^2} \right)^2 \left( \frac{k(n-k)}{n} \right) + o_P(k) + o_P(n-k);$$

and, under Eq (6.15), it now follows that, as  $n \rightarrow \infty$

$$a(\log n) T_n^{1-2} = b(\log n) + o_p(1);$$

which proves the consistency of the test.