Methods in Functional Data Analysis: Forecast Evaluation, Robust Serial Dependence Measures, and a Spatial Factor Copula Model

by

Chi-Kuang Yeh

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, 2023

© Chi-Kuang Yeh 2023

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:	Linglong Kong Professor Department of Mathematical and Statistical Sciences University of Alberta
Supervisor(s):	Gregory Rice Associate Professor Department of Statistics and Actuarial Science
	Joel A. Dubin Professor Department of Statistics and Actuarial Science School of Public Health Sciences
Internal Member:	Tony S. Wirjanto Professor Department of Statistics and Actuarial Science School of Accounting and Finance
	Samuel W.K. Wong Associate Professor Department of Statistics and Actuarial Science

Internal-External Member: Dinghai Xu
Professor
Department of Economics

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 chapters of this thesis describe and contain Chi-Kuang Yeh's work, under the supervision of Professors Gregory Rice and Joel A. Dubin, that has been published in the following peer-reviewed journals.

Chapter 2: Yeh, C.-K, Rice, G and Dubin, J.A. (2022). Evaluating Real-time probabilistic Forecasts with Application to National Basketball Association Outcome Prediction. *The American Statistician*, 76, 214–223.

Chapter 3: Yeh, C.-K, Rice, G and Dubin, J.A. (2023). Functional Spherical Autocorrelation: a Robust Estimate of the Autocorrelation of a Functional Time Series. *Electronic Journal of Statistics*, 17, 650–687.

The entirety of the related code is written in `R` and is solely authored by Chi-Kuang Yeh for these manuscripts. The code will be accessible at https://github.com/chikuang.

Abstract

With advancements in technology, new types of data have become available, including functional data, which observations in the form of functions or curves rather than scalar or vector-valued quantities. This emerging area presents unique challenges in handling intrinsically infinite-dimensional objects. In this thesis, we primarily focus on three problems, each of which has a distinct flavour of functional data analysis.

In Chapter 1, we provide an overview of the foundational concepts and methodologies that will serve as a basis for the subsequent chapters. This includes an exploration of topics such as functional data analysis, functional time series analysis, probabilistic forecasts, copula modelling, and robust methods that will be in later chapters. Additionally, we conclude this chapter by presenting a comprehensive list of the main contributions made by this thesis.

In Chapter 2, motivated by the goal of evaluating real-time forecasts of home team win probabilities in the National Basketball Association, we develop new tools for measuring the quality of continuously updated probabilistic forecasts. This includes introducing calibration surface plots, and simple graphical summaries of them, to evaluate at a glance whether a given continuously updated probability forecasting method is well-calibrated, as well as developing statistical tests and graphical tools to evaluate the skill, or relative performance, of two competing continuously updated forecasting methods. These tools are demonstrated in an application on evaluating the continuously updated forecasts published by United States-based multinational sports network ESPN on its principle webpage espn.com. This application provides statistical evidence that the forecasts published there are well-calibrated, and exhibit improved skill over several naïve models, but do not show significantly improved skill over simple logistic regression models based solely on a measurement of each teams' relative strength, and the evolving score difference throughout the game.

In Chapter 3, we propose a new autocorrelation measure for functional time series that we term "spherical autocorrelation." It is based on measuring the average angle between lagged pairs of series after having been projected onto a unit sphere. This new measure enjoys at least two complimentary advantages compared to existing autocorrelation measures

for functional data, since it both 1) describes a notion of "sign" or "direction" of serial dependence in the series, and 2) is more robust to outliers. The asymptotic properties of estimators of the spherical autocorrelation are established, and used to construct confidence intervals and portmanteau white noise tests. These confidence intervals and tests are shown to be effective in simulation experiments, and in applications model selection for daily electricity price curves, and in measuring volatility in densely observed asset price data.

In Chapter 4, we propose a new model for spatial functional data that departs from the commonly adopted assumption of normality of the errors. Instead, we assume the existence of a common process that equally affects the measurements of the data at all locations at each time point. By using general copulas, our model can accommodate heavy tails and tail asymmetry, which the existing methods may suffer from. We then derive the closed-form expression of the likelihood function when the tail dependence is generated by an exponential distribution. The simulation studies show that the parameter estimates of the proposed method accurately capture the spatial and temporal dependence when the model is correctly specified. In the case where the model is misspecified, our method is still robust in capturing the spatial dependence and the general shape of the common mean function. We close the chapter by discussing some future works and potential extensions of the proposed model.

We conclude this thesis by presenting concise summaries of each chapter and engaging in further discussions in Chapter 5. Additionally, we also offer directions for future research in each chapter, highlighting potential applications of the proposed methods. Furthermore, we explore theoretical and computational avenues that may prove beneficial to practitioners and researchers, extending the scope of the proposed methods to encompass research, applications, and beyond.

Acknowledgements

Time flies, and I never anticipated that five years would pass so swiftly. In the midst of these years, many things have unfolded, and among them, numerous colorful memories have been etched in my heart, which I hold onto with great cheerfulness. I am thankful for the presence of all these events and individuals in my life. Below are some things and people that I would like to specifically mention.

I would like to sincerely express my gratitude to my advisors, Gregory Rice and Joel A. Dubin, without whom this thesis would not have been possible. Greg has always been there for me, providing invaluable guidance, even if he occasionally misses my emails due to his busy schedule. I often refer to him as my "hero," and words cannot adequately express my appreciation. His attitude towards knowledge, research, life, and beyond is an unforgettable attribute that greatly inspires me. Joel has made our meetings enjoyable, patiently waiting and engaging us in the enjoyable aspects of life. I cherish the wonderful time we have spent together and am feeling quite melancholic that these pleasant afternoon talks will soon come to an end.

I would also like to extend my thanks to the professors who served as my committee members: Tony S. Wirjanto, Samuel W.K. Wong, Dinghai Xu, and Linglong Kong. Your insights and contributions have significantly enriched my research and enhanced this dissertation.

Apart from Greg and Joel, there are several professors in academia who have had a significant influence on me. My master's supervisor, Julie Zhou, along with Gary MacGillivray and Junling Ma, who served as my undergraduate mentors, deserve special mention. Julie, Gary, and Junling were instrumental in guiding me towards the field of statistics. Without them, I might have ended up in other disciplines such as computer science, applied or discrete mathematics, chemistry, or something else. Peijun Sang, one of my closest friends in Waterloo, has had a profound impact on my perspective towards academia. I am grateful for his advice and unwavering support throughout the years. I enjoyed the daily walks we had before and during the pandemic, as well as the countless conversations about research ideas, life, and other enjoyable and fun topics. These interactions helped me navigate through the challenging times. Additionally, I would like to express my gratitude to Laura

L.E. Cowen, who has shared invaluable advice with me in various aspects of my life. I also extend my appreciation to Paul Marriott, Pengfei Li, Richard Cook, Changbao Wu, Lisa Gao, Mary Lesperance, Martin Lysy, Yeying Zhu, Mu Zhu, Samuel W.K. Wong, Weng Kee Wong and Qinglong Tian, for our discussions, which have helped me tremendously both in academia and beyond. I am truly thankful to each and every one of you who has played a significant role in my journey, shaping who I am today.

Special thanks to Mary Lou Dufton, Greg Preston, Helen Chen, and other kind staff in the department who have made this department feel like a warm home for me. To such an extent that there are times when I would rather stay here than go home and even consider sleeping in the office (and have actually done so, many, many times). One of my happiest moments is when I ask Mary daily questions, and she's always there, patiently listening to what I have to say. Also, I am grateful that she had never dragged me out of the office even when I was sleeping there.

Regarding the funding and scholarships, I would like to express my gratitude to my advisors, the Department of Statistics and Actuarial Science, the University of Waterloo, and the Government of Ontario for their support through Ontario Graduate Scholarship (OGS), Queen Elizabeth II Graduate Scholarship in Science and Technology (QEII-GSST), President's Graduate Scholarship (PGS), Statistics and Actuarial Science Chair's Awards, among others. Without them, this thesis would once again not have been possible.

Throughout my academic journey, I have had the privilege of meeting numerous individuals whom I deeply appreciate. I extend my heartfelt appreciation to Peijun, Peiyao, Xinyi, Zhaohan, Wenyuan, Minghui, Yunruo, Zhaoxin, Jingyue, Xiyue, Mingren, Mengxiao, Yang, Wenyu, Jason, Alex, Dingding, Yechao, Fangya, Kecheng, Hwanwoo, Nam-Hwui, Yuying, Wenling, Shiyu, Feiyu, Qiuqi, Yimiao, Trang, Shaowen, Peiwei, Hongda, Wenqin, Lin, Bei, Shimeng, Boxi, Haodi, Jingyi, Jianchu, Ting Ju, Chris, Minzee, Haoxin, Junhan, Liyuan, Li-Pang, Qihuang, Jiazhen, Zixin, Keqi, Mingwei, Jie, Ce, Xiaoxue, Xianwei, Zijia, Yuyu, Meng, Yilin, Xiaobai, Lv, Keying, Marzieh, Xiaoyu and countless other lovely people. In particular, Peijun, Xiyue, and Zhaohan, your assistance with the technical aspects has been invaluable. Without your help, I might still be struggling in the dark, attempting to solve equations and understand the concepts and proofs. I would also like to express my gratitude to my academic brothers, Yijun, Chi, Yuqian and Zhaoran, for the camaraderie

and the insightful research discussions we have shared.

I would like to express my gratitude to my office mates Nam-Hwui, Chi, Dingding, Zehao and semi-office mate Xiyue, who have made our shared workspace in M3–4122 feel like a comfortable home. There were times when I even preferred to sleep in the office rather than returning home. I would also like to thank my roommates Yifan, Wei, Jiaxin, Tingshuang, Xiru, Jie, and Qinglong, with whom I have thoroughly enjoyed living with. I am especially grateful to Qinglong for sharing numerous research ideas, valuable tips, and experiences with me. Your dedication and humility have also greatly inspired me.

Additionally, I am grateful to other individuals who have contributed significantly to my life: Shengyi, Feiran, Yixin, Yu, Shuoheng, Yonglin, Buyun, Yuchi, Douglas, and Matthew from the University of Victoria; Kexin and Dan from the University of Western Ontario; Na from Queen's University; Jia-Han from Academia Sinica; Haoxuan from Simon Fraser University; Jason Tsz Him from the University of British Columbia; Sheng, Jiannan, He, Decheng, and Shaojian from high school; Chun-Tang, Chun-Yi, Kuan-Ming, Yu-Ken, Ming-Hua, Yu-Chieh, Cheng-Hsin, Wei-Chia, and our teacher, Ms. Yu-Ching Liu, from middle school, with whom I have remained friends ever since. The support of other childhood friends is also indispensable. Furthermore, I would like to particularly extend my heartfelt gratitude to Xinqi for your unwavering support and understanding throughout the years. Thank you so, so, so much.

Lastly, I would like to thank my family, including my sisters Ya-Sung and Ya-Hsin, my father Chih-Chiang, my brother-in-law Yuan-Ming, and my late mother Ming-Hui. I also want to extend my gratitude to other members of my family who have understood my situation and supported me throughout this journey, especially my late grandparents. From the bottom of my heart, I must reiterate that I cannot thank you all enough.

Dedication

To my beloved family, cherished loved ones, dear friends, and those who courageously pursue their dreams.

天道酬勤、自强不息。

Table of Contents

Examining Committee ii

Author's Declaration iv

Abstract vi

Acknowledgements viii

Dedication xi

List of Figures xvi

List of Tables xxi

1 Introduction 1

1.1 Functional Data Analysis 2

1.2 Principal Component Analysis 4

1.2.1 Principal component analysis for multivariate data 5

1.2.2 Principal component analysis for functional data 6

1.3 Functional Time Series 8

1.3.1 Time series 8

1.3.2 Linear processes . . . 13
1.3.3 Nonlinear process . . . 16
1.3.4 Functional time series . . . 17
1.3.5 Functional time series models . . . 20
1.4 Spatial Statistics . . . 22
1.5 Robust Methods in Functional Data . . . 24
1.5.1 Univariate and multivariate centring . . . 25
1.5.2 Centring for functional data . . . 26
1.6 Probabilistic Forecasts . . . 26
1.7 Copula . . . 27
1.7.1 Factor copula . . . 30
1.8 Outline of the Thesis . . . 31
1.9 Contributions . . . 32

2 Evaluating Real-time Probabilistic Forecasts with Application to National Basketball Association Outcome Prediction 34

2.1 Introduction . . . 34
2.2 Motivating Data and Competing Models . . . 38
2.2.1 Benchmark models for predicting NBA game outcomes . . . 40
2.3 Evaluating Calibration of Continuous Forecasts . . . 43
2.4 Evaluating Skill of Continuously Updated Forecasts . . . 47
2.4.1 Pointwise confidence intervals measuring the skill difference between competing methods . . . 49
2.4.2 Functional tests to measure skill aggregated across t . . . 52
2.5 Simulation Studies . . . 54

2.5.1 Basketball game simulation . . . 54
2.5.2 Models . . . 55
2.5.3 Results . . . 56
2.6 Evaluating the Skill of ESPN Forecasts . . . 60
2.7 Discussion . . . 62

3 Functional Spherical Autocorrelation: a Robust Estimate of the Autocorrelation of a Functional Time Series 64

3.1 Introduction . . . 64
3.2 Problem Statement and Methodology . . . 67
3.2.1 Definition of spherical autocorrelation . . . 68
3.2.2 Inference for ρ_h assuming μ is known . . . 69
3.2.3 Inference for ρ_h assuming μ is unknown . . . 72
3.3 Implementation and Simulation Studies . . . 74
3.3.1 White noise processes . . . 75
3.3.2 Functional time series models . . . 76
3.4 Applications . . . 81
3.4.1 Spanish electricity data . . . 84
3.4.2 Densely observed intraday asset price data . . . 85
3.5 Conclusion . . . 90

4 Factor Copula Models for Spatially Correlated Functional Data 91

4.1 Introduction . . . 91
4.2 Model . . . 94
4.2.1 Copula models for spatial data . . . 94
4.2.2 Factor copula models for spatial functional data . . . 95

4.3 Estimation 99
4.4 Simulation Studies 103
4.5 Concluding Remarks and Future works 107

5 Concluding Remarks and Future work **112**
5.1 Future Work 116

References **118**

APPENDICES **135**

A Technical Details of Chapter 2 **136**
A.1 Definition of $p_i(t)$ and Proof of Theorem 2.4.1 136
A.2 Selection of the Number of Bins using CORP Binning Method 141

B Technical Details of Chapter 3 **143**
B.1 Technical Details and Proofs 143
B.1.1 Proof of Theorem 1 143
B.1.2 Proof of Theorem 3.2.2 144
B.1.3 Proof of Theorem 3.2.3 152
B.2 Additional Simulation Results 158
B.2.1 Accuracy of Estimator $\|\hat{C}_P\|_2$ 158
B.3 Details of the Estimation of the FSAR model 160

C Technical Details of Chapter 4 **162**
C.1 Closed-form Expression for $f_{X|N}(x \mid \overline{N})$ 162
C.2 Derivation of the Distribution and Density functions 166

Glossary **168**

List of Figures

1.1 **Left**: The growth rate of weekly closing returns of the S&P 500 from January 2003 to September 2012. The large spikes centre around week 300 suggest the presence of vilatility clustering in the data. **Right**: The autocorrelation function computed up to lag 30 with 95% strong WN band. The spikes exceeding the strong WN noise bands suggests the presence of serial dependence, which needs to be addressed. 12

1.2 The concentrations of particular matter with an aerodynamic diameter of less than 10 micrometers (PM10) in ambient air in Graz-Mitte Austria from October 1, 2010 to October 11, 2010. The curves are evenly segmented into 10 sub-intervals, which represents the change in PM10 of each day. 18

1.3 **Left**: 400 simulated Brownian bridges, which are i.i.d. **Right**: Functional autocorrelation function with the 95% error bound under the assumption of data are i.i.d. The fACF at all lags are within the 95% bound indicates the independence of the simulated curves, which agrees with the data generating process. 20

2.1 Real-time probabilistic forecasts as a function of the in-game time published on `espn.com/nba` (ESPN, 2020) from a November 8, 2018 game in which the Los Angeles Lakers hosted the Minnesota Timberwolves. The black points are the points at which ESPN updated the home team winning probability forecasts due to events occurring in the game, and the red line is a piecewise constant interpolation between these forecasts to produce a real-time probabilistic forecast curve (function). 36

2.2 **Left**: the pseudo R^2 of the logistic regression model **PgRSScD**, which uses the covariates pre-game relative strength and score difference, as a function of the game time. **Right**: the variable importance of each covariate as it contributes to the pseudo R^2. 43

2.3 **Left**: a calibration plot of the ESPN forecasts at time point $t = 0.5$. **Right**: an upper calibration surface along with reference plane $f(t, p) = p$ for the continuously updated ESPN forecasts obtained by interpolating the upper bounds of each confidence interval of the calibration plots across $t \in [0, 1]$. 46

2.4 Plots of $U^{min}_{0.95}(t)$ and $L^{max}_{0.95}(t)$ against t based on $M = 10$ bins: Methods (a) historic home team win probability (**HomeWP**); (b) GLM using Score difference without intercept (**ScDnoInt**); (c) **ESPN** forecasts; (d) GLM using Pre-game relative strength and Score difference (**PgRSScD**). 48

2.5 Representative plots of $\hat{\Delta}_{1000}(t)$ with 95% confidence intervals and approximate p-values for tests of H_0 for (a) **OraBM1** versus textbfOraBM2; (b) **OraOU1** and **OraOU2**. 58

2.6 Representative plots of $\hat{\Delta}_{1000}(t)$ derived from simulated data with 95% confidence intervals and approximate p-values for tests of H_0 for (a) **PgRSScD** versus **PgRS** (b) **PgRSScD** versus **ScD** (c) **PgRSScD** versus **LS** (d) **PgRSScD** versus **PgRSLS**. 59

2.7 Plots of $\hat{\Delta}_{1213}(t)$ based on the 2018–19 season forecasts with 95% confidence intervals and approximate P-values for tests of H_0 for (a) **ESPN** versus **PgRS**; (b) **ESPN** versus **ScD**; (c) **ESPN** versus **PgRSLS**; (d) **ESPN** versus **PgRSScD**. 61

3.1 **Top**: 50 i.i.d. curves of each strong noise process with the spatial median estimate from Gervini (2008) (in black); **Middle**: The leading 7 basis functions in the Karhunen-Loéve expansion of each process; **Bottom**: The estimated fSACF of each series for lags $h \in \{1, \ldots, 30\}$ with 95% confidence bounds as in (3.9) (in blue). 77

3.2 1000 curves simulated from FAR$(1, S)$ models with different scaling parameter S. **Top**: 50 i.i.d. curves simulated based on each model; **Middle**: The fACF of each of the processes with 95% confidence bound (in blue) **Bottom**: The fSACF of each of the processes with 95% confidence bounds (3.9) (in blue). 80

3.3 Empirical rejection rate of the portmanteau test under $\mathcal{H}'_{0,H}$ of data generated according to the FAR(1,S) model based on 1000 independent simulations. **Left**: H = 1; **Right**: H = 10. The nominal level α is set to be 0.05. 82

3.4 1000 simulated curves from FAR$(2, S_1, S_2)$ models with different scaling parameters S_1 and S_2 with the error process being a Brownian bridge. **Left**: The fSACF of observations simulated from the FAR$(2, S_1, S_2)$ models with different coefficients of dependence S_1 and S_2. **Middle**: The fSACF of the residuals r_i from an FAR(2) model. **Right**: The fSACF of the residuals r_i from an FAR(1) model, which in this case is evidently under specified. . . . 83

3.5 **Left**: Top is the hourly observed annual Spanish electricity data $\{Y_i(t), i = 1, \ldots, 365, t = 1, \ldots, 24\}$ from January 1st 2014 to December 31 sgemented by day. Bottom is the differenced curves $\{Y_i(t) - Y_{i-1}(t), t \in [0, 24]\}$; **Middle**: The fACF of $Y_i(t)$ (top) and $Y_i(t) - Y_{i-1}(t)$ (bottom), respectively; **Right**: The fSACF of $Y_i(t)$ (top) and $Y_i(t) - Y_{i-1}(t)$ (bottom), respectively. The blue lines in the middle and right panels are the 95% confidence interval under the null hypothesis $\mathcal{H}_{0,h}$, for $h = 1, \ldots, 20$. 86

3.6 **Left**: The fSACF of the residual r_i after fitting by the FSAR model; using the ; **Middle**: The fSACF of the residual r_i after fitting the data using the HU model with three FPCs; **Right**: The fSACF of the residual r_i after fitting the data using the HU model with three FPCs; **Bottom** The p-values of test statistics $\hat{Q}_{364,H}$ under $\mathcal{H}'_{0,H}$ of residual using FSAR model, residual using HU model with three and five FPCs, and the differenced series $Y_i(t) - Y_{i-1}(t)$, respectively. 87

3.7 **Top**: samples of four curves from each of the four FTS considered, constructed from the intraday light sweet crude oil futures prices (CL). **Bottom**: the empirical fSACF of each to these series along with the 95% confidence interval 3.9. 89

4.1 Boxplots for the the parameter estimates coefficients of the mean curve ξ_1, ξ_2, ξ_3, ξ_4 and ξ_5 with different sample sizes of N in 500 simulation runs when the jump process is generating from an exponential distribution with parameter. 105

4.2 This figure presents the parameter estimates for σ, θ, γ, λ and β with different sample sizes of N in 500 simulation runs when the jump process is generating from an exponential distribution. The red lines denote the true parameter values. 106

4.3 Boxplots for the the parameter estimates coefficients of the mean curve ξ_1, ξ_2, ξ_3, ξ_4 and ξ_5 with different sample sizes of N in 500 simulation runs when the jumping process is generating from **Top row**: the Pareto distribution and **Bottom row**: standard Gaussian distribution. The red lines denote the true parameter values. 109

4.4 The estimated mean function versus the true mean function with different sample sizes of N for the case the tails are simulated from: **Left**: Exponential distribution, **Right**: Pareto distribution and **Bottom**: Gaussian distribution, respectively. 110

4.5 This figure presents the parameter estimates for σ, θ, γ, λ and β with different sample sizes of N in 500 simulation runs when the jump process is generating from **Top row**: Pareto distribution and **Bottom row**: Gaussian distribution. The red lines denote the true parameter values. 111

A.1 The change in number of bins across time using the CORP method: Methods (a) logit model using Score difference without intercept (ScDnoInt); (b) **ESPN** forecasts; (c) logit model based on Pre-game relative strength and Score difference (**PgRSScD**). 141

A.2 Plots of $U_{0.95}^{min}(t)$ and $L_{0.95}^{max}(t)$ against t based on the bins obtained from CORP method: Methods (a) logit model using Score difference without intercept (**ScDnoInt**); (b) **ESPN** forecasts; (c) logit model based on Pregame relative strength and Score difference (**PgRSScD**). 142

List of Tables

2.1 Summary of the data obtained from ESPN (ESPN, 2020) from the 2017–18 and 2018–19 NBA regular seasons. Raw counts represent the total number of games that ESPN provides probability forecasts for. Selected refers to those games that do not contain errors or missing values. Processed represents the data after averaging out multiple events recorded at the same game time. 39

2.2 Descriptions of benchmark models used to forecast $\hat{p}_i(t)$, listed from more naïve, to more realistic models. 42

2.3 The number of games in which a probability forecast exceeded 0.995 or was less than 0.005 at some point for each of the ESPN, PgRSScD with a logit link, and the ScDnoInt with logit link forecasts, as well as the number and proportion among these games in which the home team won. 45

2.4 Empirical rejection rates with nominal levels of 10%, 5% and 1% for the test $H_0 : \|\Delta_N\|^2 = 0$ in 1000 independent simulations. 60

3.1 The empirical rate among 1000 independent simulations of four WN processes that the $1 - \alpha$ confidence intervals (3.6) and (3.9) do not contain the estimators $\tilde{\rho}_h$ and $\hat{\rho}_h$ for $\alpha \in \{10\%, 5\%, 1\%\}$, $N \in \{100, 250, 500, 1000, 2000\}$ and several values of h. 78

3.2 Assets used in the study spanned from Jan 02, 2014, to Dec 31, 2014. . . . 88

3.3 Value of the test statistics $\hat{Q}_{249,10}$ and associated p-values computed from (3.10) for: 1) five minute log returns $R_i(t)$, 2) squared five minute log returns $R_i^2(t)$, 3) CIDR $C_i(t)$, and 4) squared CIDR $C_i^2(t)$, for each asset considered. 89

B.1 The comparison between the variance estimate in a finite-dimensional case: 1) $Z_{\alpha/2}\|C_P\|_2^2$ is normal confidence interval with true variance; 2) $t_{\alpha/2,N-1}\|C_P\|_2^2$ is t-confidence interval with true variance with degrees of freedom $N-1$; 3) $Z_{\alpha/2}\|\hat{C}_P\|_2^2$ is normal confidence interval with estimated variance; and 4) $t_{\alpha/2,N-1}\|\hat{C}_P\|_2^2$ is t-confidence interval with estimated variance with degrees of freedom $N-1$. 159

Chapter 1

Introduction

This thesis explores topics in functional time series (FTS) analysis, retrospective forecasting evaluation, and modelling of spatial-temporal functional data (FD). Functional data analysis (FDA) and time series analysis are two prominent sub-fields within statistics. FDA, a relatively newer research area, has gained significant attention in recent decades, while time series analysis remains a classical branch that continues to undergo steady developments. In recent years, these two fields merged to form a new research area known as FTS analysis, wherein the FD objects exhibit serial correlation. Similar to time series analysis, making accurate prediction of future outcomes is a primary objective in FTS. Additionally, FTS aims to quantify the extent of serial dependence within a sequence of curves.

A significant body of research has been dedicated to finding appropriate models for generating accurate forecasts and evaluating the predictive performance of forecasting methods. However, the majority of these studies has focused on multivariate data, as exemplified by works such as Box (1980) and Gneiting et al. (2008). When predictions are naturally represented as smooth curves rather than discrete observations, the evaluation criteria become less clear. Moreover, quantifying the difference in predictive capability between two continuous curves generated by distinct models remains an unresolved issue. Remarkably, to the best of our knowledge, no prior investigation has addressed the statistical testing of predictability comparisons between competing forecasters in the context of

FD.

When it comes to statistically quantifying serial correlation between functional objects, there have been some recent developments (Kokoszka et al., 2017; Rice et al., 2020; Mestre et al., 2021). However, the existing methods in the literature are still relatively limited and fail to capture certain characteristics inherent in the data, such as the "sign" of the dependency. Unlike the classical Pearson correlation or autocorrelation in classic time series analysis, which produce summary statistics based on scalar values, FD are inherently infinite-dimensional objects. Their covariance behavior is best summarized by a covariance operator, and effectively summarizing and evaluating the magnitude of such an object poses challenges. One possible approach is to employ mathematical norms; however, these norms are non-negative, resulting in the loss of the "sign" of the covariance.

Lastly, another more recently emerging area of research involves modelling data with both spatial and temporal dependence. In the context of FDA, there has been a recent focus on spatial functional data (SFD). However, to the best of our knowledge, models capable of handling situations where the data have heavy tails with asymmetric behavior have not been investigated. Fortunately, there have been some related works in this area, but mostly in the multivariate setting. Our goal is to extend those models and develop a framework to effectively model SFD in such scenarios.

In this first chapter, we provide a concise introduction to the relevant background of this thesis, including FDA, time series analysis, FTS, probabilistic forecasts, and modelling SFD for copula models. This overview serves to outline the open problems that will be addressed in subsequent chapters.

1.1 Functional Data Analysis

As technology advances, new types of data become accessible to researchers, including FD. FD can be seen as an extension of multivariate data, where the observations are elements of a general function space. However unlike multivariate data, where observations are represented by scalars or vectors, FDA involves analyzing smooth functions or curves. In recent years, FDA has gained enormous attention due to its innovative approach in

analyzing intrinsically infinite-dimensional objects and its ability to provide additional insights compared to traditional methods, for instance by incorporating information about the rate of change.

Since the publication of the first edition of Ramsay and Silverman (2005) in 1997, numerous works and comprehensive treatments of FDA have become available. These include books such as Ferraty and Vieu (2006), Ramsay and Silverman (2007), Horváth and Kokoszka (2012) and Hsing and Eubank (2015). FD techniques have also been widely applied across various fields, including medicine (Leng and Müller, 2006; Sørensen et al., 2013), finance (Hörmann and Kokoszka, 2012; Horváth and Kokoszka, 2012), and environmental studies (Ignaccolo et al., 2014; Aue et al., 2020). A review of applications of FDA in different fields can be found in Ullah and Finch (2013). The increasing interest in FDA has led to the publication of special issues in various journals, such as the "Functional Data Analysis and Related Topics" in the *Journal of Multivariate Analysis*, "Functional and Object Data Analysis" in the *Canadian Journal of Statistics*, and "High-dimensional and Functional Data Analysis" in *Computational Statistics & Data Analysis*. Review papers on FDA can be found in Wang et al. (2016) and Koner and Staicu (2023). These developments highlight the growing prominence of FDA as a powerful tool which provides a new way to analyze complex datasets.

In multivariate analysis, the i-th unit observations $\{x_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, m\}$ are either scalar or vector on discrete-time observed points at location j. In contrast, the observations in FDA are typically viewed as stochastic processes, denoted by $\{x_i(t),\ i = 1, \ldots, n,\ t \in \mathcal{T}\}$, where $\mathcal{T}$ is a problem-specific compact interval. Without loss of generality, unless stated otherwise, we consider the function $X(t)$ to be in the L^2 space over the field of real numbers $\mathbb{R}$, and the domain $\mathcal{T}$ to be a unit interval $[0, 1]$ throughout the thesis. We denote this space by $\mathcal{H} := L^2([0, 1])$ from now on. In $\mathcal{H}$, the typical metric/norm is induced by an inner product defined by

$$\langle x, x \rangle = \|x\|_2^2 = \int_0^1 |x(t)|^2 dt < \infty.$$

We further assume that $\mathcal{H}$ is *separable*, which means that there exists a countable subset of $\mathcal{H}$ such that this subset is dense in $\mathcal{H}$. Consequently, we may express every $x \in \mathcal{H}$ as $x = \sum_{j=1}^{\infty} c_j e_j$, where $\{e_1, e_2, \cdots\}$ is a countable orthonormal basis. In addition, with this

orthonormal basis, we may express any $x \in \mathcal{H}$ as $x = \sum_{j=1}^{\infty} \langle x, e_j \rangle e_j$, and its norm may be written as

$$\|x\|^2 = \sum_{j=1}^{\infty} |\langle x, e_j \rangle|^2. \tag{1.1}$$

Equation (1.1) is known as *Parseval identity*, which provides a useful way to calculate the norm by projecting the elements onto the orthonormal basis.

Therefore, the data can be expressed by using a basis expansion as follows:

$$X_i(t) = \sum_{j=1}^{\infty} c_{ij} B_j(t) \approx \sum_{j=1}^{K} c_{ij} B_j(t), \quad i = 1, 2, \cdots, N, \quad t \in [0, 1], \tag{1.2}$$

where $B_j(t)$ denotes basis functions. The salient point here is that the function X itself is inherently infinite-dimensional, yet it can be well-approximated by finite-dimensional projections along with associated coefficients. The selection of basis functions can be tailored to suit specific requirements of practitioners or characteristics of the observations, such as their shape or other attributes. For instance, the Fourier basis is commonly preferred when the observed curves exhibit periodic behavior, while spline basis functions, particularly B-splines, are suitable for non-periodic data due to their desirable properties such as compact support (i.e., a finite number of supports for each sub-interval in the domain), which helps alleviate computational burdens. Numerous other basis systems are also available for diverse applications, including wavelet, monomial, and power bases, as discussed in Chapter 21 of Ramsay and Silverman (2005). Note that, it is essential to determine the number of basis functions, denoted as K, in the expansion as it requires some pre-specifications. Before delving into the methodology for selecting K, one might contemplate how to choose, in a sense, the "optimal" basis functions among various options for the observations $X_i(\cdot)$. In this regard, one approach is to employ a principal component analysis to guide the selection process which we will go through in the next section.

1.2 Principal Component Analysis

One of the most widely-used tools in FDA is *functional principal component analysis* (FPCA). Principal component analysis (PCA) is an adaptive descriptive tool and dimen-

sionality reduction technique commonly employed in statistics and various other disciplines. Since its original conception by Pearson (1901), PCA has gained immense popularity due to its geometric interpretation and practical utility across diverse fields; see, e.g., Wold et al. (1987), Abdi and Williams (2010), and Jolliffe and Cadima (2016) for comprehensive reviews, as well as reference books including Jolliffe (1986), Mardia et al. (1979), and Gnanadesikan (2011). Notably, when applying PCA to FD objects, numerous excellent reviews are available; e.g., Chapter 8 in Ramsay and Silverman (2005), Chapter 3 in Horváth and Kokoszka (2012), Section 11.4 in Kokoszka and Reimherr (2017), and Section 3.a in Jolliffe and Cadima (2016).

The significance of FPCA lies in its versatility across various areas of FDA, including exploratory data analysis (Hyndman and Shang, 2010), modelling (Ramsay and Dalzell, 1991; Ramsay and Silverman, 2005), forecasting (Hyndman and Ullah, 2007; Aue et al., 2015; Shang, 2017), and dimensionality reduction, to name only a few. At a higher level of understanding, FPCA involves iteratively projecting infinite-dimensional sample curves onto directions that contain the most statistical information or variability within the data. Consequently, the leading functional principal components (FPCs) can capture a substantial portion of the "information" associated with a given dataset, while the remaining components contribute only a minimal amount of information, achieving dimension reduction.

1.2.1 Principal component analysis for multivariate data

Before presenting FPCA, we briefly review PCA in the context of multivariate data. For a dataset with N observations and p features, we define a data matrix $\tilde{\mathbf{X}} \in \mathbb{R}^{N\times p}$, where the i-th column $\tilde{\mathbf{x}}_i$ corresponds to the i features of the observations. Let $\tilde{\mathbf{X}}$ have $\mathbb{E}(\tilde{\boldsymbol{X}}_i) = \mu_i$, and $\mathbb{V}\text{ar}(\tilde{\boldsymbol{X}}) = \mathcal{S}$. We can centre the dataset by setting $\boldsymbol{X}_i = \tilde{\boldsymbol{X}}_i - \mu_i$, representing the centred observations for each column. This convention does not affect the solution since covariance and correlation values are independent of shifting. Next, we seek a vector of principal component (PC) scores by performing a singular value decomposition on the centred data matrix $\boldsymbol{X}$. Let $(\cdot)^\top$ and $\text{tr}(\cdot)$ denote a transpose and a trace, respectively.

Recall that any matrix $\mathbf{M} \in \mathbb{R}^{N \times p}$ with rank r (where $r \leq \min(N, p)$) can be expressed as

$$\mathbf{M} = \mathbf{LDR}^\top,$$

where $\mathbf{L} \in \mathbb{R}^{N \times r}$ and $\mathbf{R} \in \mathbb{R}^{p \times r}$ have orthogonal columns, and $\mathbf{D} \in \mathbb{R}^{r \times r}$ is a diagonal matrix. The matrix $\mathbf{D}$ represents singular values of $\mathbf{M}$ and contains non-negative square roots of non-zero eigenvalues of matrices $\mathbf{M}^\top\mathbf{M}$ and $\mathbf{MM}^\top$. Additionally, let's assume that diagonal elements of $\mathbf{D}$ are arranged in decreasing order. We can then express $\mathcal{S}$ as

$$\mathcal{S} = \frac{\boldsymbol{X}^\top \boldsymbol{X}}{N} = \frac{\mathbf{RDL}^\top\mathbf{LDR}^\top}{N} = \frac{\mathbf{RD}^2\mathbf{R}^\top}{N},$$

where $\mathbf{D}^2$ represents a diagonal matrix with squared singular values (or eigenvalues in a decreasing order), and $\mathbf{R}$ is a matrix whose columns are eigenvectors. These eigenvectors are PCs of the data. The projection of data onto the PCs, $\mathbf{XR} = \mathbf{LDR}^\top\mathbf{R} = \mathbf{LD}$, is known as PC scores. Note that the denominator can use $N - 1$ instead of N, but the difference is often negligible, especially when the sample size N is large.

1.2.2 Principal component analysis for functional data

Let $X_i(\cdot)$ be a square integrable process with point-wise mean $\mathbb{E}X_i(t) = \mu(t),\ t \in [0, 1]$ and covariance function $c(t, s) = \mathbb{E}[(X_i(t) - \mu(t))(X_j(s) - \mu(s))],\ t, s \in [0, 1]$. By Mercer's Theorem (Mercer, 1909), $c(\cdot, \cdot)$ 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 $\phi(\cdot)$'s and λ's are the eigenfunctions and eigenvalues of $c(\cdot, \cdot)$, respectively. The eigenvalues λ_j's are arranged in a decreasing or non-increasing order (i.e. $0 \leq \cdots \leq \lambda_2 \leq \lambda_1$) and eigenfunctions $\phi_j(\cdot)$'s are orthonormal, i.e., $\langle \phi_k, \phi_\ell \rangle = 1$ when $k = \ell$, and zero otherwise.. Then we can express independent and identically distributed (i.i.d.) copies of a square integrable function $X \in [0, 1]$ with $\lambda_1 > \lambda_2 > \cdots$ as

$$X_i(t) - \mu(t) = \sum_{j=1}^{\infty} \varepsilon_{ij}\eta_j(t), \quad i = 1, \cdots, N, \tag{1.3}$$

where $\varepsilon_{ij} = \langle X_i - \mu, \eta_j \rangle$ and η_j are called a FPC score or simply a score, and the FPCs, respectively. Decomposition (1.3) is known as Karhunen-Loève expansion after Karhunen (1946) and Loève (1946). The scores are uncorrelated across j with mean $\mathbb{E}\varepsilon_{ij} = 0$, $\mathbb{V}\text{ar}(\varepsilon_{ij}) = \lambda_j^2$ and $\mathbb{C}\text{ov}(\varepsilon_j, \varepsilon_k) = 0$, for $j \neq k$.

In practice, since the orthogonal eigenfunctions $\eta_k(\cdot)$ for $k \geq 1$ form a basis for $L^2(\mathcal{T})$, the first K FPCs encompass the majority of variability in the curve. Therefore, the first K terms typically provide a solid approximation to the infinite sum using a finite number of terms. As a result, we may express (1.3) by truncating it up to the first K terms

$$X_i^{(K)}(t) - \bar{X}_N(t) = \sum_{j=1}^{K} \hat{\varepsilon}_{ij} \hat{\eta}_j(t), \quad j = 1, \cdots, N. \tag{1.4}$$

The estimated FPCs are unique up to sign, so we will need to assume that the sign has been taken into account,

$$\hat{\eta}_j \coloneqq \text{sgn}\left(\langle \hat{\eta}_j, \eta_j \rangle\right) \hat{\eta}_j,$$

where the $\text{sgn}(\cdot)$ is the sign function that returns the sign of a real number. Note that $\text{sgn}\left(\langle \hat{\eta}_j, \eta_j \rangle\right)$ cannot be computed from the data. Hence, we will assume that the sign is aligned between η_j's and $\hat{\eta}_j$'s and omit writing it out from now on. The estimates for the FPCs and the scores are obtained by solving the eigenfunctions

$$\int_0^1 \hat{c}(t,s)\hat{\eta}_j(s)ds = \hat{\lambda}_j \hat{\eta}_j(t), \quad j = 1, \ldots, N, \quad , t \in [0,1],$$

where $\hat{c}(t,s) = N^{-1} \sum_{i=1}^{N} (X_i(t) - \hat{\mu}(t))(X_i(s) - \hat{\mu}(s))$, $t, s \in [0,1]$ and $\hat{\mu}(t) = N^{-1} \sum_{i=1}^{N} X_i(t)$ are sample covariance function and sample mean, respectively. The scalar K in (1.4) may be chosen by using a cumulative percentage of total variance (CPV) criterion since the total variability explained by the first K empirical FPCs is

$$\text{CPV}(K) = \frac{\sum_{\ell=1}^{K} \hat{\lambda}_\ell}{\sum_{\ell=1}^{N} \hat{\lambda}_\ell}, \tag{1.5}$$

where $\hat{\lambda}$'s are empirical FPCs. There are other selection methods available in the literature, including information criteria (Shibata, 1981; Yao, 2007), resampling methods (Hall and Vial, 2006; Bathia et al., 2010) and cross-validation (Leng and Müller, 2006).

1.3 Functional Time Series

Time series is a classic and distinct area in statistics and econometrics where measurements of random responses are recorded sequentially across over time. One may think of a time series as a stochastic process, which is a collection of random variables, $\{X(t)\}_{t\in\mathbb{Z}}$, indexed by time t. The unique structure and practical utility of time series analysis have led to its widespread application in various domains, such as hydrology (Salas et al., 1980; Lane, 2002), meteorology (Duchon and Hale, 2012), engineering (Wiener, 1949), and many others.

Numerous comprehensive references, including books, have been published on the topic of time series analysis. Some notable examples include Brockwell and Davis (1991), Cryer and Chan (2008), Shumway and Stoffer (2017) and Hamilton (1994). Additionally, there also exists a scholarly reviewed journal specifically focused on time series analysis, aptly named the *Journal of Time Series*, along with closely related journals such as the *Journal of Forecasting* and the *International Journal of Forecasting*.

In this section, we aim to provide a concise overview of FTS analysis, exploring fundamental concepts and notions that will serve as a cornerstone for this thesis.

1.3.1 Time series

Prior to delving into the topic of FTS, it is necessary to outline some fundamental concepts rooted in classical time series analysis, hereafter referred to as time series for brevity. The principal objective of time series analysis is to provide a quantitative description of a random mechanism inherent in the underlying process that are observed over time. Typically, this objective is achieved by leveraging the available data to draw inferences within the observed timeframe. Making accurate predictions, which we refer to as forecasts in the time series context, beyond the observed timeframe is also of interest. To achieve such goals, it requires the assumption that the underlying structure of the data generating process remains somewhat stable throughout the entire domain, in which we require the concept of *stationarity*. This assumption plays a fundamental role in the analysis, as it enables us to analyze the sampled series in the presence of a single realization of the series.

Furthermore, without the stationarity assumption, it becomes difficult to forecast future values as anything becomes possible. Usually two types of stationarity are considered in time series analysis, namely strict stationarity and the weak stationarity.

Definition 1.3.1 (Stationary) *A time series $\{X_t\}_{t\in\mathbb{N}}$ is (weakly) stationary if*

1. $\mathbb{E}X_t = \mu$ *for all* $t \in \mathbb{N}$, *and*
2. $\mathbb{C}\mathrm{ov}(X_t, X_s) = \mathbb{C}\mathrm{ov}(X_0, X_{|t-s|})$.

That is, stationarity means that the mean function $\mathbb{E}X_t$ is constant and independent of time, and the covariance function

$$\gamma_X(t,s) \coloneqq \gamma(t,s) = \mathbb{C}\mathrm{ov}(X_t, X_s) = \mathbb{E}[(X(t)-\mu(t))(X(s)-\mu(s))], \tag{1.6}$$

only depends on the difference $|t-s|$. When the covariance function acts on the same series at two different time points t and s, it is known as autocovariance, while the cross-covariance refer to the two time points between two distinct series. The subscript is an indication of the series X, which may be dropped when the series that the autocovariance function acts on is clear. Note that the autocovariance function is symmetric in time, i.e. $\gamma(s,t) = \gamma(t,s)$. There is another strong notion of stationarity that is commonly used, namely the strict stationarity.

Definition 1.3.2 (Strict Stationary) *A time series $\{x_t\}_{t\in\mathbb{N}}$ is strictly stationary if its finite dimensional distributions are shift invariant. Namely, if $i_1, \ldots, i_d \in \mathbb{Z}$ and $h \in \mathbb{Z}$, then*

$$\left(X_{i_1}, \ldots, X_{i_d}^\top\right) \overset{D}{=} \left(X_{i_1+h}, \ldots, X_{i_d+h}^\top\right).$$

One may think that strict stationarity implies stationarity, but actually neither implies the other in general. This fact is well-documented in the standard textbooks. For example, a strictly stationary sequence that does not have the second moments is not weakly stationary. However if $X_t \in L^2$ (i.e. $\mathbb{E}X_t^2 < \infty$) and is strictly stationary, then X_t is stationary. Conversely, if X_t is stationary and a Gaussian process, then X_t is strictly stationary. As

we have mentioned in Section 1.1, the FD in this thesis are assumed to be in $\mathcal{H}$, where the process being strictly stationary does not imply weak stationarity.

In most real life scenarios, tine series raw data are often non-stationary. In those scenarios, one typically employs some data-preprocessings to obtain an approximately stationary series. These preprocessing techniques often include detrending, differencing or deseasonalizing the series. When the data has been transformed into stationary series, one may employ a model selection procedure to find a model which best captures the characteristics of the given sample of observations. Once a model is selected, one often wishes to perform statistical inference on the model, including goodness of fit, forecasting, and parameter estimation.

Time series analysis encompasses two principal approaches, namely *time domain* and *frequency domain*. The time domain approach considers temporally lagged relationships of the series. In contrast, the frequency domain focuses on sinusoids. For the purpose of this thesis, our focus will solely be on the time domain. However, readers with an interest in exploring the frequency domain may refer to the monographs by Brillinger (2001) and Bloomfield (2004). There is a variety of methods developed in the time series literature that are for model selection and forecasting. Among which, there are linear time series methods, non-linear time series models, and state space models or hidden Markov models. This thesis will only look at the first two topics, while readers interested in state space models may consider Commandeur and Koopman (2007) and Durbin and Koopman (2012).

There are several ways of quantifying and identifying the serial dependence structures in time series. Perhaps the most commonly used technique is an *autocorrelation function* (ACF). The ACF is a scaled version of the autocovariance function (1.6), and is defined as

$$\rho(s,t) = \frac{\gamma(s,t)}{\sqrt{\gamma(s,s)\gamma(t,t)}}, \qquad \text{where } s,t \in \mathbb{N}. \tag{1.7}$$

Under the assumption of stationarity, assuming $s \geq t$, say $t = s + h$, where $h \geq 0$, the notation may be simplified as $\rho(h) \coloneqq \rho(t,s)$. We refer to the difference in the two time points h as a *lag* from now on. The ACF is a tool to examine serial dependence in time series and identify suitable models for time series. One may estimate the autocovariance

function and the ACF at lag h consistently using their sample counterparts, as

$$\hat{\gamma}(h) = \frac{1}{N}\sum_{t=1}^{N-h}(X_t - \bar{X}_N)(X_{t+h} - \bar{X}_N) \quad \text{and} \quad \hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}, \quad h = 1, 2, \cdots, N.$$

Under the mild assumptions that X_t is i.i.d. and $\mathbb{E}|X_t|^4 < \infty$, if X_t is the random error process, the asymptotic distribution of the sample ACF is given as

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

where $\mathcal{N}$ is a normal distribution, and $\xrightarrow{D}$ denotes convergence in distribution. The technical details of the proof can be found in Theorem A.7 of Shumway and Stoffer (2017). Consequently, if the observed process is white noise (WN), then $100(1-\alpha)\%$ of the sample ACF are expected to lie within the limits of the corresponding confidence interval. The ACF is hence a useful tool to reduce complex time series process to a WN.

For instance, the left panel in Figure 1.1 displays the growth rate of weekly closing returns of the S&P 500 index from January 3, 2003 to September 28, 2012. It shows that this series has constant mean around zero, while the variation is roughly the same except for a short time period around week 300. The latter characteristic indicates that the assumption that the series is a WN process is questionable. The right panel of the figure shows the estimated ACF with strong WN bound. We observe spikes that extend beyond the confidence bands, indicating that the series is not a *white noise process*. In fact, the series is more likely to exhibit weak WN characteristics due to the presence of volatility clustering in the data. We will discuss certain models designed to address the scenario of weak WN in the subsequent section, namely Section 1.3.3. However, it's important to emphasize that this aspect is not the central focus of this thesis.

Definition 1.3.3 (White noise process) *If $\{\varepsilon_t\}_{t\in\mathbb{N}}$ has expectation zero $\mathbb{E}X_t = 0$ for all t, and covariance $\mathbb{C}\text{ov}(\varepsilon_t, \varepsilon_s) = \sigma_\varepsilon^2$ only when $s = t$ and zero otherwise. Then we refer to such a sequence as a white noise process, which is denoted as* $\text{WN}(0, \sigma_\varepsilon^2)$.

The ACF offers several advantages, such as its ease of interpretability, its connection to the Pearson correlation function in non time series contexts, and its visual representation.

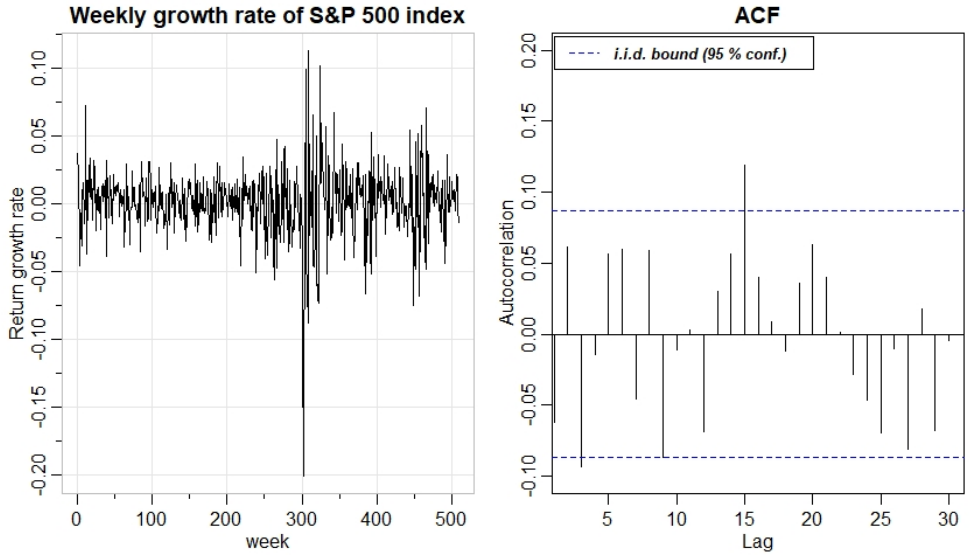


Figure 1.1: **Left**: The growth rate of weekly closing returns of the S&P 500 from January 2003 to September 2012. The large spikes centre around week 300 suggest the presence of vilatility clustering in the data. **Right**: The autocorrelation function computed up to lag 30 with 95% strong WN band. The spikes exceeding the strong WN noise bands suggests the presence of serial dependence, which needs to be addressed.

However, it is important to acknowledge that the ACF has its limitations. One of the criticisms it faces is that it solely provides information regarding the correlation at a pre-selected lag, thereby limiting its usefulness in examining the broader dependence structure. Consequently, a class of tests has been developed to address and examine groups of ACFs collectively, rather than on an individual basis. This approach recognizes that even if none of the ACFs at the selected lags exhibit a significant deviation from a WN, a group of ACFs may collectively yield a large test statistic, indicating a strong serial dependence.

Notable examples of such tests are referred to as "*portmanteau tests*," and include the *Ljung-Box test*. The portmanteau test states that, under a set of mild assumptions, the test statistic for assessing the significance of autocorrelation coefficients of the samples $X_1, \ldots, X_N$ from lags 1 to H is denoted as $Q_{N,H} \coloneqq N \sum_{j=1}^{H} \hat{\rho}(j)$, which approximately follows a chi-squared distribution with a specific number of degrees of freedom. The Ljung-

Box test operates similarly. However, it is important to note that we will revisit this discussion after introducing the model, as the degrees of freedom are contingent upon the model being employed.

1.3.2 Linear processes

Time series is a broad area, and to narrow our scope, we focus on the linear processes which we will use in the later chapters. This class includes some of the most widely used time series models, such as an autoregressive moving average (ARMA) models, which provides a general and flexible framework for studying stationary processes. It is originated from a seminar book of Whittle (1951), and later it is extended to accommodate nonstationary time series in a groundbreaking book (Box and Jenkins, 1970) and known as an autoregressive integrated moving average (ARIMA) model. The aim of these models is to analyze the structure of time series data and use time series models to fit the time series given data.

The data is usually assumed to follow a typical model such that

$$X_i = g(X_{i-1}, \dots) + \varepsilon_i, \tag{1.8}$$

where $g(\cdot)$ is linear in the past observations $X_{t-1}, \dots$, and ε_t is a random fluctuation that is independent of $g(\cdot)$. In time series, the random error is usually assumed to be a WN.

If a time series model fits a dataset well, the residual, defined as $r_t \coloneqq X_t - \hat{X}_t$ where $\hat{X}_t$ is the predicted value, namely the forecasts, r_t would behave similarly to the WN ε_t under model (1.8). Hence, it is of interest to use a WN test to check the model fitting.

Definition 1.3.4 (Autoregressive (AR) process) *A time series $\{X_t\}_{t\in\mathbb{N}}$ is said to be an autoregressive process of order p, written as AR(p), if*

1. *it is stationary, and*
2. *if for every t,*

$$X_i - \mu = \phi_1(X_{i-1} - \mu) + \cdots + \phi_p(X_{i-p} - \mu) + \varepsilon_t, \tag{1.9}$$

where $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$ given that $\phi_p \neq 0$.

Definition 1.3.5 (Autoregressive moving average (ARMA) process) *A time series* $\{X_t\}_{t\in\mathbb{N}}$ *is said to be an ARMA*(p, q) *process if*

1. *it is stationary, and*
2. *if for every* t,

$$X_i - \mu = \phi_1(X_{i-1} - \mu) + \cdots + \phi_p(X_{i-p} - \mu) + \tau_1\varepsilon_{i-1} + \cdots + \tau_q\varepsilon_{i-q}, \tag{1.10}$$

where $\phi_p, \tau_q \neq 0$, $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$ *given that* $\phi_p \neq 0$.

Sometimes, there are seasonal or periodic components augmented to the models. Before introducing the model for this, it is easier to write the models in a compact form by introducing what is often called a *backshift or lag operator*. It is a commonly used way to express linear time series models and is widely used in textbooks. The backshift operator B is defined as follows:

$$BX_t = X_{t-1},$$

which can be further extended to

$$B^kX_t = X_{t-k}, \text{ for } k \in \mathbb{N}.$$

The backshift operator can be used to express the ARMA type of models more compactly. First we write the AR(p) model as

$$\Phi(B)X_t \coloneqq (1 - \phi_1B - \phi_2B^2 - \cdots - \phi_pB^p)X_t = \varepsilon_t.$$

Sometimes the $\Phi(B)$ operator is referred as an *autoregressive operator*. Similarly, we express the noise part in a concise form as

$$\Psi(B) \coloneqq (1 - \psi_1B - \psi_2B^2 - \cdots - \psi_pB^p)\varepsilon_t.$$

In practice, time series observations often exhibit a high degree of persistence. To assess the "whiteness" of the data or residual, various statistical tests have been developed, including the Ljung-Box test mentioned in the previous section. It is common for the original time series, denoted as $\{X_i\}_{i\in\mathbb{N}}$, to be nonstationary, while the d-th differenced

series, such as the first-order differenced series $\{X_i - X_{i-1}\}_{i\in\mathbb{N}}$, is stationary. To address this characteristic, the ARIMA model extends the ARMA model. Instead of analyzing the original time series, the ARMA model is applied to the d-th differenced series, denoted as $\{X_i - X_{i-d}\}_{i\in\mathbb{N}}$, resulting in the ARIMA$(p, d, q)$ model.

Moreover, it is common for time series to display seasonal and cyclic patterns, as seen in series related to air pollution or stock prices. These seasonal behaviors can be incorporated using the backshift operator. For instance, in temperature data, a yearly cycle might be meaningful. Therefore, by indexing the series with month, we can use $\{X_i - X_{i-12}\}_{i\in\mathbb{N}}$ to capture a seasonal lag. For a purely seasonal autoregressive moving average model, ARMA$(P, Q)_s$, we may write it in the following form

$$\Phi_P(B^s)X_i = \Theta(B^s)\varepsilon_i,$$

where the following operators

$$\Phi_P\left(B^s\right) \coloneqq 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps},$$

and

$$\Theta_Q\left(B^s\right) \coloneqq 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs},$$

are called a *seasonal autoregressive operator* and a *seasonal moving operator*, respectively. In this model, there is one thing which is defined as the seasonal difference of order D:

$$\nabla_s^D X_t \coloneqq (1 - B^s)^D X_t,$$

where $D \in \{1, 2 \ldots, \}$. Integrating the seasonal operator and the differencing operator into the ARIMA model, following the notation given on page 150 in Shumway and Stoffer (2017), we now can define a seasonal ARMA model of the differenced series as follows:

Definition 1.3.6 (Seasonal autoregressive integrated moving average (SARIMA) process) *A time series $\{X_t\}_{t\in\mathbb{N}}$ is said to follow an SARIMA model of order p, d, q, P, D, Q, and seasonal period s, abbreviated as $(p, d, q) \times (P, D, Q)_s$ if*

1. *it is stationary, and*

2. *if for every* t,

$$\Phi_P\left(B^S\right)\Phi_p(B)\nabla_s^D\nabla^d X_i = \Theta_Q(B^s)\theta_q(B)\varepsilon_i, \tag{1.11}$$

where $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$.

We will use these models in Chapter 3. Once the model is specified, one often wishes to assess its goodness of fit. Let N denote the number of data points, and $\hat{X}_i$ represent the fitted value for $i \in \{1, \ldots, N\}$. We can examine the residuals, defined as $r_i \coloneqq X_i - \hat{X}_i$, to evaluate how well the model captures the underlying dependence structure. If the residuals exhibit behavior similar to that of a WN process, it suggests that the model adequately captures the underlying serial dependence in the data. Specifically, under the null hypothesis that the residuals r_i behave similarly to a WN ε_i, the sampled autocorrelation function will converge to a standard normal distribution with a rate of $N^{-1/2}$. Furthermore, for an ARMA(p, q) model with a pre-specified value of H such that $N > H \geq (p + q + 1)$, the portmanteau test statistic will asymptotically follow a $\chi^2_{H-(p+q)}$ distribution, where $H - (p + q)$ represents the degrees of freedom.

A modified and improved version of the Box-Pierce test (Box and Pierce, 1970) was proposed by Ljung and Box (1978) to improve poor finite sample performances of the original test. This enhanced test is given as

$$Q_{N,H} \coloneqq N(N+2)\sum_{j=1}^{H}\frac{\hat{\rho}_j^2}{N-j},$$

where H is the maximum lag that are being tested, N is the sample size, and $\hat{\rho}_j$ is the sample autocorrelation. Note that the asymptotic distribution of this test statistics $Q_{N,H}$ as N tends to infinity is still $\chi^2_{H-(p+q)}$ under the null hypothesis that the time series data is i.i.d..

1.3.3 Nonlinear process

The class of nonlinear processes/models finds extensive applications across various fields, in particular in financial econometrics. Among the widely utilized nonlinear time series

models, an autoregressive conditional heteroskedasticity (ARCH) and a generalized autoregressive conditional heteroskedasticity (GARCH) models are prominent. These models are motivated by the observation that the volatility or heteroskedasticity of data tends to exhibit clustering behavior over short periods of time, which traditional ARMA models are unable to adequately capture. However, it is important to note that the investigation of these models falls beyond the scope of this thesis. Interested readers seeking are encouraged to refer to Chapter 3 of Tsay (2005) and Francq and Zakoian (2019).

1.3.4 Functional time series

Similar to the spirit of time series, when FD objects are obtained sequentially over time, it is natural to expect them to exhibit serial correlation. We refer to this type of data as FTS. Figure 1.2 displays the concentrations of particular matter with an aerodynamic diameter of less than 10 micrometers, denoted PM10, in ambient air taken in Graz-Mitte, Austria from October 1, 2010 to October 11, 2010. The curve is segmented into 10 subintervals, each represents a single day of pollution. As one may imagine, air pollution from each day affects the next day; and so we expect these curves to be serially dependent. This is one of the key characteristics that separates FTS from most other scenarios in FDA which temporal dependency between the observed values must be taken into consideration. The focus of FD literature are centred around i.i.d. observations. Without incorporating temporal dependency, the corresponding estimation and inferential procedures may be inappropriate.

Since the monograph of Bosq (2000) which focuses on the theory of linear functional time series, FTS literature has seen an upsurge in exposure; see, for example, Hörmann and Kokoszka (2012) and the references therein. See also Chapters 13 – 18 in Horváth and Kokoszka (2012) and Chapter 8 in Kokoszka and Reimherr (2017) for textbook treatments.

Akin to scalar or multivariate time series, forecasting and performing statistical inference are two of the most important aspects of FTS. Similarly to the scalar or multivariate counterparts, many procedures in the FTS literature requires a given FTS object to be stationary. To address this fundamental issue, Horváth et al. (2014) proposed a stationarity test. Perhaps the most used model in FTS literature is the functional autoregressive model

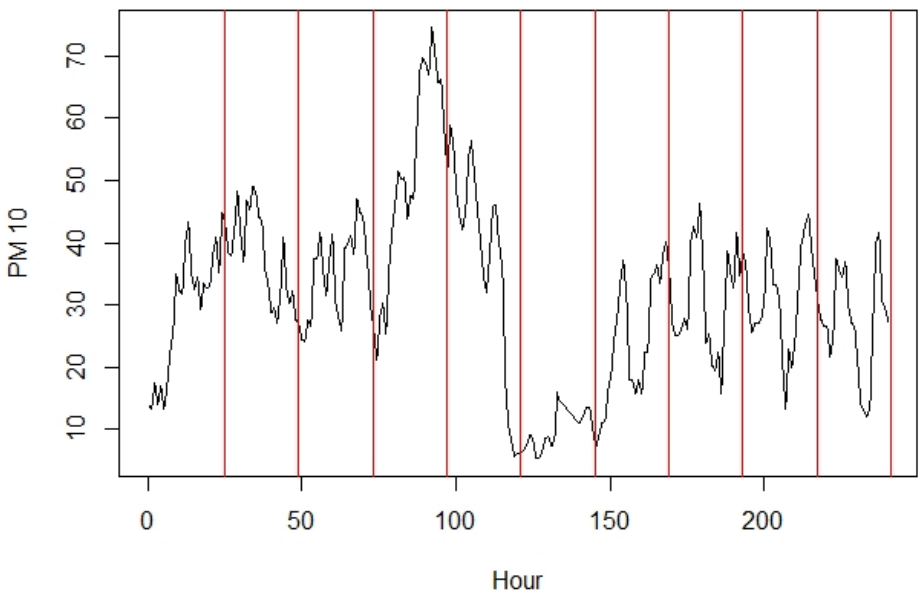


Figure 1.2: The concentrations of particular matter with an aerodynamic diameter of less than 10 micrometers (PM10) in ambient air in Graz-Mitte Austria from October 1, 2010 to October 11, 2010. The curves are evenly segmented into 10 sub-intervals, which represents the change in PM10 of each day.

(FAR) , especially the FAR(1) model. We will elaborate those autoregressive processes in the next section. One of our aims is to quantify serial dependency of FTS. In particular, the previous work of investigating the correlation structure are centred around the "norm" approach, which misses the "sign" of the dependency.

The state-of-the-art investigation of the WN assumption is centred around the (auto)covariance operator. Gabrys and Kokoszka (2007) propose a portmanteau test for FD based on lagged autocovariance operator and FPCA; also see Chapter 7 of Hörmann and Kokoszka (2012) for the descriptions of this method. Horváth et al. (2013)'s approach uses the sum of the L^2 norm of the estimated correlation functions. Kokoszka et al. (2017) propose an-

other version of the portmanteau test to account for potential presence of conditional heteroscedasticity in previous settings. The above references mentioned thus far mainly lie in the time-domain, and literature about independence test for FTS on the spectral domain are scanty until recently. The independence test on the frequency-domain has recently been investigated based on a spectral density operator; see a review paper of WN tests for FTS by Kim et al. (2023). When the interest is to identify the presence of serial correlation at an arbitrary lag h, Mestre et al. (2021) proposed a functional version of the ACF shown in (1.7) which is helpful in model identification and WN diagnosis. Let a lagged autocovariance function and its empirical estimate at lag h be respectively defined as

$$c_h(t,s) = \mathbb{E}[(X_i(t) - \mu(t))(X_{i+h}(s) - \mu(s))], \quad \text{and}$$
$$\hat{c}_h(t,s) = \frac{1}{N}\sum_{i=1}^{N-h}[(X_i(t) - \bar{X}_N(t))(X_{i+h}(s) - \bar{X}_N(s))], \quad t,s \in [0,1],$$

where $\bar{X}_N(\cdot) = N^{-1}\sum_{i=1}^{N} X_i(\cdot)$ is a point-wise sample mean function. Following the work of Kokoszka and Reimherr (2017) and Mestre et al. (2021), who study a functional autocorrelation function (fACF) at lag h as

$$\rho_h = \frac{\|c_h\|}{\int_0^1 c_0(t,t)dt}, \quad \text{where } \|c_h\| = \left(\int_0^1\int_0^1 c_h^2(t,s)dtds\right)^{1/2}, \tag{1.12}$$

with their respective large sample properties. Through simulations and empirical studies, they further show that the proposed fACF is able to extract information on the serial dependence structure at specified lags. This information is useful to model selection for FTS. The left panel in Figure 1.3 shows 400 simulated i.i.d. Brownian bridges, i.e., the conditional Brownian motion $B(t)$ where $B(1) = B(0) = 0$; The right panel shows the fACF of the simulated Brownian bridges with 95% error bound. The estimated fACF at all lags are within the bound indicating that there is no significant serial dependence, which agrees with the underlying data generating process. Despite the usefulness of FPCA in detecting the correlation between FTS objects, unlike the ACF shown in Figure 1.1, fACF is not able to capture the negative dependence as shown in Figure 1.3. This information may have both theoretical and practical significance, such as in medical or environmental applications, where the "sign" is often as important as the magnitude.

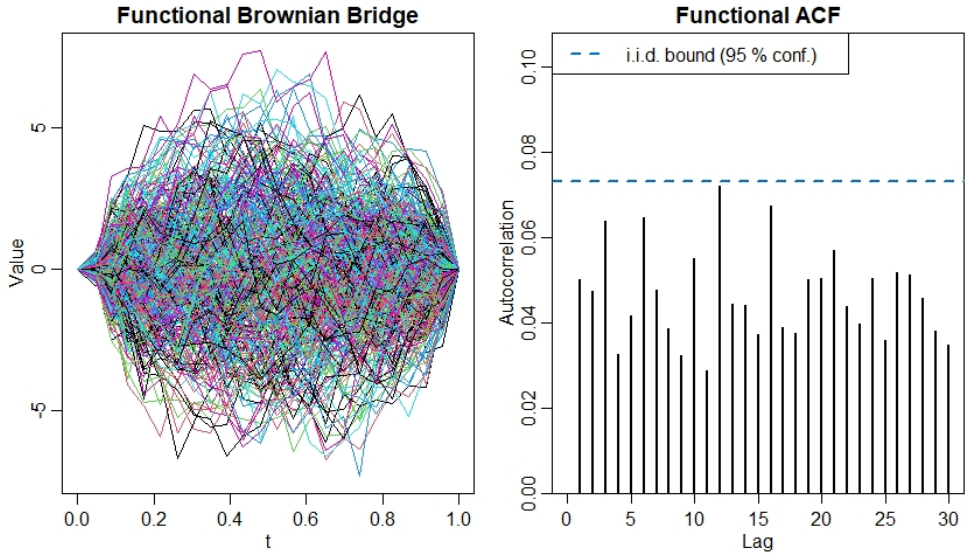


Figure 1.3: **Left**: 400 simulated Brownian bridges, which are i.i.d. **Right**: Functional autocorrelation function with the 95% error bound under the assumption of data are i.i.d. The fACF at all lags are within the 95% bound indicates the independence of the simulated curves, which agrees with the data generating process.

1.3.5 Functional time series models

A variety of FTS models and other general linear processes have been extensively studied in the literature. One of the earliest monographs to explore the theoretical aspects of these linear processes in Hilbert and Banach spaces is Bosq (2000). These models can be viewed as extensions of the linear processes taught in elementary time series analysis courses. Among these models, the functional autoregressive model is particularly widely used. Functional autoregressive processes, or models, are an extension of the autoregressive model commonly used for univariate or multivariate time series. To provide a brief overview, let $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ represent i.i.d. WN functions in a function space, typically the Hilbert space of square integrable functions $\mathcal{H}$. Let $\Phi_j(\cdot)$ be a linear operator mapping $\mathcal{H}$ to $\mathcal{H}$ for each j. The functional autoregressive model of order p, denoted as FAR(p), is defined as

follows:

$$(X_i - \mu) = \sum_{j=1}^{p} \Phi_j \left(X_{i-j} - \mu\right) + \varepsilon_i,$$

where μ is the mean of the process, which is often estimated by $\hat{\mu}(\cdot) = \bar{X}_N(\cdot) = N^{-1} \sum_{i=1}^{\top} X_i(\cdot)$.

The FAR model extends the AR models (1.9) in the univariate time series, where the variables are now considered to have values in $\mathcal{H}$. Extensive research has been conducted to investigate this class of models, especially the FAR(1) case. The FAR models are commonly referred to as *Autoregressive Hilbertian processes with order p* (ARH(p)), as described in the influential book by Bosq (2000). Notably, this book provides extensive limiting theories pertaining to covariance operators and the associated kernels in Chapters 4 to 6. In particular, we briefly outline a method for estimating the kernels in FAR(1) case, which is based on Section 8.2 in Kokoszka and Reimherr (2017) and Section 13.2 in Horváth and Kokoszka (2012). Without loss of generality, we suppose that $\mu = 0$. Otherwise, we may take $\tilde{X} \coloneqq X - \mu$ and work with $\tilde{X}$ in the following procedure. We first rewrite the FAR(1) model as

$$X_{i+1}(t) = \Phi(X_i)(t) + \varepsilon_{i+1}(t) = \int \phi(t,s) x_i(s) ds + \varepsilon_{i+1}(t).$$

Note we require the kernel $\phi(\cdot,\cdot)$ to be square integrable, i.e. $\|\phi\|_2 < \infty$. Suppose that λ_i's are the eigenvalues and v_i be the eigenvector of the covariance operator C, that is $C(v_j) = \lambda_j v_j$. Using FPCA $\hat{v}_j, j = 1, \ldots, p$, the estimator of the kernel operator and its associated kernels are

$$\widehat{\Phi}_p(x) = \frac{1}{N-1} \sum_{k=1}^{N-1} \sum_{j=1}^{p} \sum_{i=1}^{p} \hat{\lambda}_j^{-1} \left\langle x, \hat{v}_j \right\rangle \left\langle X_k, \hat{v}_j \right\rangle \left\langle X_{k+1}, \hat{v}_i \right\rangle \hat{v}_i \tag{1.13}$$

and

$$\hat{\phi}_p(t,s) = \frac{1}{N-1} \sum_{k=1}^{N-1} \sum_{j=1}^{p} \sum_{i=1}^{p} \hat{\lambda}_j^{-1} \left\langle X_k, \hat{v}_j \right\rangle \left\langle X_{k+1}, \hat{v}_i \right\rangle \hat{v}_j(s) \hat{v}_i(t), \tag{1.14}$$

respectively. That is, the relationship between (1.13) and (1.14) is as follows:

$$\widehat{\Phi}_p(x)(t) = \int \hat{\phi}_p(t,s) x(s) ds. \tag{1.15}$$

The detailed derivation of the result may be found in Section 8.2 in Kokoszka and Reimherr (2017) and Section 13.2 in Horváth and Kokoszka (2012).

As for the general FAR(p) case, there are more general methods for estimating the kernels, and we refer to Bosq (2000) and Horváth and Kokoszka (2012) for an in-depth discussion. There are other linear FTS models including functional autoregressive moving average model (Bosq, 2014), functional autoregressive conditional heteroskedasticity (FARCH) model (Hörmann et al., 2013), and functional generalized autoregressive conditional heteroskedasticity model (Aue et al., 2017).

As for available statistical software, there exist various software implementations in R programming language that are specifically designed to perform FTS analysis. Noteworthy examples include the **ftsa** package (Shang, 2013; Hyndman and Shang, 2022) and the **far** package, both of which provide functionalities related to FTS analysis. Additionally, there is the WN tests for FD package, **wwntests** (Kim and Petoukhov, 2022), which offers WN tests specifically tailored for FD. Another valuable resource is the **freedom.fda** package (Hörmann and Kidziński, 2017), which implements a version of the FPCA specifically designed for FTS, based on the work of Hörmann et al. (2015). Furthermore, the fACF method that was mentioned in (1.12) is included in the **fdaACF** package by Mestre et al. (2020).

1.4 Spatial Statistics

Spatial statistics is another sub-field within the broader discipline of statistics, focusing on datasets that possess a spatial dimension. Its unique structure lies in the pronounced interdependence observed between data points located in close proximity on a grid or spatial domain. This spatial proximity gives rise to a heightened degree of dependence among nearby data points. In the past, rsearchers applied the spatial statistics techniques to a vast variety of applied fields, spanning from demography, transportation, environmental studies, ecology and beyond.

In this section, we offer a concise overview of essential concepts, aiming to provide a preliminary understanding that will pave the way for subsequent chapters, in particular

Chapter 4. It is, however, important to note that this section is not intended to serve as a comprehensive review. We only focus on *geostatistical spatial data*, which refers to data comparised of multiple quantities/points in a region of interest $\mathcal{D}$. There are other spatial structured data type, for instance, *Lattice/regional data* where the data are measured not at its exact location, but rather as a region; and *point pattern analysis data* refers to the study of spatial locations of occurrences of the event of interest. The main difference of geostatistical spatial data and point pattern analysis data is that the former treat the locations as an explanatory variable and the value as a response, while the latter treat both values as a response. For a comprehensive understanding of the subject, we refer readers to the monograph by Cressie (1993). As a brief introduction, let $\boldsymbol{s}_k$ denote the k-th (spatial) location. $\boldsymbol{s}_k$ may lie in a d-dimensional Euclidean space $\mathbb{R}^d$, or another arbitrary space denoted by $\mathcal{D}$. Then the observation, $X(\boldsymbol{s}_k)$ is a measurement taken of the location $\boldsymbol{s}_k$, and the dataset is usually denoted by $\boldsymbol{X}^K := \{\boldsymbol{X}(\boldsymbol{s}_k),\ \boldsymbol{s}_k \in \mathcal{D},\ k = 1, \ldots, K\}$. Those observations are viewed as a realization of a random process referred to as a random field. A more detailed treatment of random field can be found in Hristopulos (2020). Similar to time series, in spatial statistics, there is a concept of stationarity, which also plays an important role. As in Definition 1.3.2, a random field $\boldsymbol{X}^K$ is strictly stationary if for any spatial locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k$ such that for any shift $\boldsymbol{\delta}$, we have

$$\{\boldsymbol{X}(\boldsymbol{s}_1), \ldots, \boldsymbol{X}(\boldsymbol{s}_k)\} \overset{D}{=} \{\boldsymbol{X}(\boldsymbol{s}_1 + \boldsymbol{\delta}), \ldots, \boldsymbol{X}(\boldsymbol{s}_k + \boldsymbol{\delta})\}.$$

Similarly the stationary is defined as if the mean $\mathbb{E}\boldsymbol{X}(\boldsymbol{s})$ remains the same for all locations, and the covariance between two locations $\mathbb{C}\text{ov}(\boldsymbol{X}(\boldsymbol{s}), \boldsymbol{X}(\boldsymbol{s}+\boldsymbol{\delta})$ only depends on the distance $\boldsymbol{\delta} > \boldsymbol{0}$.

Recently, SFD have received more attention, see Section 5 in Koner and Staicu (2023), and Chapter 9 in Kokoszka and Reimherr (2017). The distinct difference is that now $\boldsymbol{X}(\boldsymbol{s})$ is a random function. In this thesis, we assume, similarly to Section 1.1, that all the SFD of this type lie in $\mathcal{H}$, and denoted by $\boldsymbol{X}(\boldsymbol{s}; t)$ for the observation $\boldsymbol{X}(\boldsymbol{s})$ at time t. In addition, we assume that each $X(\boldsymbol{s}; \cdot)$ such that $\mathbb{E}\left\|\boldsymbol{X}(\boldsymbol{s}; \cdot)\right\|^2 = \int \mathbb{E}X^2(\boldsymbol{s}; t)dt < \infty$, and is strictly stationary. These two conditions imply the mean function is constant across all locations, i.e. $\mathbb{E}\boldsymbol{X}(\boldsymbol{s}; \cdot) = \mu(\cdot)$ and covariance $C(\boldsymbol{\delta}; t, u) = \mathbb{C}\text{ov}(\boldsymbol{X}(\boldsymbol{s}; t), \boldsymbol{X}(\boldsymbol{s} + \boldsymbol{\delta}; u))$. In the spatial statistics literature, a popular approach to estimate this covariance function is

to use something called a semivariogram, but a detailed discussion of it is out of the scope of this thesis. Interested readers may refer to Section 9.1 in Kokoszka and Reimherr (2017) and references therein. One common assumption, known as *isotropy*, is usually employed.

Definition 1.4.1 (Isotropy) *For a geostatistical data* $\{\boldsymbol{X}(\boldsymbol{s})\}$*, the correlation between two observations* $\mathbb{C}\text{ov}(\boldsymbol{X}(\boldsymbol{s}), \boldsymbol{X}(\boldsymbol{s}'))$ *depends only on* $\|\boldsymbol{s} - \boldsymbol{s}'\|_2$.

In Chapter 4, we will adopt the assumptions of strict stationarity, isotropy, and square integrability in our analysis. These assumptions will be restated and discussed in detail within that chapter.

As for the available software in `R` language, the **geofd** package developed by Giraldo et al. (2020) offers functions to perform predictions within the field of spatial statistics, which sometimes known as *kriging* in the context of spatial statistics. The **fdaPDE** package (Arnone et al., 2023) allows for estimation of partial differential equations that capture the spatial dependence while leveraging the availability of prior information.

1.5 Robust Methods in Functional Data

Robust statistics is another classic area in statistics which considers the case where the assumptions we made do not hold. It has been around 90 years since, perhaps the first person who noted this was Pearson (1931) and termed "robustness" by Box (1953), as mentioned in the review paper by Huber (1972). There is a vast amount of studies of robustness properties among different subareas in statistics, as assumptions are often violated in real life applications. Huber (2011) in the *International encyclopedia of statistical science* is perhaps the most popular reference to this field. Maronna et al. (2019) provides a textbook length description of this subject, including ways to measure the robustness, robust regression methods and data with temporal dependence; with implementations in its associated `R` package **RobStatTM** (Salibian-Barrera et al., 2023).

Centring the data before proceeding to subsequent steps is a common statistical practice that offers several advantages. For example, pre-centring the data in a regression analysis

can ensure identifiability by allowing the estimators of other parameters to be distinguished from the intercept term. Additionally, assuming that the data is centred around the origin makes it easier to explain important concepts and simplify notation for readers. Nevertheless, centring the dataset is not a straightforward process in many settings and presents a challenge during estimation steps. In Chapter 3, we use a *geometric median*, which may be referred as *spatial median* to centre the FD. The purpose of this section is to elucidate the background of the multidimensional median that are used for centring the data in the problems considered in this thesis.

1.5.1 Univariate and multivariate centring

The common practice before performing any statistical analysis including inference or prediction is to analyze the trend or the average μ of the data and the fluctuation or the random noise ε, separately. An immediate question would arise is how to estimate the centre of distributions. When centring the data $X_1, \ldots, X_N$, the sample mean $\bar{X}_N \coloneqq N^{-1} \sum_{i=1}^{N} X_i$ is perhaps the most commonly used statistic. On the other hand, it is commonly known that $\bar{X}$ itself is not robust, as one contaminated value may change it to an arbitrary degree. To address the lack of robustness of the sample mean $\bar{X}$, the median is often used in the literature of robust statistics. The median $\tilde{X}_n$, defined as the "middle" point of the sorted data, is much more robust against outliers or contaminated points. To be specific, for one-dimensional data, the median is defined as

$$\tilde{X}_N = \begin{cases} X_{(N+1)/2} & \text{if } N \text{ is odd,} \\ (X_{(N/2)} + X_{(N/2)+1})/2 & \text{otherwise} \end{cases}$$

It turns out that, the L_1 median of the data points $X_1, \ldots, X_N$ can be found by solving the following optimization problem

$$\arg\min_{c \in \mathbb{R}^p} \mathbb{E}\left[\|X - c\|_E - \|X\|_E\right],$$

where $\|\cdot\|_E$ is a standard Euclidean norm of a vector. In the univariate case where the dimension $p = 1$, we have

$$\arg\min_{c \in \mathbb{R}} \mathbb{E}\left[|X - c| - |X|\right].$$

Given the samples $x_1, \ldots, x_n$, one may find a sample median $\mu_{\text{Med}} \in \mathbb{R}^d$ by solving the above optimisation problem. It is well-documented that centring the data using the median will be more robust compared to using the mean. A few review papers on this topic are available in the literature, including Small (1990), Chapter 1 in Becker et al. (2014) for general overview, and Fritz et al. (2012) reviews the computational algorithms.

1.5.2 Centring for functional data

For FD, as it may be viewed as a natural extension of multivariate data, the centring is usually used with the functional mean which is defined as

$$\mu(t) = \mathbb{E}X(t), \quad t \in [0, 1].$$

In many cases, we may see the asymptotic rate of replacing μ by its estimate using the sample mean $\hat{\mu} = \bar{X}_N$. However, it is not always to case. For a broader class of robust estimations or a general notion such as data depth, see Fraiman and Muniz (2001), Cuevas et al. (2007), Serfling and Wijesuriya (2017) and Gijbels and Nagy (2017).

1.6 Probabilistic Forecasts

Probabilistic prediction is a type of prediction which estimates the probability of a future event. There exists a rich literature on this topic, see e.g., Gneiting and Katzfuss (2014) for a detailed review of predictive models and methods, and of evaluating the quality of probabilistic forecasts. We provide a brief overview here.

When it comes to accessing the quality of forecasts, one may wonder what is the best way to evaluate the predictions. Murphy and Winkler (1987) and Gneiting et al. (2007) argued that the optimal probabilistic forecasts maximize the sharpness of the forecasted distributions, subject to their calibration. Sharpness, also known as refinement, explains the confidence level between the forecasts or the concentration of the resulting predictions, while the calibration refers to the alignment/agreement between the forecasts and the actual observed outcome distribution. A reliability diagram is a commonly used tool

in accessing the calibration, in which it plots the conditional events frequency against binned predictive probabilities. The deviation from the diagonal reference line indicates a degree of miscalibration. When it comes to measures of sharpness, we may use the graphical or numerical summaries of the width of prediction interval Gneiting et al. (2007), or the sharpness diagram (Bremnes, 2004). Throughout the development of probabilistic forecasts, other measurement of quality such as resolution and bias also existed in the literature, but is not the focus here; for further details, one can read, e.g., Murphy and Winkler (1992). The two indices, sharpness and calibration, may be measured separately, but more commonly, we may estimate them jointly via overall performance measures. One such measures is a scoring rule which can access two metrics simultaneously.

The scoring rule measures the overall performance of probabilistic forecasts. It assigns a numerical value based on the realized value of the outcome and the predictive probabilistic distribution. Let $\mathcal{F}$ be a generic convex class of probability distribution on $\mathbb{R}$. A scoring rule is a function such that $S : \mathcal{F} \times \mathbb{R} \to \bar{\mathbb{R}}$, where $\bar{\mathbb{R}}$ is the extended real line. In the definition of the scoring rule, it is difficult to compare two distinct forecasting methods, or to find the "optimal" strategy. To address these issues, we introduce a proper scoring rule (Murphy, 1972; Gneiting et al., 2007). The proper scoring $L : \mathcal{F} \times \mathbb{R} \to \bar{\mathbb{R}}$ is proper relative to the class $\mathcal{F}$ if

$$L(G, G) \leq L(F, G),$$

for all $F, G \in \mathcal{F}$. It is strictly proper if the equality only holds when $F = G$. We should be cautious at using improper scoring rules as they may lead to false conclusions about the predictive quality (Gneiting et al., 2007; Hilden and Gerds, 2014; Gneiting, 2011). Some common proper scoring rules can be found in Gneiting et al. (2007), in which a few examples are continuously ranked probability score (Matheson and Winkler, 1976; Hersbach, 2000), Brier score (Brier, 1950), and logarithmic score (Good, 1952; Bernardo, 1979).

1.7 Copula

Given the exponential growth of data volume and the increasing complexity of its features, there is a growing demand within the academic community to devise a flexible modelling

framework capable of accommodating such diverse datasets. This issue has garnered significant attention in recent years. One of the approaches is to use *copula* modelling. Copulas provide a versatile and potent approach when it comes to modelling a complex dependence strcture in multivariate data. Copula is a Latin word that means link or tie. It is widely used in the multivariate statistics literature as it provides a flexible approach on the multivariate distribution class. Often there exists dependence between the variables themselves, for instance, measurements of moisture and temperature, may have some dependence. Perhaps the most important characteristic/advantage of copulas is that it allows one to separate univariate marginal distribution functions from the dependence structure of a joint distribution of all of the variables together. To fix ideas, we first state a few notations. Let $\boldsymbol{x}_d = (x_1, \ldots, x_d)^\top$ be a d-dimensional vector, f_i and F_i be a marginal probability density function and distribution function for each $i = 1, \ldots, n$, and F be the joint distribution function. A d-dimensional copula is represented as $C : [0,1]^d \to [0,1]$. Particular, if C is an absolutely continuous copula distribution function, then the corresponding copula density can be obtained by taking a partial derivative of each components as

$$c(u_1, \ldots, u_d) \coloneqq \frac{\partial^d}{\partial u_1 \cdots \partial u_d} C(u_1, \ldots, u_d).$$

In this thesis, we assume that the random variables are all absolutely continuous. The relationship between the multivariate distribution and the copula was first established by Sklar (1959). Following the copula definition stated in Nelsen (1999), Sklar's fundamental representation theorem is as follows:

Theorem 1.7.1 (Sklar's theorem in d-dimensions) *Let F be an d-dimensional distribution function with margins $F_1, \ldots, F_d$. Then there exists a d-copula (d-dimensional copula) C such that for all $\boldsymbol{x} \in \bar{\mathbb{R}}^d$,*

$$F_{\boldsymbol{x}}(x_1, \ldots, x_d) = C\left(F_1(x_1), \ldots, F_d(x_d)\right). \tag{1.16}$$

If all F_i's are continuous, then C is unique. Otherwise, C is uniquely determined on the Range $\operatorname{Ran} F_1 \times \cdots \times \operatorname{Ran} F_d$. *Conversely, if C is a d-copula, and $F_1, \ldots, F_n$ are distribution functions, then the function $F(\cdot)$, defined by (1.16), is a d-dimensional distribution function with margins $F_1, \ldots, F_n$.*

From Sklar's theorem 1.7.1, the relationship between the density function and the *copula density* $c(\cdot)$ may be defined as follows:

$$f(x_1, \ldots, x_d) \coloneqq c(F_1(x_1), \ldots, F_d(x_d)) \prod_{j=1}^{d} f_j(x_j).$$

Given that copulas involve assessing the joint dependence while also accounting for the univariate marginal distributions, it becomes crucial to select an appropriate dependence measure. Although the Pearson correlation coefficient $\mathbb{C}\text{ov}(X, Y)/(\sigma_X \sigma_Y)$, denoted as $\text{Corr}_{Pear}(X, Y)$, is commonly employed, its value can change depending on monotone transformations of the marginal distributions, making it less suitable for use in copula. On the other hand, two widely used measures, namely Kendall's tau and Spearman's rho, are invariant under such transformations, thus making them particularly appropriate for the copula.

Spearman's rho is a nonparametric correlation measure that is similar to the Pearson correlation but acting on the ordered samples. Denote $R_{x,j}$ and $R_{y,j}$ as ranked value from smallest to largest of x's and y's, respectively. Spearman (1904) defined a rank correlation coefficient as

$$\rho_{R(x),R(y)} = 1 - \frac{6}{N(N^2-1)} \sum_{j=1}^{N} D_j^2,$$

where $D_j = R_{x,j} - R_{y,j}$, for $j = 1, \ldots, N$ assuming that there are no ties in both x's and y's.

Kendall's tau is another popular nonparametric correlation statistics based on the sign statistics instead of the rank which is used in Spearman's rho. It is defined as

$$\tau \coloneqq \frac{2}{N(N-1)} \sum_{i<j}^{N} \text{sgn}(x_i - x_j) \,\text{sgn}(y_i - y_j),$$

where $\text{sgn}(\cdot)$ is the sign function. There is an extensive amount of work in the copula literature that used those two correlation measure to examine the correlation between the variables. Since copula is a flexible framework, it also is capable of modelling the heavy tail and asymmetric distributions such as the skew-normal, skew-t, Pareto, three parameters

generalized extreme value, and so on. There are many different kinds of copula functions, depending on the definition of the univariate variables. Some instances include Gaussian copula, Clayton copula and survival copula. Since this topic is rather broad, we refer interested readers to consult the references: For review papers of copula, see Genest and Favre (2007) and Czado and Nagler (2022); and the monograph of Nelsen (1999), Joe (1997) and Joe (2014) provide in-depth theoretical treatments. In particular, we extend one particular kind of the copula class, *factor copula*, to FD in Chapter 4.

1.7.1 Factor copula

Factor copulas are a class of copulas that extends Gaussian factor models, allowing for modelling dependence beyond the Gaussian framework. They offer significant advantages, particularly in terms of ease of interpretability, as the factor model provides a clear understanding of the relationships. Factor copulas incorporate copulas of latent variables, sometimes referred to as factors, with an additive structure. They have gained popularity in modelling dependent data structures, including those in the temporal domain (Hull and White, 2004; Oh and Patton, 2017) and the spatial domain (Krupskii and Genton, 2017; Krupskii et al., 2018). Krupskii and Joe (2013) introduced a vine copula construction that employs bivariate linking functions. The p-factor copula model is defined as follows:

$$C(\boldsymbol{u}) = \int_{[0,1]^p} \prod_{j=1}^{d} C_{j|V_1,\ldots,V_p}\left(u_j \mid v_1,\ldots,v_p\right) \mathrm{d}v_1 \cdots \mathrm{d}v_p, \quad \boldsymbol{u} \in [0,1]^d,$$

where $U_1,\ldots,U_d$ are assumed to be conditionally independent given p-dimensional latent variables $V_1,\ldots,V_p$, which, without loss of generality, are assumed to be i.i.d. uniform with support from 0 to 1. The conditional cumulative distribution function (CDF) of U_j given V_i's is denoted as $F_{j|V_1,\ldots,V_p} = C_{j|V_1,\ldots,V_p}$. For theoretical properties and detailed descriptions, see Section 3.10 in Joe (2014) and the references therein.

Although this approach reduces the number of dependence parameters to a low-dimensional representation ($O(d)$ instead of $O(d^2)$), the selection and interpretation of the linking functions can be challenging. In contrast, the interpretability offered by the former class

provides a full explanation of the dependence structure between variables, which is advantageous in applications where interpretability is a primary focus. In addition, for a good reference on the details of the copula, see the book by Joe (2014). In Chapter 4, we follow the formulation in Krupskii and Genton (2017) of which they extend replicated spatial data in Krupskii et al. (2018) to a spatial-temporal process. Our goal in Chapter 4 is to extend this formulation to the SFD. In Krupskii et al. (2018), they proposed a model with only spatial dependence as follows:

$$X(\boldsymbol{s}) = Y(\boldsymbol{s}) + V_0, \quad \boldsymbol{s} \in \mathbb{R}^d, \tag{1.17}$$

where Y is a Gaussian process and V_0 is a common factor that is independent on $Y(\cdot)$ or location $\boldsymbol{s}$. Krupskii and Genton (2017) further extend this model to incorporate the temporal dependence while still considering the spatial dependence, and the model is as follows:

$$X(\boldsymbol{s};t) = Y(\boldsymbol{s};t) + \beta(\boldsymbol{s};t)\mathcal{E}_{\mathcal{P}(t)}, \quad \boldsymbol{s} \in \mathbb{R}^d,\ t \in \mathbb{R}_+. \tag{1.18}$$

Hofert et al. (2018) provide a comprehensive book along with their associated package, *copula*, which offers extensive examples of various copula types, including elliptical, Archimedean, and extreme-value copulas, among others. In the book and the accompanying manual, Hofert et al. (2023) provide numerous graphical and coding examples. The first version of the copula package was developed by Jun Yan in 2005, who later wrote a paper explaining some early concepts of the package and introducing basic examples. Yan (2007), also includes a discussion and comparison of other related packages for copula problems. Moreover, Joe (2014) provides an associated `R` code with a monograph, implemented in Fortran90 or `C`, to efficiently compute copulas. Additional references for some of the functions used in this software can be found in Krupskii (2014). It is also possible to compute copulas using other software platforms, such as MATLAB (Górecki et al., 2020) and in Python (e.g., the **copulas** package).

1.8 Outline of the Thesis

In this thesis, we focus on the problems centred around FDA. In Chapter 2, we develop a new statistical test and graphical representations for measuring the quality of continuously

updated probabilistic forecasts. This work has been published in Yeh et al. (2022).. In Chapter 3, we introduce novel ways to quantify the correlation between curves, which we term *functional spherical autocorrelation*. It can be seen as the functional counterpart of the autocorrelation function in the time series data framework. This method is helpful in identifying a suitable model, diagnosing the existing correlation, and understanding the underlying dependence structure. This work has been published in Yeh et al. (2023). In Chapter 4, we extend the factor copula model to FDA. This new modelling approach provides notable advantages over existing methods, as it accommodates tail dependence and addresses the issue of tail asymmetry. Lastly, in Chapter 5, we wrap up the thesis with summary and discussions of preceding chapters. We also outline potential avenues for future research and provide concluding remarks.

1.9 Contributions

The contribution of this thesis extends beyond the classical FDA, which mainly studies independent random functions. It falls within streams of study that record multiple functional observations according to either time series and/or spatial designs. The analysis required for these FD must account for complex dependencies between functional observations and calls for more advanced methods to analyze the data. This thesis introduces new methods and models that bridge this gap. Specifically, the summary of the main contributions in this thesis is as follows.

1. We develop a significance test on examining the predictive power of competing continuously updated probabilistic forecasts.

2. We develop identification tools to quantify the correlation in functional time series objects. Our method is able to capture the "direction" of dependency, which previous methods were not able to achieve.

3. Our version of the autocorrelation function for FTS which we termed "functional spherical correlation function," is more robust against outliers compared to the existing method.

4. We develop a portmanteau test to examine the significance of the group autocorrelation at multiple time lags.

5. We develop a model to handle SFD, which is more flexible at handling the case when there are some fat tails, or asymmetric tails.

6. We provide full implementations of the proposed methods in user-friendly computational packages in the `R` programming platform.

Chapter 2

Evaluating Real-time Probabilistic Forecasts with Application to National Basketball Association Outcome Prediction

2.1 Introduction

Probabilistic predictions and forecasts are ubiquitous in modern society, and many individuals consider, and make decisions based on, such forecasts on a routine basis. For example, in the United States probability of precipitation forecasts became widely publicly available starting in the late 1960's, and are now a critical factor in countless people's daily decisions (Murphy, 1998; Council, 2006). Over time the number and scope of probabilistic forecasts readily accessible to the public has increased at a steady pace, and now covers prediction of phenomena ranging from sports (Silver et al., 2019), to politics (Erikson and Wlezien, 2012), to medicine (Spiegelhalter, 1986), to geology (Gomberg, 2015), among many others, some more exotic (Rowe and Beard, 2018), areas.

Many such forecasts are made initially well before the event in question occurs, and are

then continuously updated as new information becomes available. The example that we focus on throughout this chapter is a basketball game outcome prediction in the National Basketball Association (NBA). Websites like `espn.com`, the main web page of the United States-based multinational sports network, ESPN, publish and update, in real time, probabilistic forecasts of the home team winning for each NBA game played. Although the method by which ESPN produces these forecasts is largely proprietary, ostensibly initial probability forecasts of the home team winning are constructed based on information that is available before the game starts, e.g. the usual home court advantage in the NBA, relative team strength, player injuries, etc., and then after the game commences and progresses these forecasts are updated with new information such as the score, game time remaining, ball possession, fouls, in-game player injuries, etc. The resulting probabilistic forecasts and their fluctuations may be viewed as a curve that is a function of the in-game time; see Figure 2.1 for an example. Such curves arise in any similar continuously updated probabilistic forecasting task, and are evidently not unique to the basketball game outcome prediction; see e.g. Silver (2020).

Natural questions to ask when faced with any probabilistic forecast, including those that are continuously updated, are "are these forecasts accurate?" and "could these forecasts be improved upon?". Following the seminal work of Murphy and Winkler (1987, 1992) on evaluating the quality of probabilistic forecasts in meteorology, evaluating a method for producing probabilistic forecasts is often broken into the tasks of measuring its *calibration* and *skill*. A model is deemed well-calibrated if its forecasts are compatible with the observed outcomes. In other words, a model that predicts an outcome with a given probability is well-calibrated if the relative frequency that the outcome occurs matches the probability forecast in the long run. A model is deemed to have higher skill than a competing model if its predictions are "sharper" or "more concentrated" than its competitor. For example, a model that forecasts the probability of a rainy day in New York City on a given day with the long run background rate of rainy days (which happens to be about 33.1% for New York City), will in the long run be well-calibrated, but has less skill than a model, perhaps based on more complete weather data, that correctly predicts rainy days with probabilistic forecasts of zero and one. Excellent reviews and more in-depth discussions of these concepts can be found in Gneiting et al. (2007) and Gneiting and Katzfuss (2014).

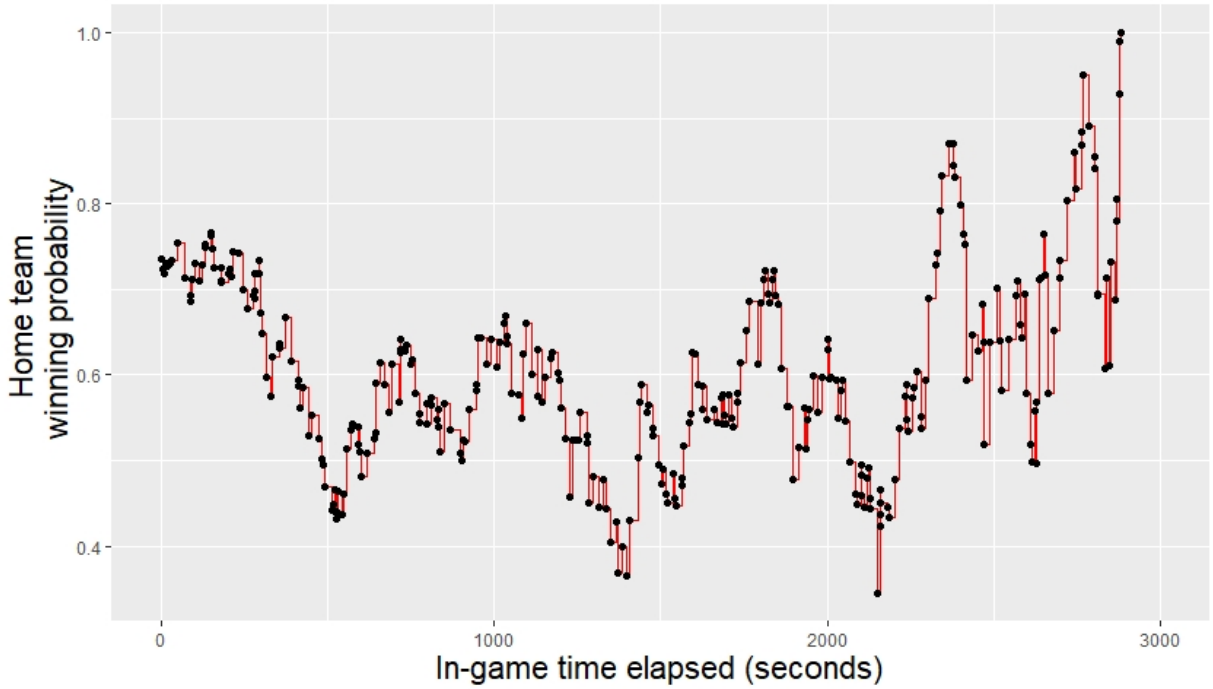


Figure 2.1: Real-time probabilistic forecasts as a function of the in-game time published on `espn.com/nba` (ESPN, 2020) from a November 8, 2018 game in which the Los Angeles Lakers hosted the Minnesota Timberwolves. The black points are the points at which ESPN updated the home team winning probability forecasts due to events occurring in the game, and the red line is a piece-wise constant interpolation between these forecasts to produce a real-time probabilistic forecast curve (function).

The goal of this chapter is to develop simple, easily interpretable, tools for evaluating the calibration and skill of methods to produce continuously updated probabilistic forecasts, and to apply these methods to evaluate the probabilistic forecasts pertaining to the NBA basketball game outcome prediction published on ESPN. In terms of evaluating model calibration of probabilistic forecasts, standard tools are reliability diagrams and calibration plots, in which outcome frequencies are plotted against binned forecast probabilities; see Gneiting et al. (2007). Below we show how such curves can be extended in the continuously updated case to calibration surfaces, and how such surfaces can be summarized to show at

a glance whether a given method is well-calibrated. In order to evaluate the relative skill of one continuously updated forecasting model against another, we employ the method of Lai et al. (2011) to construct confidence intervals for the average loss difference measured by the Brier score (Brier, 1950) between the two models at given time points throughout the updating process. Estimating these intervals pointwise can be used to construct a simple graphical summary of the relative skill of one model versus another. In order to measure the cumulative statistical significance of differences observed in such a plot, we develop a new significance test for the skill differences aggregated across time based on a novel large sample result for estimating continuous loss difference curves.

For the purpose of demonstrating these methods and evaluating ESPN's forecasts, we introduce a number of "competing" continuously updated forecasting methods for basketball outcome prediction. Some are designed to be "straw men" for the purpose of demonstration, whereas others are based on logistic or probit generalized linear models making use of in-game information such as the score difference. We show that using the proposed methods that ESPN's model is generally well-calibrated, and exhibits significantly better skill than some naïve models, although it does not demonstrate superiority over relatively simple logistic regression models based on the score difference and relative team strength alone.

The rest of the chapter is organized as follows. Section 2.2 introduces the details of the ESPN forecasting data that we consider, as well as some competing forecasting methods that we develop and use for the purpose of comparison. Section 2.3 discusses the construction of calibration surfaces for such forecasts, as well as simple graphical summaries of these surfaces. Section 2.4 explains the proposed methods to evaluate the relative skill of two sets of real-time probabilistic forecasts. A Monte Carlo simulation study of these methods is given in the Appendix 2.5. A detailed comparison of the ESPN forecasts as well as those of the proposed models is given in Section 2.6. Technical details are provided in the Appendix A.1.

2.2 Motivating Data and Competing Models

The specific data that we consider are play-by-play records and real-time probabilistic forecasts of NBA regular season games downloaded from `espn.com/nba` (ESPN, 2020). The NBA is a major professional basketball league, which is often referred to as one of the "Big Four" professional sports leagues in North America. Since 2004, except for the lockout season in 2011 and the COVID-19-influenced seasons in 2020 and 2021, the NBA is comprised of 30 teams, with each team playing a schedule of 82 games in the regular season.

Starting in the 2017–18 NBA season, ESPN Analytics began providing real-time in-game probabilistic forecasts of the home team winning for each NBA game played; an example of the forecasts from one game is shown in Figure 2.1. The data available from ESPN are quite rich, including real-time information about details such as substitutions, fouls, and ball possession. We consider here only a subset of these data that includes the real-time probabilistic forecasts provided by ESPN, as well as the evolution of the score throughout the game, for the 2017–18 and 2018–19 seasons. These data are updated each time there is an "event" in the game, which includes primarily score changes, fouls, and changes of possession. A typical game features between 460 to 480 events.

We excluded a small portion of these data from our analysis due to two issues. Quite often, multiple events will occur at the same instant in a game. One of the main examples that contributes to this is multiple players substituting at the same time. Although these events are all logged at the same time point, they occur in the dataset in an ordered sequence. The forecasted probabilities published by ESPN during such an event are typically contingent on this order. Therefore, we simply average the forecasts together in such a scenario to produce a probabilistic forecast at that instant. We also tried a number of other ways to handle this situation, such as using the first or last probabilistic forecast among the events recorded, and the difference in the results was negligible.

The second issue is due to games that go to overtime. If two teams' scores are tied at the end of the 48-minute regulation game time, the teams will play an extra five-minute overtime period. For such games we remove the overtime period from the analysis, and only consider probabilistic forecasts up to the end of the game so that they are comparable

Table 2.1: Summary of the data obtained from ESPN (ESPN, 2020) from the 2017–18 and 2018–19 NBA regular seasons. Raw counts represent the total number of games that ESPN provides probability forecasts for. Selected refers to those games that do not contain errors or missing values. Processed represents the data after averaging out multiple events recorded at the same game time.

Season	Mode	Games	Events	Max. events	Min. events	Avg. # of events
17–18	Raw	1158	530032	606	234	457.7133
	Selected	1137	517983	572	240	455.5699
	Processed	1137	354749	375	173	312.0343
18–19	Raw	1229	583443	700	124	474.7299
	Selected	1213	572546	598	366	472.0082
	Processed	1213	396991	385	241	327.2803

to those that arise from games that did not go to overtime. Overtime games represent slightly less than 10% of the total games. Additionally, a small number of data points were discarded due to evident defects or excessive missing values.

The remaining data that we analyze are summarized in Table 2.1, and in each season there are more than 1100 games with a total of over 350,000 play-by-play records available. Below we use the data from the 2017–18 season as training data for our own models, and then we produce and evaluate forecasts for the 2018–19 season.

Letting N denote the total number of games with forecasts that we wish to evaluate, so $N = 1213$ when we consider the 2018–19 forecasts, the data may then be denoted as $\hat{p}_i^{ESPN}(t)$, $1 \leq i \leq N$, $t \in [0, 1]$ representing the probabilistic forecasts of the home team winning in the $i'th$ game at intra-game time t. We assume that the game time parameter t is normalized to be between zero and one so that it represents the proportion of the game completed. Although these forecasts are only available when events occur, due to the fact that events are very dense throughout the game, we complete these forecasts as a piecewise constant function set to the last probability forecast in-between events to produce full probability forecast curves over $[0, 1]$, which also makes them more comparable from one game to the next. The high frequency of the events also helps minimize bias from piecewise

constant interpolation. This is illustrated in Figure 2.1. We also consider the data $H_i(t)$ and $A_i(t)$, $1 \leq i \leq N$, $t \in [0,1]$, denoting the home team score and away team score in the $i'th$ game, respectively, at proportion t of the game. In our analysis below we frequently make use of the score difference $ScD_i(t) = H_i(t) - A_i(t)$, $1 \leq i \leq N$, $t \in [0,1]$.

The goal of the methods that we develop below is to evaluate the quality of the forecasts $\hat{p}_i^{ESPN}(t)$, $1 \leq i \leq N$, $t \in [0,1]$. To do this we also develop a number of benchmark models that are used for the purpose of comparison.

2.2.1 Benchmark models for predicting NBA game outcomes

We use the following notations below. We let Y_i denote the indicator random variable that the home team wins in the i-th game, so that $Y_i = 1$ if the home team wins the i-th game, and $Y_i = 0$ if the home team loses the i-th game. We are interested in forecasting or estimating the probability $p_i(t)$ that the home team wins, given the information up to time t in the game, so that

$$p_i(t) = \mathbb{P}(Y_i = 1 \mid \text{all information up to time } t \text{ in game } i).$$

A more formal definition of $p_i(t)$ is given in Appendix A. $\hat{p}_i^{ESPN}(t)$ is in principle an estimate (forecast) of $p_i(t)$. In order to evaluate the quality of these forecasts, we consider a number of competing benchmark models, progressing from naïve to more realistic ones. The benchmark models that we consider are for the most part generalized linear models (GLMs) for binary response data, which are often termed logistic regression models. Using $g(\cdot)$ to denote the GLM link function, which we take to be the logit link, see e.g. Chapter 4 of McCullagh and Nelder (1989), each of our models is of the form

$$g(p_i(t)) = \beta_0(t) + \sum_{j=1}^{d} \beta_j(t) X_j(t),$$

where the terms $X_j(t)$ denote covariates that are used to predict $p_i(t)$. All GLMs were fit pointwise over a discrete grid of 721 equally spaced time points t, which corresponds to a 4 second in-game resolution, using the `R` programming language, specifically the `glm`

function in the stats package, version 4.0.2. For each model we used the 2017–18 season data as training data, and then produced rolling forecasts on the 2018–19 season data to compare to the ESPN forecasts.

The most complicated covariates that we consider to build these benchmark models are the score difference $ScD_i(t)$, and some measure of the relative strength of the teams, which we term RS_i. We also consider the "leading status" covariate, which encodes whether or not the home team is winning at intra-game time t. We define this as the covariate $LS_i(t)$, which takes the value 1 if $ScD_i(t) > 0$, -1 if $ScD_i(t) < 0$, and 0 if $ScD_i(t) = 0$.

In terms of defining RS_i, there is a number of ways of evaluating the relative strength of teams, including using the Elo rating system (Elo, 1978), which has been extensively used to rate the strength of basketball teams (see Silver 2014; Silver and Fischer-Baum 2015; Silver et al. 2019) and determine odds in betting markets. We instead use as a proxy of the relative team $RS_i = \hat{p}_i^{ESPN}(0)$, the pre-game probability of the home team winning as forecast by ESPN. We considered a number of alternate metrics to define RS_i, and found that generally the results and conclusions of the analyses did not change materially, and so we use this quantity to avoid the need to introduce new metrics.

Descriptions of the basic benchmark models that we considered to forecast $p_i(t)$ are collected in Table 2.2, and are listed in order from naïve to more realistic ones. We note that the GLMs with intercept terms are able to implicitly model the home team advantage, which refers to the phenomenon that in the NBA the home team tends to win a higher percentage of games than the away team, and so a model based on the score difference but without an intercept term, like **ScDnoInt** in Table 2.2, could be expected to be poorly calibrated, at least at the beginning of the game. As mentioned, each GLM model is fit pointwise over the game-time parameter t. This allows that, for example, the effect, as determined by the models, of relative team strength, home team advantage, and score difference can evolve throughout the game. Diagnostic plots of the pseudo R^2 (McFadden, 1973) and relative variable importance over the course of the game of the variables in our "least naïve" model, which uses both the score difference and relative team strength and is termed **PgRSScD**, are displayed in Figure 2.2. It is clear from this figure that the models' predictions improve as the game progresses, evidently since ultimately the score difference covariate determines the winner, and also that the relative importance

of the score difference versus team strength changes inversely as the game progresses; relative team strength is the most important predictor early in the game, but becomes less important later in the game as the score difference becomes more informative.

Table 2.2: Descriptions of benchmark models used to forecast $\hat{p}_i(t)$, listed from more naïve, to more realistic models.

Model	Covariates	Description
CF	None	A constant forecast for all intra-game times of 1/2, i.e. $\hat{p}_i(t) = 0.5$ for all $t \in [0, 1]$. Abbreviation is for **C**oin **F**lip.
HomeWP	None	A constant forecast for all intra-game times set to the observed home team win rate in the prior 10 NBA regular season, spanning 2008–17. This amounts to forecasting $\hat{p}_i(t) = 0.593$ for all $t \in [0, 1]$.
PgRS	RS_i	GLM for the home team win probability in terms of **P**re**g**ame **R**elative team **S**trength, as measured by $\hat{p}^{ESPN}(0)$.
LS	$LS_i(t)$	GLM for the home team win probability in terms of the **L**eading **S**tatus for the home team.
ScDnoInt	$ScD_i(t)$	GLM for the home team win probability in terms of the **Sc**ore **D**ifference between the home and away teams. **No Int**ercept term is included in the model.
ScD	$ScD_i(t)$	GLM for the home team win probability in terms of the **Sc**ore **D**ifference between the home and away teams.
PgRSLS	RS_i, $LS_i(t)$	GLM for the home team win probability in terms of **p**re**g**ame **R**elative team **S**trength as measured by $\hat{p}^{ESPN}(0)$, and the **L**eading **S**tatus.
PgRSScD	RS_i, $ScD_i(t)$	GLM for the home team win probability in terms of **p**re**g**ame **R**elative team **S**trength as measured by $\hat{p}^{ESPN}(0)$, and the **Sc**ore **D**ifference.

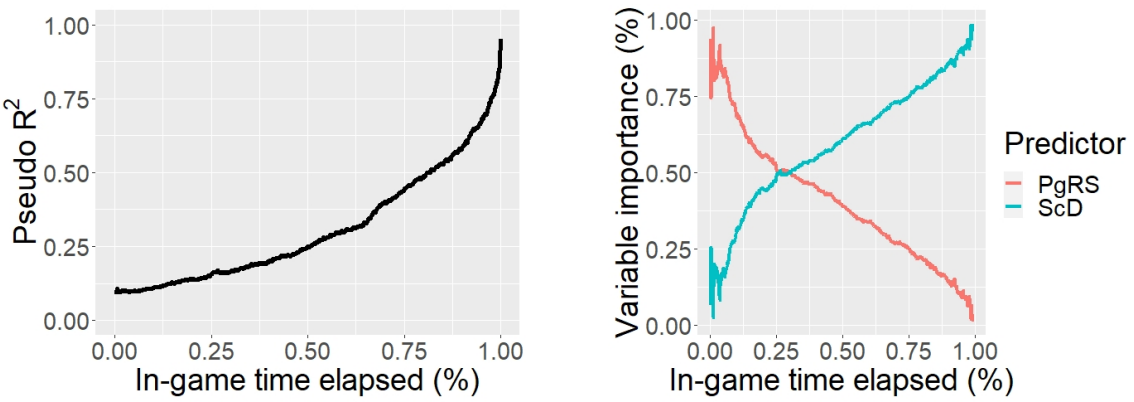


Figure 2.2: **Left**: the pseudo R^2 of the logistic regression model **PgRSScD**, which uses the covariates pre-game relative strength and score difference, as a function of the game time. **Right**: the variable importance of each covariate as it contributes to the pseudo R^2.

2.3 Evaluating Calibration of Continuous Forecasts

We now turn to the task of evaluating the calibration for a given set of continuously updated forecasts $\hat{p}_i(t)$ with realized outcomes Y_i, $i = 1, \ldots, N$. As mentioned in the introduction, traditionally when evaluating such forecasts one often considers what are called *calibration plots* or *reliability diagrams*. A calibration plot is a plot of binned probabilistic forecasts against the conditional event frequency associated with forecasts in a given bin. Since well-calibrated forecasts should have that the event frequency matches the forecast probability, calibration may then be measured by comparing these points against a 45-degree diagonal reference line. Large departures from this line thus indicate poor calibration (cf. Dawid, 1986, Murphy and Winkler, 1992 and Ranjan and Gneiting, 2010). We refer the reader to Gneiting and Katzfuss (2014) and the references therein for a more comprehensive discussion.

One obvious method then to check for calibration of continuously updated forecasts is to produce a calibration plot for each $t \in [0, 1]$ based on the pairs $(\hat{p}_i(t), Y_i)$. While this is in essence what we propose, there are two main challenges in doing so. First traditionally when producing calibration plots, the forecasted probabilities are binned into

fixed bins, commonly by deciles. For example, often the event frequency corresponding to all forecasted probabilities between $[0, 0.1)$ are compared to 0.05, similarly for $[0.1, 0.2)$ to 0.15, and so on. With continuously updated forecasts, and as with the ESPN forecasts, it is typical that the forecasts fluctuate so that for certain time points t the forecasts cluster around some fixed values, and are hence far from being uniformly distributed in such fixed bins. In the case of the ESPN forecasts, near the end of game the majority of forecasts are clustered around 0 and 1. Fixed bins will often have the problem that the forecasts within them are not uniformly distributed within the bin. Second having constructed calibration plots for each t, one must examine a large number of such plots to pinpoint whether a method appears to be well-calibrated, or to diagnose whether there is a subset of times t at which the method is more or less calibrated than others. A simple summary of the many calibration plots produced would therefore be useful.

In order to address the first point, we propose to use adaptive bins in constructing the calibration plots. Specifically, for each t, suppose that we wish to construct M bins for the forecasts $\hat{p}_i(t)$. By calculating the ranked forecasts $\hat{p}_{(i)}(t)$, $i = 1, \ldots, N$, we may group them into M bins so that $\hat{p}_{(i)}(t)$ is in bin j if $(\lfloor N/M \rfloor (j-1) + 1) \leq i < \lfloor N/M \rfloor j$. We denote the collection of $\hat{p}_i(t)'s$ in the j'th bin as Bin_j. In simpler terms, the forecasts at a given time t are grouped into M bins based on their rank. As a reference point or summary of the forecasts in the j'th bin, we use $\tilde{p}_j(t) = \text{Median}\left(\hat{p}_{(i)}(t),\ \ (\lfloor N/M \rfloor (j-1) + 1) \leq i < \lfloor N/M \rfloor j\right)$. A calibration plot at time t is constructed then by comparing $\tilde{p}_j(t)$ to $\bar{Y}_j(t) = Average(Y_i, \text{ such that } \hat{p}_i(t) \in Bin_j)$. Letting n_j denote the number of forecasts in Bin_j, a 95% confidence interval for the mean of the events in Bin_j is constructed as

$$\frac{n_j \bar{Y}_j(t) + \kappa^2/2}{n_j + \kappa^2} \pm \frac{\kappa n_j^{1/2}}{n_j + \kappa^2} \left(\bar{Y}_j(t)(1 - \bar{Y}_j(t)) + \kappa^2/(4n_j)\right)^{1/2},$$

where $\kappa = z_{\alpha/(2M)}$, and z_β denotes a β quantile of the standard normal distribution. Here we have applied a Bonferroni correction to the significance level according to the number of bins used M. α is typically taken to be 5% so that 95% confidence intervals are computed. This interval is often referred to as Wilson interval; see Brown et al. (2001) and Wilson (1927). We have noticed significant improvements by using the Wilson interval in this setting due to event frequencies that approach zero and one near the end of the game,

and so we use these intervals in calibration plot construction below. Additionally, so that no bins contain predominately zero or one probability forecasts, we discard all forecasts to produce a calibration plot at a given t, and corresponding events, if $\hat{p}_i(t) < 0.005$ or $\hat{p}_i(t) > 0.995$, and we analyze those forecasts for calibration separately; those results are summarized in Table 2.3.

Table 2.3: The number of games in which a probability forecast exceeded 0.995 or was less than 0.005 at some point for each of the ESPN, PgRSScD with a logit link, and the ScDnoInt with logit link forecasts, as well as the number and proportion among these games in which the home team won.

	ESPN		PgRSScD		ScDnoInt	
$\hat{p}_i(t)$ for some t	> 0.995	< 0.005	> 0.995	< 0.005	> 0.995	< 0.005
Total Games	555	343	591	358	589	374
Home team wins	553	1	588	0	587	0
Proportion	0.9964	0.0029	0.9949	0.0000	0.9966	0.0000

A calibration plot may then be made by plotting the bin references $\tilde{p}_j(t)$ against the above confidence intervals, and comparing these intervals to the reference line $y = x$. The method is deemed well-calibrated at a given significance level if the reference line generally goes through each interval. An example of this is shown based on the ESPN forecasts for $t = 0.5$ using $M = 10$ bins in the left hand panel of Figure 2.3a at the 95% confidence level. By linearly interpolating the upper bounds of these intervals across both the reference proportions $\tilde{p}_j(t)$ as well as t, we may construct an "upper calibration surface", $U_{1-\alpha}(t, p)$. The upper 95% calibration surface $U_{0.95}(t, p)$ is displayed in the right hand panel of Figure 2.3b for the ESPN forecasts from the 2018–19 season. A lower surface $L_{1-\alpha}(t, p)$ may be similarly constructed by linearly interpolating the lower bounds. A continuously updated forecasting method may then be deemed well-calibrated at a given significance level if the reference plane $f(t, p) = p$, for $t, p \in [0, 1]$, is contained between both surfaces.

Before proceeding, we note that an interesting alternative that we implemented, with results presented in Appendix A.2, is to choose the bins, both their locations and number, adaptively based on an isotonic regression as discussed in Dimitriadis et al. (2021). In our

application to the ESPN forecasts this leads to similar results, and so we omit them here.

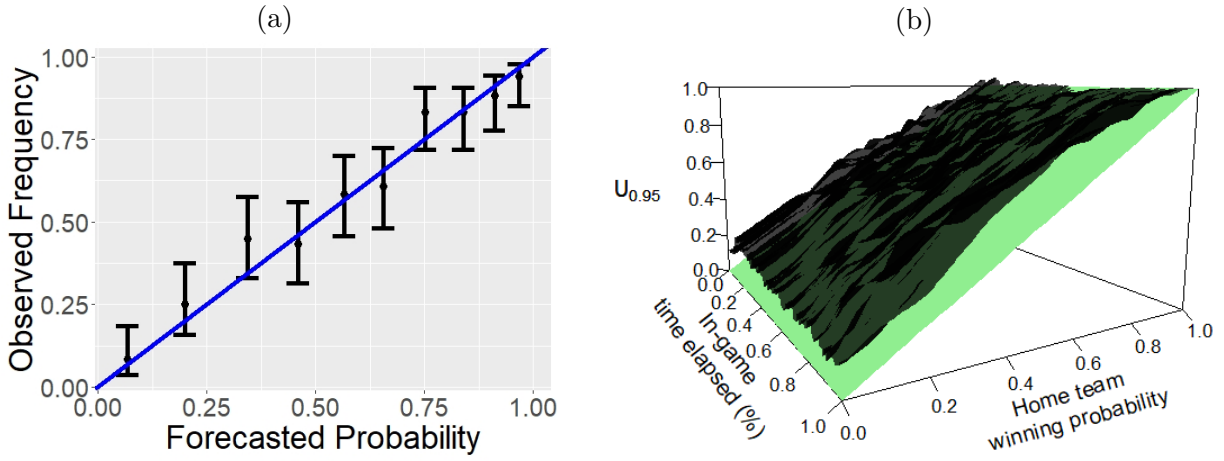


Figure 2.3: **Left**: a calibration plot of the ESPN forecasts at time point $t = 0.5$. **Right**: an upper calibration surface along with reference plane $f(t, p) = p$ for the continuously updated ESPN forecasts obtained by interpolating the upper bounds of each confidence interval of the calibration plots across $t \in [0, 1]$.

Although such surface plots are informative, it can be challenging to infer quickly based on these plots whether a given method appears to be well-calibrated. In order to produce a more easily interpretable summary of such calibration surface plots, we instead consider a plot for each t of the minimum distance over p between the reference plane $f(t, p) = p$, and the upper and lower calibration surfaces. Specifically, we consider plots of the functions $U^{min}_{1-\alpha}(t) = \min_{1\leq j\leq M} U_{1-\alpha}(t, \tilde{p}_j(t)) - \tilde{p}_j(t)$ and $L^{max}_{1-\alpha}(t) = \max_{1\leq j\leq M} L_{1-\alpha}(t, \tilde{p}_j(t)) - \tilde{p}_j(t)$ against t. If the upper and lower confidence surfaces contain the reference plane $f(t, p) = p$, then $U^{min}_{1-\alpha}(t)$ should always lie above zero, and $L^{max}_{1-\alpha}(t)$ should always lie below zero. Points with respect to t at which this does not hold can be used to identify times for which a given method does not appear to be well-calibrated. We note here that in this case, due to the high degree of fluctuations in continuously updated probabilistic forecasts for basketball prediction, we have found it also useful to aid in the interpretation of these plots to smooth them with respect to t using a simple moving average smoother over 5% of the game times.

These summary plots calculated based on the ESPN forecasts as well as from the benchmark models **HomeWP**, **ScDnoInt**, and **PgRSScD** are shown in Figure 2.4; see Table 2.2 for a description of these models. From these we see that the naïve model **HomeWP**, which predicts that the home team will win each game at all times t with the prior 10-year historic win rate of home teams in the NBA, is well-calibrated, as expected. Similar plots (not shown) for the method **CF**, which simply predicts that the home team will win with probability 50%, show that this method is not well-calibrated. Considering this plot for the method **ScDnoInt** (Panel (b) in Figure 2.4), we see that the forecasts are poorly calibrated at the beginning of the game, but calibration improves towards the end of the game. This is expected since the corresponding logit model is taken to be free of an intercept term, and so the model is unable to capture the home team advantage which should force the forecast probabilities to favour the home team at the beginning of the game. Both the ESPN forecasts and those from the model **PgRSScD**, which incorporates team strength as well as the score difference, demonstrated generally good calibration for all game times.

A summary of the outcomes corresponding to games in which a probability forecast at some intra-game time exceeded 0.995 or was less than 0.005 is given in Table 2.3 for the ESPN forecasts as well as for the forecasts based on the models **PgRSScD** and **ScDnoInt**. The empirical home team win rate closely matched these thresholds in each case suggesting that the forecasts for each of these methods are reasonably well-calibrated at these extremal probability forecast levels.

2.4 Evaluating Skill of Continuously Updated Forecasts

In this section we consider methods to produce pointwise confidence intervals for the difference in skill, measured by the Brier score, between two competing methods as a function of the intra-game parameter t, which can be used to identify intra-game times where one method appears to perform significantly better than a benchmark. We also introduce a method to evaluate the statistical significance of differences in skill aggregated across all

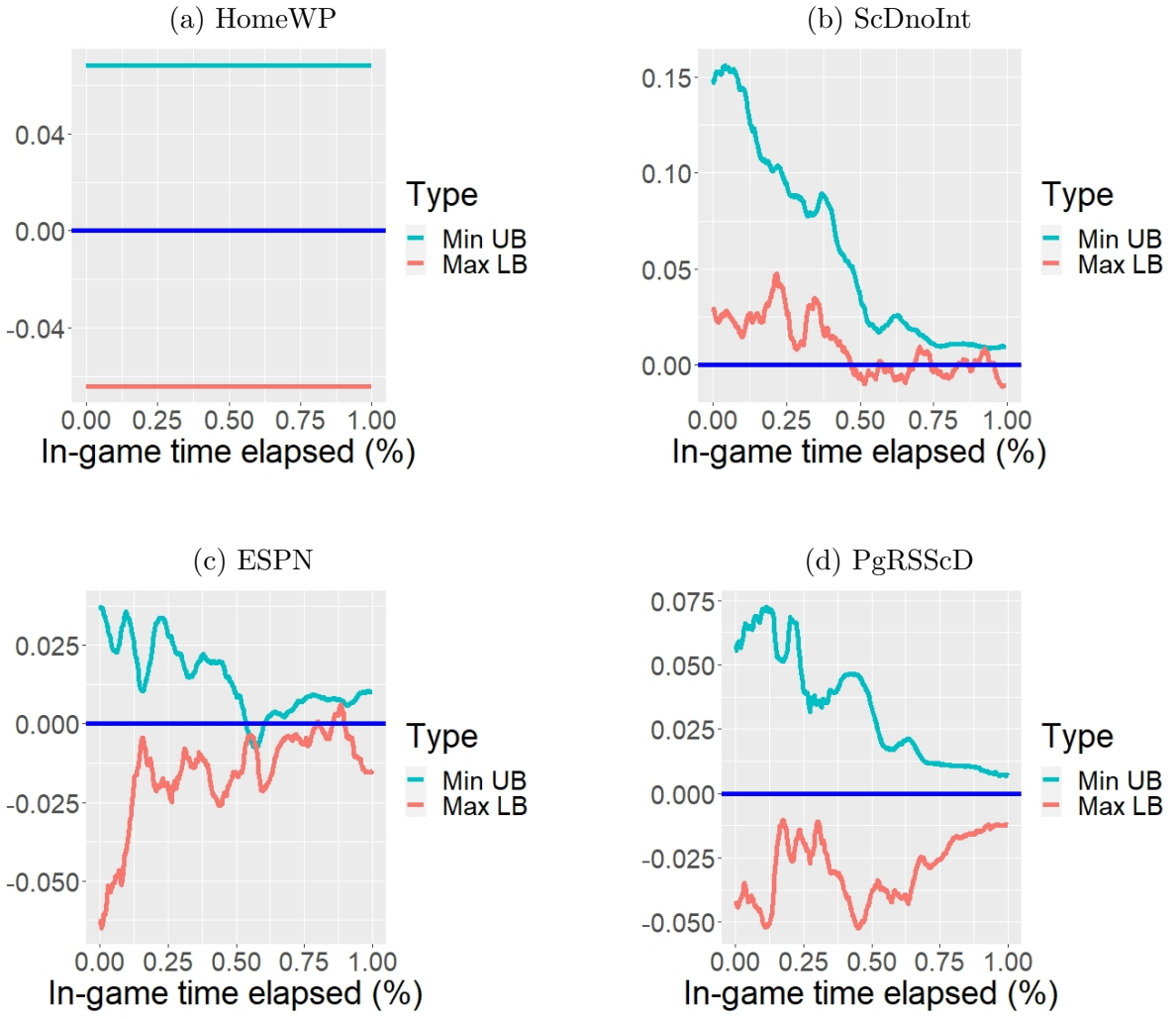


Figure 2.4: Plots of $U_{0.95}^{min}(t)$ and $L_{0.95}^{max}(t)$ against t based on $M = 10$ bins: Methods (a) historic home team win probability (**HomeWP**); (b) GLM using Score difference without intercept (**ScDnoInt**); (c) **ESPN** forecasts; (d) GLM using Pre-game relative strength and Score difference (**PgRSScD**).

intra-game times.

2.4.1 Pointwise confidence intervals measuring the skill difference between competing methods

As described in the introduction, the skill of a probabilistic forecasting method generally refers to its acuity relative to a competing or benchmark method. Formally this can be measured by defining a loss function, or a scoring rule, used to measure the accuracy of a given probabilistic forecast based on the realized events. Perhaps the most frequently used loss function is $L(a, b) = (a - b)^2$ and defines the Brier score (Brier, 1950). We use this loss function below, but the following results generalize to any loss function that has a linear equivalent, which means that (1) $L'(x, b)$ is linear in x, and (2) $L'(x, b) - L(x, a)$ does not depend on x. This includes a Kullback-Leibler Divergence (Kullback and Leibler, 1951), and a Good's log-score, among others; see Bickel (2007) and Lai et al. (2011).

Following the work of Lai et al. (2011), ideally any method to forecast $p_i(t)$ would satisfy the condiitons that $L(p_i(t), \hat{p}_i(t))$ is small, and further when averaged over all forecasts would minimize

$$L_N(t) = \frac{1}{N} \sum_{i=1}^{N} L\left(p_i(t), \hat{p}_i(t)\right). \tag{2.1}$$

Since the underlying true probabilities $p_i(t)$ are unobservable, a sensible estimate of $L_N(t)$ is obtained by replacing these probabilities with their point estimates based on the realizations Y_i to produce

$$\hat{L}_N(t) = \frac{1}{N} \sum_{i=1}^{N} L(Y_i, \hat{p}_i(t)). \tag{2.2}$$

The quantity $\hat{L}_N(t)$ captures the skill of the forecasting method at a given time point t as described above, since a given forecasting method is ascribed generally lower losses, or higher scores, if $\hat{p}_i(t)$ is closer Y_i, the latter of which takes the values 0 and 1.

Suppose that we wish to compare two methods, call them method A and method B, respectively, for producing continuously updated probabilistic forecasts. We denote such forecasts by $\hat{p}_i^A(t)$ and $\hat{p}_i^B(t)$, and we compare them based on the corresponding realized

events Y_i, $1 \leq i \leq N$. This can be done by comparing their average losses defined in (2.2). Specifically, we consider a function of t

$$\hat{\Delta}_N(t) = \frac{1}{N} \sum_{i=1}^{N} \left[L(Y_i, \hat{p}_i^A(t)) - L(Y_i, \hat{p}_i^B(t)) \right], \tag{2.3}$$

which can be viewed as an estimate of the true loss difference

$$\Delta_N(t) = \frac{1}{N} \sum_{i=1}^{N} \left[L(p_i(t), \hat{p}_i^A(t)) - L(p_i(t), \hat{p}_i^B(t)) \right]. \tag{2.4}$$

Larger than zero values of $\hat{\Delta}_N(t)$ favour method B at the given t, whereas negative values show favour for method A. In order to measure the statistical significance of any deviations of $\hat{\Delta}_N(t)$ from zero, we may view $\hat{\Delta}_N(t)$ as an estimator for $\Delta_N(t)$. By constructing suitable confidence intervals for $\Delta_N(t)$ based on $\hat{\Delta}_N(t)$, we may evaluate whether observed deviations suggest the superiority of one model over another, and further construct simple graphical summaries that illustrate the skill of one model compared to another as a function of t. To construct such confidence intervals, we first define the variance of $\hat{\Delta}_N(t)$ as

$$s_N^2(t) = \frac{1}{N} \sum_{i=1}^{N} \delta_i^2(t) p_i(t)(1 - p_i(t)), \tag{2.5}$$

where $\delta_i(t) = \left[L(1, \hat{p}_i^A(t)) - L(0, \hat{p}_i^A(t)) \right] - \left[L(1, \hat{p}_i^B(t)) - L(0, \hat{p}_i^B(t)) \right]$. The following result is proven in Lai et al. (2011) and stated for each $t \in [0, 1]$.

Theorem 2, Lai et al. (2011): Suppose that for each $t \in [0, 1]$ that $s_N^2(t)$ converges in probability to a positive constant as $N \to \infty$, and that each of the variables $A_i(t) = L(Y_i, \hat{p}_i^A(t)) - L(p_i(t), \hat{p}_i^A(t))$ and $B_i(t) = L(Y_i, \hat{p}_i^B(t)) - L(p_i(t), \hat{p}_i^B(t))$ forms martingale difference sequences. Then, for each t,

$$\frac{\hat{\Delta}_N(t) - \Delta_N(t)}{s_N(t)} \xrightarrow{D} \mathcal{N}(0, 1),$$

where $\xrightarrow{D}$ denotes convergence in distribution, and $\mathcal{N}(0, 1)$ denotes a standard normal random variable.

We note that the above result is similar to the main asymptotic result underpinning tests for equivalent predictive accuracy discussed in Diebold and Mariano (2002), see also Gneiting and Katzfuss (2014). The two main conditions of the above theorem are that (1) $s_N^2(t)$, the variance of $\Delta_N(t)$, should for large N behave like a positive constant, and (2) that the forecast loss differences should behave like martingale difference sequences. The first condition can be thought of as a non-degeneracy condition – this result only holds if the forecasts of the two methods to be compared do not coincide entirely. It is not valid for two methods that produce equivalent, or almost equivalent, forecasts. This is, in general, a reasonable assumption if the forecasts being compared are from entirely different models, or if one or both sets of forecasts to be compared come from unknown models, as with the ESPN forecast data, since it is unlikely in this case that they will produce forecasts that coincide. This assumption is in question when comparing forecasts of nested models, e.g. comparing two GLM models whose only difference is that a covariate is included or excluded. A more in-depth discussion of comparing forecasts from nested models, and the problems that arise from it, may be found in Clark and McCracken (2015). Regarding the second condition, this is almost always satisfied when using a loss function with a linear equivalent and constructing genuine forecasting methods that must be based on available (past) information, rather than information from the unknown future, as discussed on page 2361 of Lai et al. (2011) (see also page 356 of Seillier-Moiseiwitsch and Dawid, 1993).

The above results suggest constructing a $100(1-\alpha)\%$ confidence interval for $\Delta_N(t)$ as

$$\hat{\Delta}_N(t) \pm z_{1-\alpha/2}\frac{s_N(t)}{\sqrt{N}}. \tag{2.6}$$

Note that since $p_i(t)(1-p_i(t))$ in the definition of $s_N^2(t)$ is unobserved, we may replace it with the upper bound of $1/4$ in both (2.5) and (2.6) to obtain a conservative confidence interval. Plots as a function of t of $\hat{\Delta}_N(t)$ and the corresponding conservative confidence intervals can be used to evaluate the relative skill of one model compared to another. Points t at which the associated confidence interval $1-\alpha$ confidence interval for $\Delta_N(t)$ do not contain zero indicate a significant improvement at the level α of the average loss of one method over another. Examples of plots of this form may be found in Figure 2.7, which we will discuss in more detail in Section 2.6. We note again that due to the high degree of

fluctuations in continuously updated probabilistic forecasts for basketball prediction, we have found it also useful to aid in interpretation of these plots to smooth them with respect to t using a simple moving average smoother over 5% of the game times.

2.4.2 Functional tests to measure skill aggregated across t

Although the above confidence intervals can be used to evaluate whether two methods exhibit similar or significantly different skill at any given time point t, it is often also of interest to evaluate whether two continuously updated models have approximately equal predictive power when discrepancies between them are aggregated across $t \in [0, 1]$. For example, it might be that one method exhibits similar but somewhat better skill at each game time that when viewed in aggregate suggests the superiority of one model over another. Conversely, it is also possible one method exhibits apparently improved performance at a single game time t_0, although when viewed in aggregate across $t \in [0, 1]$ this improvement may appear rather insignificant.

To make this more precise, we formulate the null hypothesis of equal predictive power aggregated across $t \in [0, 1]$ of two methods as

$$H_0 : \ \|\Delta_N\|^2 = 0,$$

where $\|\cdot\|^2$ is a standard squared L^2 norm of a function, so that $\|f\|^2 = \int_0^1 f^2(t)dt$. The hypothesis H_0 posits then that the two methods to be compared exhibit approximately on average (across t) equal skill. Let

$$Z_N(t) = \sqrt{N}\hat{\Delta}_N(t).$$

A measure of global discrepancy across time between the two forecasting methods may be obtained by considering $\|Z_N\|^2$. In order to determine the large-sample properties of Z_N that would inform determining appropriate significance levels and estimating p-values for tests of H_0 based on $\|Z_N\|^2$, we make use of the following result, which we state rigorously and prove in the Appendix A.1.

Theorem 2.4.1 *Under H_0 and conditions analogous to those of Theorem 2 of Lai et al. (2011), see Appendix A.1, there exists an infinite sequence of constants $\{\lambda_i,\ i \geq 1\}$ that satisfies*

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq 0, \text{ and } \sum_{i=1}^{\infty} \lambda_i < \infty,$$

so that

$$\|Z_N\|^2 \xrightarrow{D} \sum_{i=1}^{\infty} \lambda_i \chi_i^2(1),$$

where $\chi_i^2(1)$, $i = 1, 2, \ldots$ are independent and identically distributed χ^2 random variables with one degree of freedom. Moreover, the constants $\{\lambda_i,\ i \geq 1\}$ can be conservatively estimated by the eigenvalues of the function

$$\hat{C}_{cons}(t,s) = \frac{1}{N} \sum_{i=1}^{N} [\hat{p}_i^A(t) - \hat{p}_i^B(t)][\hat{p}_i^A(s) - \hat{p}_i^B(s)].$$

That is, with $\{\hat{\lambda}_i,\ i = 1, \ldots, N\}$ defined so that $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_N \geq 0$ and satisfying the condition that there exist functions $\hat{\phi}_i(t)$, $i = 1, \ldots, N$, $t \in [0,1]$ with $\|\hat{\phi}_i\|^2 = 1$, such that

$$\hat{\lambda}_i \hat{\phi}_i(t) = \int_0^1 \hat{C}_{cons}(t,s) \hat{\phi}_i(s) ds, \tag{2.7}$$

then for any fixed $j \geq 1$, $\mathbb{P}(\hat{\lambda}_j \geq \lambda_j) \to 1$ as $N \to \infty$.

This result suggests a simple way of conducting an approximate and conservative test of the hypothesis H_0.

Step 1: Evaluate $\|Z_N\|^2$.

Step 2: Estimate $\hat{C}_{cons}$, and the eigenvalues satisfying (2.7).

Step 3: Estimate the distribution of the random variable $Q_D = \sum_{i=1}^{D} \hat{\lambda}_i \chi_i^2(1)$ where D is a large number (below we take $D = 10$ and have found this choice generally adequate). This can be done easily with a Monte-Carlo simulation, or using a numerical method of Imhof (1961).

Step 4: Calculate an approximate and conservative p-value of the test of H_0 as $p = \mathbb{P}(Q_D \geq \|Z_N\|^2)$.

This p-value combined with the confidence intervals in (2.6) allow for a detailed evaluation, both at particular game times t and across all $t \in [0, 1]$, of the relative skill of competing continuously updated methods.

We studied this test as well as the pointwise confidence intervals introduced in the next Section in which an extensive simulation study is performed on synthetic basketball game data, and found that both methods performed generally as expected, and exhibited power to differentiate models with differing skills levels.

2.5 Simulation Studies

In this section we present the setup and results of a simulation study to investigate the performance of the methods proposed in the main document, and illustrate their application in a controlled setting.

2.5.1 Basketball game simulation

We generate data that resembles the NBA game data that we have described in Section 2.2 of the thesis. This entails generating random quantities that serve the role of the score difference and the initial relative team strength of the teams to play. Let $\{W_i(t), t \in [0, 1]\}, i = 1, \ldots, N$ be independent standard Brownian motions, where again N is the total number of games. To represent relative team strength, we consider $RS_i \sim a \times \text{Unif}(-1, 1) + c$, where $\text{Unif}(-1, 1)$ denotes a uniform random variable on $[-1, 1]$, and $a, c \in \mathbb{R}$ are constants that we use to calibrate the simulated data. With this defined, we model the score difference as $ScD_i(t) = tRS_i + W_i(t)$, and hence define the indicator variables (analogous to the home team winning) $Y_i = 1$ if $ScD_i(1) > 0$, and $Y_i = 0$ otherwise. Namely, the score difference is modelled as a Brownian motion with drift determined by RS_i; positive RS_i makes it easier for the home team to win, while negative RS_i has the

opposite effect. The constants a and c defining RS_i were selected so that the home team win probability is approximately 59% in order to match the past-10-year historical home team win rate in the NBA. Similar and simple Brownian motion models for NBA scores have been extensively studied; see Stern (1994), Gabel and Redner (2012), and Chen and Fan (2018).

Under these settings, it is relatively straightforward to show that

$$
\begin{aligned}
p_i(t) &= \mathbb{P}\left(Y_i = 1 \mid \text{all information up to time } t \text{ in game } i\right) \\
&= \mathbb{P}(Y_i = 1 \mid F_{i,t}) = \mathbb{E}\left[\mathbb{1}_{\{ScD(1)>0\}} \mid F_{i,t}\right] \\
&= \mathbb{P}(ScD_i(1) > 0 \mid F_{i,t}) \\
&= \mathbb{P}(ScD_i(1) - ScD_i(t) + ScD_i(t) > 0 \mid F_{i,t}) \\
&= \mathbb{P}(RS_i + W_i(1) - [tRs_i + W_i(t)] + ScD_i(t) > 0 \mid F_{i,t}) \\
&= \mathbb{P}((1-t)RS_i + [W_i(1) - W_i(t)] + ScD_i(t) > 0 \mid F_{i,t}) \\
&= \mathbb{P}(W_i(1-t) > -[(1-t)RS_i + ScD_i(t)] \mid F_{i,t}) \\
&= \mathbb{P}(\sqrt{1-t}Z > -[(1-t)RS_i + ScD_i(t)] \mid F_{i,t}) \qquad (W_i(1-t) \sim \mathcal{N}(0, 1-t)) \\
&= \mathbb{P}(Z > -[(1-t)RS_i + ScD_i(t)]/\sqrt{1-t}) \\
&= 1 - \Phi(-[(1-t)RS_i + ScD_i(t)]/\sqrt{1-t}) \\
&= \Phi\left(\frac{(1-t)RS_i + ScD_i(t)}{\sqrt{1-t}}\right),
\end{aligned}
$$

where Φ denotes the standard normal distribution function. We call this true probability function an "**Oracle**" model or probability below.

2.5.2 Models

We first consider the problem of constructing two "models" for this simulation experiment that have equal forecasting power and satisfy the null hypothesis H_0 of equal forecasting accuracy. This is achieved in this setting by perturbing the Oracle probability by independent random noise that has the same distribution in each model to be compared. We considered two such "models" in our simulation experiments:

The **Ora**cle model perturbed by standard **B**rownian **m**otion (**OraBM**):

$$\hat{p}_i(t) = \Phi\left(\frac{ScD_i(t) + RS_i(1-t) + BM_i(t)}{\sqrt{1-t}}\right),$$

where $\{BM_i(t), t \in [0,1]\}, i = 1, \ldots, N$ are independent standard Brownian motions.

The **Ora**cle model perturbed by time homogeneous **O**rnstein–**U**hlenbeck process (**OraOU**):

$$\hat{p}_i(t) = \Phi\left(\frac{ScD_i(t) + RS_i(1-t) + OU_i(t)}{\sqrt{1-t}}\right),$$

where $\{OU_i(t), t \in [0,1]\}, i = 1, \ldots, N$ are independent time-homogeneous standard Ornstein–Uhlenbeck processes. We may generate the $OU_i(t)$ from the standard Brownian motion as follows: $OU_i(t) = \exp(-t/2)BM_i(e^t)$.

We use the notation **OraBM1** and **OraBM2** (similarly **OraOU1** and **OraOU2**) to denote two probabilistic forecasts generated according to the above models with independent noise terms. Since the noise term in both forecasts are equal in distribution, they intuitively have equal predictive power and satisfy H_0. We compare these "models", as well as the GLM models introduced in Section 2.2.1. To fit these GLM models in a way that is similar to our analysis of the ESPN forecast data, we first generate two independent "seasons" following the description above with N games in each season, one of which we refer to as training data, and the other we call the testing data. We fit the competing models introduced in Section 2.2.1 on the training data with the link function $g(\cdot)$ taken to be the Probit link, and then compare the model forecasts on the test data. We note that with the Probit link the model **PgRSScD** is correctly specified in the sense that $\Phi^{-1}(p_i(t))$ is for each fixed t a linear function of RS_i and $ScD_i(t)$.

2.5.3 Results

For each setting of $N = 100$, 250 and 500 and model comparison considered, we simulated data independently 1000 times and applied the test of H_0 described for comparing forecast skill in Section 2.4.2. The results in terms of empirical rejection rates of H_0 are summarized for each comparison made in Table 2.4.

We first compare the oracle models perturbed by noise **OraBM1** versus **OraBM2**, and **OraOU1** versus **OraOU2**. Figure 2.5 gives a representative example of plots of $\hat{\Delta}_{1000}(t)$ as a function of t, with a conservative 95% pointwise confidence interval for $\Delta_{1000}(t)$, across all t, and an approximate p-value of the test of H_0 for these comparisons. We see in this example that the plots show clearly that the two models have approximately equal skill at all time points t, and further that any discrepancies between the models across t are not significant. We observed in this case that the empirical rejection rates of H_0 for this example tended to be close to, though typically slightly below, the nominal levels of 10%, 5%, and 1%, respectively, for each sample size N used to define the size of the training and testing sets considered, which shows that both the asymptotic result that motivates the test of H_0 seems to be reasonably accurate for the sample sizes and examples that we considered, and that the conservative approximations made do not produce a test that is too conservative in practice. We believe that this might be due to the fact that the way that we have generated the synthetic game data leads to probabilities $p_i(t)$ that are close to 1/2, in which case the conservative approximation is ineffectual. We also have tried adjusting the parameters of the RS_i variable in the simulation so that the home team win rate would be closer to 90%, and found that this made the empirical rejection rates somewhat lower, as expected. We also compared the oracle probability forecasts against the models with noise added, and found that generally the test rejected H_0 at a high rate over all significance levels in this case, as expected.

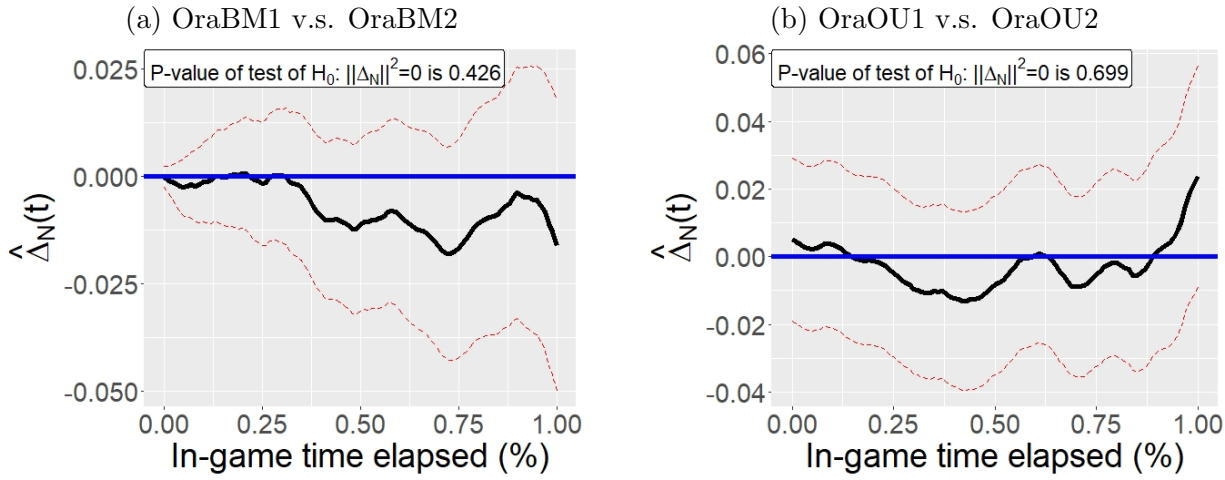


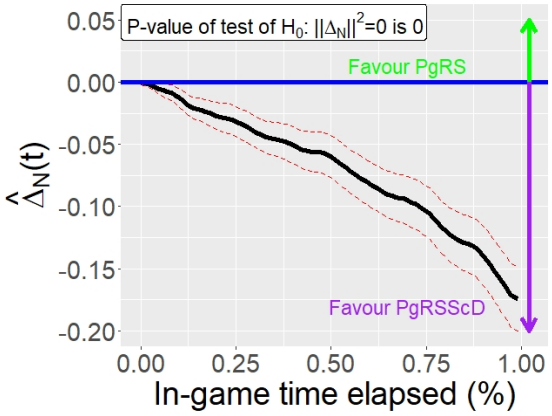
Figure 2.5: Representative plots of $\hat{\Delta}_{1000}(t)$ with 95% confidence intervals and approximate p-values for tests of H_0 for (a) **OraBM1** versus textbfOraBM2; (b) **OraOU1** and **OraOU2**.

In terms of comparing the proposed GLM models, similar representative plots comparing the correctly specified model **PgRSScD** to several naïve competitors are displayed in Figure 2.6. When comparing **PgRSScD** to models that do not adjust their forecasts as the game progresses, our test was always able to distinguish **PgRSScD** as having higher skill. From Figure 2.6, we can see that the expected advantages of **PgRSScD** over the competitor models are transparent in the plot. The relative skill of **PgRSScD** improves over the course of the game compared to models that do not incorporate score information, and for models that do incorporate score information the relative improvement of **PgRSScD** decays as the game progresses. We do see in Table 2.4 that once the sample size reaches 500 the proposed test of H_0 is generally able to distinguish between the correctly specified and naïve models with empirical power approaching one.

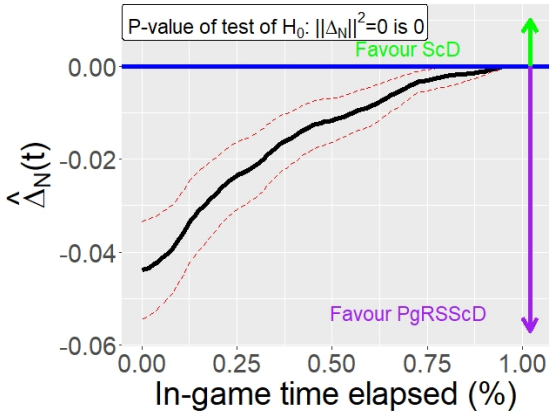
Overall, we found that the proposed test seems to perform well and as expected in many controlled examples, and tends to be conservative. While the test is certainly powerful to differentiate poorly performing models as having low skill compared to competitors, it might struggle to differentiate competitive models without a large sample size ($N \geq 500$ in the examples considered). This along with the graphical tools proposed allows one to easily

identify the relative skill of two competing continuously updated probabilistic forecasting models.

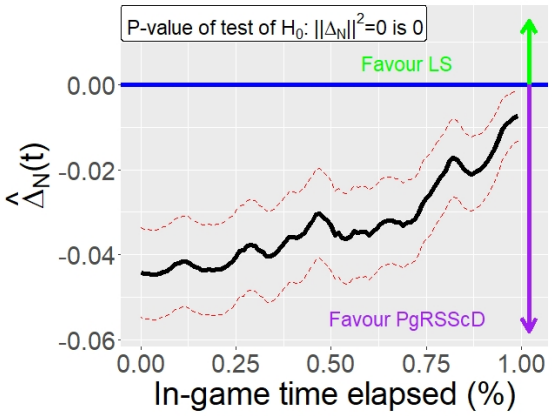
(a) PgRS v.s. PgRSScD



(b) ScD v.s. PgRSScD



(c) LS v.s. PgRSScD



(d) PgRSLS v.s. PgRSScD

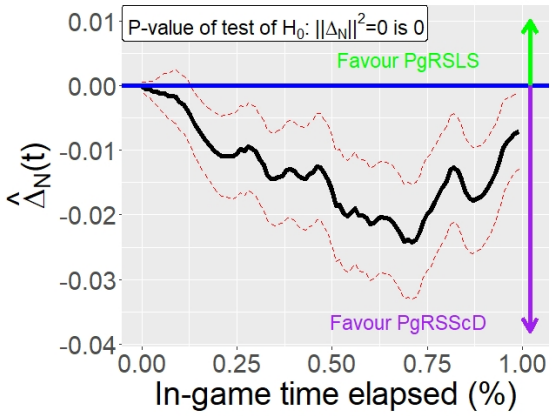


Figure 2.6: Representative plots of $\hat{\Delta}_{1000}(t)$ derived from simulated data with 95% confidence intervals and approximate p-values for tests of H_0 for (a) **PgRSScD** versus **PgRS** (b) **PgRSScD** versus **ScD** (c) **PgRSScD** versus **LS** (d) **PgRSScD** versus **PgRSLS**.

Table 2.4: Empirical rejection rates with nominal levels of 10%, 5% and 1% for the test $H_0 : \|\Delta_N\|^2 = 0$ in 1000 independent simulations.

Competing Models	N=100			N=250			N=500		
	10%	5%	1%	10%	5%	1%	10%	5%	1%
Ora v.s. OraOU	0.997	0.993	0.960	1.000	1.000	1.000	1.000	1.000	1.000
Ora v.s. OraBM	1.000	0.996	0.963	1.000	1.000	1.000	1.000	1.000	1.000
OraOU1 v.s OraOU2	0.096	0.043	0.007	0.085	0.046	0.005	0.083	0.037	0.002
OraBM1 v.s OraBM2	0.072	0.028	0.007	0.081	0.034	0.006	0.089	0.030	0.003
PgRSScD v.s. PgRS	1.000	0.998	0.972	1.000	1.000	1.000	1.000	1.000	1.000
PgRSScD v.s. ScD	0.510	0.377	0.176	0.831	0.745	0.509	0.995	0.982	0.907
PgRSScD v.s. LS	0.795	0.704	0.415	0.990	0.967	0.898	1.000	1.000	1.000
PgRSScD v.s. PgRSLS	0.820	0.648	0.253	1.000	0.999	0.958	1.000	1.000	1.000

2.6 Evaluating the Skill of ESPN Forecasts

In this section, we apply the methods described in Section 2.4.1 to evaluate the skill of ESPN's continuously updated probabilistic forecasts.

Figure 2.7 shows plots $\hat{\Delta}_{1213}(t)$ with conservative 95% confidence intervals, as well as approximate p-values of the test of H_0 for comparisons of the ESPN forecasts with the naïve models **PgRS**, **ScD**, **LS**, and **PgRSLS**. These plots suggest that, in aggregate, the ESPN forecasts significantly outperform these models. The specific points in the game at which the ESPN forecasts exhibit higher skill compared to these benchmarks is also clear in the plots. For the models that use relative team strength as encoded by ESPN's initial home win probability as a covariate, the relative skill is similar to ESPN's forecasts early in the game, and similarly those that make use of the score difference improve relative to the ESPN forecasts towards the end of the game. For example, **ScDnoInt**, the model based on the score difference alone, is strongly outperformed by the ESPN forecasts early in the game, but their forecasts have indistinguishable skill towards the end of the game.

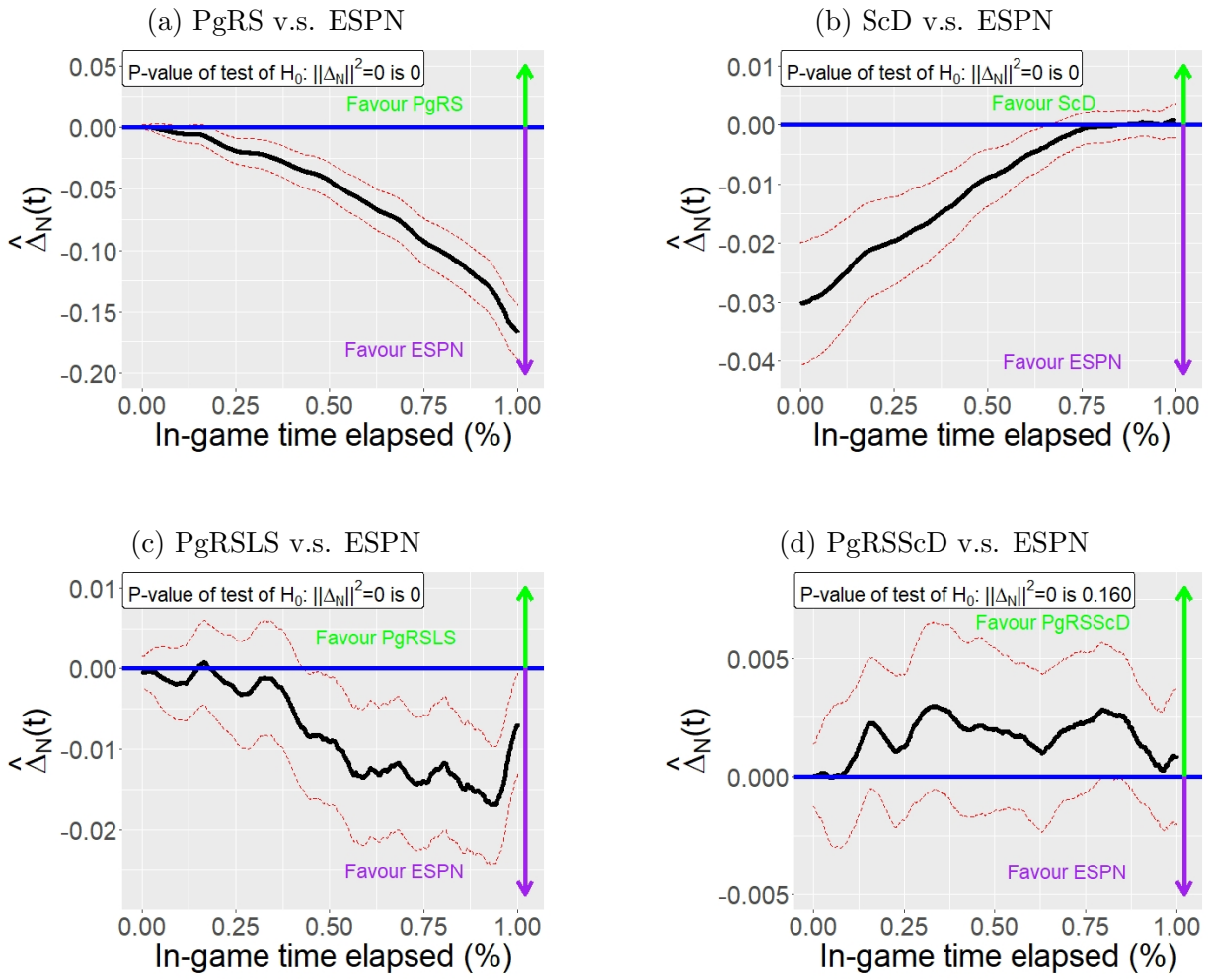


Figure 2.7: Plots of $\hat{\Delta}_{1213}(t)$ based on the 2018–19 season forecasts with 95% confidence intervals and approximate P-values for tests of H_0 for (a) **ESPN** versus **PgRS**; (b) **ESPN** versus **ScD**; (c) **ESPN** versus **PgRSLS**; (d) **ESPN** versus **PgRSScD**.

We also compared the ESPN forecasts to those of the somewhat less naïve model **PgRSScD** using a logit link. Figure 2.7 shows a plot $\hat{\Delta}_{1213}(t)$ based on the 2018–19 season with conservative 95% confidence intervals, as well as approximate P-values of the test of H_0 for this comparison. In absolute terms, the estimated skill as measured by the

Brier score generally favoured the simple logit model **PgRSScD**, with the exception of the last moments in the game. However, we do see from this analysis that the difference is apparently not statistically significant at the 5% level at any game time point based on the conservative confidence interval estimates, nor is it significant when the difference is aggregated across time points. We found it interesting that the ESPN's sophisticated proprietary model, which ostensibly makes use of more nuanced information about the game status and more sophisticated models, did not significantly outperform a simple logistic regression model. One might draw the conclusion based on this that whatever additional information used by ESPN's model in producing these forecasts is not clearly beneficial for the purpose of forecasting, except for perhaps in the final moments of the game.

2.7 Discussion

Motivated by evaluating forecasts of NBA basketball games, we have developed graphical tools and statistical tests for assessing the calibration and relative skill of continuously updated probabilistic forecasts. These were studied via a simulation study of synthetic "basketball games", and applied to evaluating and comparing the forecasts published on ESPN and a number of competing models. In terms of calibration, the ESPN forecasts, as well as forecasts produced from simple logistic regression models using the in-game score difference and/or pre-game relative strength of teams as covariates, appear reasonably well-calibrated. In terms of skill, the ESPN forecasts exhibited significantly higher skill over naïve models, but did not demonstrate superiority over simple logistic regression models based on the score difference and relative team strength.

We conclude with a few remarks about some ideas that we have considered but chose not to include in this chapter, and treat them as avenues for future work. It is noteworthy that the confidence intervals defined in (2.6) may be made narrower and less conservative by using auxiliary information to improve the approximation of replacing $p_i(t)(1-p_i(t))$ with the upper bound $1/4$. We have considered a number of methods to achieve this, including employing auxiliary models and covariates to estimate $p_i(t)(1-p_i(t))$ solely in

the variance estimation step, but found both that this changed the intervals little, and lead to poor performance near the end of the game. For a discussion of similar methods, see Sections 3.4 and 3.5 in Lai et al. (2011).

The basic reason why nested models often cannot be compared by means of the confidence intervals introduced and the test of H_0 derived using Theorem 2.4.1 is that nested models may lead to a covariance kernel $\hat{C}_{cons}$ that is approximately (and asymptotically) degenerate. It might be interesting to adapt this result, and the corresponding intervals and test, to this case, for instance using methods similar to those described in Clark and McCracken (2015).

Chapter 3

Functional Spherical Autocorrelation: a Robust Estimate of the Autocorrelation of a Functional Time Series

3.1 Introduction

Functional data analysis (FDA) as a field has grown considerably over the past three decades, likely owing to increasing interest in analyzing high-dimensional data that arise from continuously observing processes over various domains, including time, space, and frequency. We refer the reader to Ramsay and Silverman (2005) and Horváth and Kokoszka (2012) for textbook length treatments of FDA. In many cases of interest, functional data (FD) are collected sequentially over time. For example, continuous observations of the electricity price in a given region might be interpolated and fashioned into daily electricity price curves. Such series of FD objects are referred to as functional time series (FTS), and methods for analyzing and FTS have also been steadily developed in recent years; see Bosq (2000), Hörmann and Kokoszka (2010), and Chapter 8 of Kokoszka and Reimherr (2017) for an overview of FTS analysis.

When analyzing, or evaluating a model for, any time series, one often begins by computing and plotting a measure of a sample autocorrelation of the series, along with corresponding confidence intervals computed assuming that the series forms a strong white noise. This is usually done in order to evaluate the serial dependence structure of the series or model residuals in order to inform further modelling, or serve as validation that a given time series model appears to fit the data well.

In terms of estimating the serial dependence structure of a functional time series, available methods to date are based on measuring the magnitude of the lagged autocovariance operator of the series. In order to fix ideas, suppose that $\{X_i(t), \quad i \in \mathbb{Z}, \quad t \in [0,1]\}$ is a sequence of stochastic processes whose sample paths are in $L^2[0,1]$, the space of real-valued square-integrable functions defined on the unit interval. For example, $X_i(t)$ might denote the electricity price in a certain region on day i at time t, where t is normalized to lie in $[0,1]$. Conceptually a FTS can be thought of as an observed stretch of such a sequence of length N, $X_1, \ldots, X_N$. As defined in, for example, Panaretos and Tavakoli (2013), when the series $\{X_i(t), \quad i \in \mathbb{Z}, \quad t \in [0,1]\}$ is stationary, the *autocovariance kernel at lag h* of the series is

$$C_h(t,s) = \mathbb{E}\left[\left(X_0(t) - \mathbb{E}X_0(t)\right)\left(X_h(s) - \mathbb{E}X_h(s)\right)\right].$$

Using a standard L^2-norm to measure the magnitude of C_h, a fACF may be defined as the mapping $h \mapsto \rho_h^{(F)}$, where

$$\rho_h^{(F)} = \frac{\|C_h\|}{\int C_0(t,t)dt}, \text{ and } \|C_h\|^2 = \iint C_h^2(t,s)dtds.$$

Mestre et al. (2021) considered this definition of the fACF, and studied its use in performing model selection for functional SARIMAX models. Kokoszka et al. (2017) also considered the properties of this fACF by assuming that the underlying FTS is conditionally heteroscedastic.

Although this functional autocorrelation function (fACF) is useful in many situations, it has several drawbacks. Among them are that $\rho_h^{(F)}$ describes only the magnitude of C_h, and is always non-negative. Any information about the "sign" or "direction" of the correlation in the series is lost in computing $\rho_h^{(F)}$. Additionally, estimating $\rho_h^{(F)}$ is highly prone to the influence of outliers. Although the definition of the fACF only presumes the existence of

two moments of the underlying series, i.e. $\mathbb{E}\|X_i\|^2 < \infty$, in order to perform inference on $\rho_h^{(F)}$, one often requires four moments through the condition that $\mathbb{E}\|X_i\|^4 < \infty$. Worse still, in many cases of interest, for instance in the context of fitting volatility models to FTS, one often wishes to estimate the autocorrelation structure of transformations of the original series. A popular transformation to consider in this setting is X_i^2, and in this case the implicit number of moments required in order for the usual confidence intervals for $\rho_h^{(F)}$ to be valid is reflected in the condition that $\mathbb{E}\|X_i\|^8 < \infty$.

To address some of these shortcomings, we propose a new autocorrelation measure for functional time series that we dub a *spherical autocorrelation function*. It is constructed by computing the average angle, as encoded by the inner product of projections of the series onto the unit sphere in $L^2[0,1]$, between lagged pairs in the series that have been centred by the spatial (geometric) median of the series. This measure both captures "signed" serial dependence in the series, and is also robust in the sense that it does not require high moment conditions to define and estimate. We show that this new autocorrelation function can be consistently estimated when the underlying series is stationary, and satisfies the central limit theorem assuming that the series is a strong white noise.

The main theoretical challenge that we need to overcome in doing so lies in handling the effect of estimating the spatial median, since the natural estimator of the spherical autocorrelation is not easily linearized in terms of the spatial median, and existing spatial median estimators are defined in terms of iterative optimization procedures, and do not have closed-form expressions; see e.g. Gervini (2008) and Cardot et al. (2013). We establish though that under primitive conditions, which are satisfied by existing functional spatial median estimators, the estimator that we propose for the spherical autocorrelation is still consistent and asymptotically Gaussian when the underlying series is a strong white noise.

These results are used to construct asymptotically valid confidence intervals and portmanteau white noise tests for the spherical autocorrelation function. As confirmed in simulation experiments and two real data applications, these confidence intervals and tests appear to perform well in many settings, and are useful in exploring the dependence structure, and performing model selection, with functional time series.

There are a number of papers related to this work that generally address the problems

of estimating the correlation between FD objects, and robust covariance estimation with FD. Leurgans et al. (1993) studied the canonical correlation between paired FD, see also page 28 in Boente and Kudraszow (2022), and Ramsay and Silverman (2005) put forward a robust estimator for the canonical correlation of FD. Boente et al. (2019) develop a robust, spherical covariance operator for FD, which is also based on centring the data using the spatial median.

Dubin and Müller (2005) is perhaps the most related to the present work. They develop a correlation measure between paired FD based on projecting onto the unit sphere referred to as "dynamical correlation". Some drawbacks to their measure are that it is computed by centring the data with the sample mean, which causes the measure to potentially take non-zero values for measuring the correlation between independent FD. Moreover, asymptotic normality of the sample statistic is established only when the mean is either known or constant. Weighted empirical likelihood methods are used to construct confidence intervals for the dynamical correlation in Sang et al. (2019).

The rest of the chapter is organized as follows. Section 3.2 gives a definition of the spherical autocorrelation and its natural estimator. That section also includes statements of the asymptotic properties of the estimator, as well as definitions of confidence intervals and portmanteau tests. The practical implementation of the methods are discussed in Section 3.3, as well as the results of several simulation experiments. Two data applications are provided in Section 3.4, and we close the chapter with concluding remarks in Section 3.5. All proofs and some additional implementation details and simulation results are placed in an Appendix following these main sections.

3.2 Problem Statement and Methodology

We begin by introducing some notation that will be used throughout the chapter. We consider a FTS $\{X_i\}_{i\in\mathbb{Z}} \coloneqq \{X_i(t) : i \in \mathbb{Z},\ t \in \mathcal{T}\}$, where $\mathcal{T}$ is a compact interval, from which we assume that we have observed a stretch of length n, $X_1, \ldots, X_n$. Without loss of generality, we assume that the argument t is re-scaled so that $\mathcal{T} = [0, 1]$. Here each variable X_i is viewed as a stochastic process whose sample paths lie in the Hilbert space

$\mathcal{H} = L^2[0,1]$ of real valued square integrable functions equipped with an inner product

$$\langle f, g \rangle = \int f(t)g(t)dt,$$

and corresponding norm $\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}$, where $\int = \int_0^1$. Although the methods and results below would hold equally well when $\mathbb{H}$ is any separable Hilbert space, we take $\mathcal{H} = L^2[0,1]$ since this both covers the applications that we consider, and also coincides with the space considered for many existing methods for robust estimation of the centre and covariance of FD; see Gervini (2008) and Boente et al. (2019).

3.2.1 Definition of spherical autocorrelation

In the event that the series $\{X_i\}_{i \in Z}$ is strictly stationary, our main goal is to quantify the autocorrelation present in the series using the observations $X_1, \ldots, X_N$. The method that we propose is based on examining the angle between suitably centred, lagged pairs of the series X_i and X_{i+h}. This angle is encoded by the inner product of the projections of these pairs onto the unit sphere. We denote the projection of $x \in \mathbb{H}$ onto the unit sphere as $S(x) := x/\|x\|$ for $x \neq 0$, and $S(0) = 0$. $S(\cdot)$ is often referred to as a *spatial sign*. The centred and projected observations may then be written as

$$S(X_i - \mu) = \frac{X_i - \mu}{\|X_i - \mu\|}, \quad i = 1, \cdots, N, \tag{3.1}$$

where μ is a suitably defined "centre" of the process X_i, which we discuss in detail below. Note that $S(x)$ is bounded as it is easy to see that $\|S(x)\| = 1$. The following defines a new notion of autocorrelation for a FTS.

Definition 3.2.1 Suppose that the series $\{X_i\}_{i \in \mathbb{Z}}$ is strictly stationary. Given a fixed time lag h, we define a *spherical autocorrelation* at lag h as

$$\rho_h^{(\mathrm{S})} = \rho_h = \mathbb{E}\left[\langle S(X_0 - \mu), S(X_h - \mu) \rangle\right]. \tag{3.2}$$

A *functional spherical autocorrelation function* (abbreviated fSACF) is a function $h \mapsto \rho_h^{(S)}$ for $h \in \{0, 1, \ldots\}$.

It is immediate from its definition that the fSACF shares a number of similar features with the classic autocorrelation function of a univariate time series: i) $-1 \leq \rho_h \leq 1$, ii) $\rho_0 = 1$, and iii) the fSACF is symmetric around the origin, i.e., $\rho_h = \rho_{-h}$. Another clear consequence of its definition is that ρ_h is well-defined for any stationary process in $\mathbb{H}$. In particular, no moment conditions are required in defining the fSACF.

In order to maintain its analogy to other classic notions of autocorrelation and aid in its interpretation, it would also be an appealing feature of the fSACF when the variables are independent at lag h, i.e. when X_0 and X_h are independent, $\rho_h = 0$. It turns out that this property depends primarily on how the observations are centred by μ. A natural candidate is to take $\mu(t) = \mathbb{E}X_i(t)$, the mean function of the observations, however this does not lead to a fSACF possessing this property.

What appears instead to be the correct "centre" in order to preserve this property is take μ to be the functional spatial median:

Definition 3.2.2 A *spatial median* (also sometimes referred to as a *geometric median*) μ of a random variable $X \in \mathcal{H}$ is any element of $\mathbb{H}$ satisfying

$$\mu = \arg\min_{y \in \mathcal{H}} \mathbb{E}\left[\|X - y\| - \|X\|\right].$$

See Kemperman (1987), Gervini (2008), Cardot et al. (2013), and Boente et al. (2019). As noted in equation (2) of Gervini (2008), when $\mathbb{P}(X = \mu) = 0$, the definition of the spatial median implies that $\mathbb{E}S(X - \mu) = 0$. As a result when X_0 and X_h are independent and μ is the spatial median, we have by Fubini's theorem that $\rho_h = \langle \mathbb{E}S(X_0 - \mu), \mathbb{E}S(X_h - \mu) \rangle = 0$, so long as $\mathbb{P}(X_i = \mu) = 0$.

Going forward we thus take μ in the definition of ρ_h to be the functional spatial median as in Definition 3.2.2.

3.2.2 Inference for ρ_h assuming μ is known

Although in practice the spatial median μ must be estimated from the sample in order to produce feasible estimators of ρ_h, we think that it is useful to begin the analysis by

considering estimators of ρ_h, and their properties, assuming that μ is known. In the next subsection we will consider the properties of these estimators when μ is replaced with typical estimators of the spatial median. Presenting the results in this order allows us to separate the properties of the underlying process $\{X_i\}_{i\in\mathbb{Z}}$ that are needed to perform inference on ρ_h from those that are required to handle the effect of estimating μ.

Assuming that μ is known, a natural estimator of ρ_h is

$$\tilde{\rho}_h = \frac{1}{N}\sum_{i=1}^{N-h} \langle S(X_i - \mu), S(X_{i+h} - \mu)\rangle \,. \tag{3.3}$$

When analyzing a FTS, we are often interested in evaluating whether the series, perhaps comprised of model residuals, appears to be a white noise, or instead appears to exhibit significant autocorrelation at some lags, which may inform subsequent modelling. As such our main aim is to approximate the distribution of $\tilde{\rho}_h$ if the underlying series forms a strong white noise, i.e., if a sequence of independent and identically distributed FD objects, and also show that $\tilde{\rho}_h$ is consistent for ρ_h if the underlying series is strictly stationary.

If the series $\{X_i\}_{i\in\mathbb{Z}}$ is strictly stationary and ergodic, then $\tilde{\rho}_h \xrightarrow{a.s.} \rho_h$ by the mean ergodic theorem. Our first result then concerns the asymptotic distribution of $\tilde{\rho}_h$ when the underlying series is a strong white noise. Let

$$C_P(t,s) = \mathbb{E}\left\{[S(X_0 - \mu)(t)][S(X_0 - \mu)(s)]\right\}, \quad t,s \in [0,1],$$

denote a covariance kernel of the observations projected onto the unit sphere in $\mathbb{H}$.

Theorem 3.2.1 *Suppose that $\{X_i\}_{i\in\mathbb{Z}}$ is a strong white noise in $\mathcal{H}$ such that $\mathbb{P}(X_0 = \mu) = 0$. Then*

$$\sqrt{N}\tilde{\rho}_h \xrightarrow{D} \mathcal{N}(0, \|C_p\|_2^2), \tag{3.4}$$

where $\|\cdot\|_2$ is the Hilbert-Schmidt norm of a kernel. Moreover, if for a positive integer H $\tilde{R}_H = (\tilde{\rho}_1, \ldots, \tilde{\rho}_H)^\top$, then

$$\sqrt{N}\tilde{R}_H \xrightarrow{D} \mathcal{N}_H(0, \|C_p\|_2^2 \mathbb{I}_H), \tag{3.5}$$

where $\mathcal{N}_H(m, \Sigma)$ denotes an H-dimensional normal random vector with mean vector m and covariance Σ, and $\mathbb{I}_H$ is the H-dimensional identity matrix.

The proof of Theorem 3.2.1 is given in the Appendix B.1.1. This result may be used to justify the asymptotic consistency of tests of the hypotheses:

$\mathcal{H}_{0,h}$: For a fixed positive integer h, $\rho_h = 0$, and

$\mathcal{H}'_{0,H}$: For a fixed positive integer H, $\rho_h = 0$ for all $h \in \{1, \ldots, H\}$.

In particular, an asymptotically sized α test of $\mathcal{H}_{0,h}$ is to reject when $|\sqrt{N}\tilde{\rho}_h| > \|C_P\|_2 z_{\alpha/2}$, where z_q is the q critical value of the standard normal distribution. Similarly, a $1-\alpha$ confidence set for the fSACF assuming that the series is a strong white noise is

$$-\frac{z_{1-\alpha/2}}{\sqrt{N}\|C_P\|_2} \leq \tilde{\rho}_h \leq \frac{z_{1-\alpha/2}}{\sqrt{N}\|C_P\|_2}. \tag{3.6}$$

Comparing the estimated fSACF $\tilde{\rho}_h$ to its 95% confidence bounds assuming that the underlying series follows a strong white noise leads to a simple visual summary that may be used to measure at a glance the serial dependence structure of a FTS; see e.g. Figure 3.1 below.

Testing $\mathcal{H}'_{0,H}$ is often referred to as "portmanteau testing". As a result of Theorem 3.2.1, the test statistic is

$$\tilde{Q}_{N,H} = \|\sqrt{N}\tilde{R}_H\|_E^2 = N\sum_{h=1}^{H}\tilde{\rho}_h^2 \xrightarrow{D} \|C_P\|_2^2\chi^2(H),$$

where $\|\cdot\|_E$ is a standard Euclidean norm of a vector, and $\chi^2(H)$ denotes a χ^2 random variable with H degrees of freedom. An approximate p-value of a test of $\mathcal{H}'_{0,H}$ may be computed as $p = \mathbb{P}(\|C_P\|_2^2\chi^2(H) > \tilde{Q}_{N,H})$.

3.2.3 Inference for ρ_h assuming μ is unknown

In practice of course the spatial median μ is typically not known, and must instead be estimated from the sample. Given its definition, it is natural to estimate μ with

$$\hat{\mu}_{ideal} = \arg\min_{\mu \in \mathcal{H}} \sum_{i=1}^{N} \|X_i - \mu\| \,. \tag{3.7}$$

Although there is no closed-form solution for the estimator satisfying (3.7), several authors have proposed iterative procedures to approximate $\hat{\mu}_{ideal}$, and have shown that these lead to consistent estimators with quantifiable convergence rates. For example, Gervini (2008) proposes an approach to approximately solve (3.7) using Gateaux differentials, and Cardot et al. (2013) propose an approach based on stochastic gradient decent. Both of these approaches result in an estimator $\hat{\mu}$ approximately satisfying (3.7). Using $\hat{\mu}$ in place of μ in (3.3) leads to the feasible estimator

$$\hat{\rho}_h = \frac{1}{N} \sum_{i=1}^{N-h} \langle S(X_i - \hat{\mu}), S(X_{i+h} - \hat{\mu}) \rangle \,. \tag{3.8}$$

We now aim to establish conditions under which $\hat{\rho}_h$ also satisfies Theorem 3.2.1 and is asymptotically consistent to ρ_h. Or, in other words, conditions under which the effect of estimating the spatial median μ is asymptotically negligible. We begin by establishing an analogue of Theorem 3.2.1 under the following additional assumptions.

Assumption 3.2.1 *The sequence $\sqrt{N}(\hat{\mu} - \mu)$ is uniformly tight in $\mathcal{H}$.*

Assumption 3.2.2 *There exists constants $C_1, C_2 > 0$ so that for all $u \in \mathcal{H}$ with $\|u\| \leq C_1$, $\mathbb{E}\|X_i - (\mu + u)\|^{-2} \leq C_2$.*

For a definition of uniform tightness in a metric space see Billingsley (1968). Assumption 3.2.2 coincides with Assumption A3 in Cardot et al. (2013), where it is discussed extensively. It holds, for example, for Gaussian processes under mild conditions, as well as for many other processes possessing exponentially decaying small-ball probabilities; see Li and Shao

(2001) and Nazarov (2009). The stochastic gradient decent-based estimator of Cardot et al. (2013) also satisfies Assumption 3.2.1 under Assumption 3.2.2 and an additional assumption entailing that the spatial median is unique. The estimator of the spatial median proposed in Gervini (2008) also satisfies Assumption 3.2.1 under similar conditions.

Theorem 3.2.2 *Suppose that $\{X_i\}_{i\in\mathbb{Z}}$ is a strong white noise in $\mathcal{H}$, and that Assumptions 3.2.1 and 3.2.2 hold. Then $\hat{\rho}_h$ satisfies Theorem 3.2.1 in place of $\tilde{\rho}_h$.*

Theorem 3.2.2 shows that under Assumptions 3.2.1 and 3.2.2, the same confidence sets and portmanteau tests based on $\tilde{\rho}_h$ remain asymptotically valid when based on $\hat{\rho}_h$ instead.

We now turn to the asymptotic consistency of $\hat{\rho}_h$, which we establish under the following assumptions.

Assumption 3.2.3 $\hat{\mu} \overset{a.s.}{\to} \mu$ *as* $N \to \infty$.

Assumption 3.2.4 *For any* $h \geq 1$, $\mathbb{E}\left[\|X_0\|^{-1}\|X_h\|^{-1}\right] < \infty$.

Assumption 3.2.5 $\mathbb{E}\|X_0\|^{-1} < \infty$.

Theorem 3.2.3 *Suppose $\{X_i\}_{i\in\mathbb{Z}}$ is strictly stationary and ergodic, and that Assumptions 3.2.3-3.2.5 hold. Then $\hat{\rho}_h \overset{a.s.}{\to} \rho_h$ as $n \to \infty$.*

Assumption 3.2.3-3.2.5 are analogous in a time series setting of the assumptions in Theorem 1 of Boente et al. (2019), and the proof makes use of some of the same ideas. Assumption 3.2.5 essentially ensures that the terms in the sum defining $\hat{\rho}_h$ have bounded expectation, and from a technical perspective appears in bounding terms in the sum where X_i or X_{i+h} have large deviations from $\hat{\mu}$. When the data are a strong white noise Assumption 3.2.4 is implied by Assumption 3.2.5. Theorem 3.2.3 implies that when $\rho_h \neq 0$ for some $h \in \{1, \ldots, H\}$, both tests of $\mathcal{H}_{0,h}$ and $\mathcal{H}'_{0,H}$ described at the end of Section 3.2.2 are consistent when based on $\hat{\rho}_h$.

3.3 Implementation and Simulation Studies

In this section we provide some details on implementation, and then present the results of several simulation studies that aimed to evaluate the proposed inferential procedures for the fSACF.

In the simulations and applications below, we assume that all FDs are observed on a common grid $t_j \in [0,1]$, $j = 1, \ldots, M$. In our applications to electricity price data and intraday asset price data below, each FD object is observed at $M = 24$ and $M = 78$ points, respectively, and for all simulated examples we generated the data on $M = 101$ equally spaced points. In these cases all inner products and norms involved in producing $\hat{\rho}_h$ can be estimated by simple Riemann integration. For instance, with $t_0 = 0$

$$\|X_i - \hat{\mu}\|^2 \approx \sum_{j=1}^{M} [X_i(t_j) - \hat{\mu}(t_j)]^2 [t_j - t_{j-1}]$$

and

$$\begin{aligned} &\langle S(X_i - \hat{\mu}), S(X_{i+h} - \hat{\mu}) \rangle \\ &\approx \frac{1}{\|X_i - \hat{\mu}\| \|X_{i+h} - \hat{\mu}\|} \sum_{j=1}^{M} [X_i(t_j) - \hat{\mu}(t_j)][X_{i+h}(t_j) - \hat{\mu}(t_j)](t_j - t_{j-1}). \end{aligned}$$

If the FDs are not observed on a common grid, then one may use interpolation techniques to estimate the integrals required to compute $\hat{\rho}_h$; see e.g. Chapter 2 of Ramsay and Silverman (2005).

In order to estimate $\|C_P\|_2$, we use the estimator

$$\hat{C}_P(t,s) = \frac{1}{N} \sum_{j=1}^{N} S(X_j(t) - \hat{\mu}(t)) S(X_j(s) - \hat{\mu}(s)),$$

and estimate its norm based on the discrete sample with

$$\|\hat{C}_P^2\|_2^2 \approx \sum_{j,k=1}^{M} \hat{C}_P^2(t_j, t_k)[t_j - t_{j-1}][t_k - t_{k-1}].$$

Additional simulations evaluating the accuracy of these estimators are provided in Appendix B.2, which suggested that they work well so long as the resolution of the data is not too low. With these estimators, the confidence sets for $\hat{\rho}_h$ assuming the series is a strong white noise, and p-values for the portmanteau test, may be calculated respectively as

$$-\frac{z_{1-\alpha/2}}{\sqrt{N}\|\hat{C}_P\|_2} \leq \hat{\rho}_h \leq \frac{z_{1-\alpha/2}}{\sqrt{N}\|\hat{C}_P\|_2}, \tag{3.9}$$

and

$$\hat{Q}_{N,H} = N\sum_{h=1}^{H}\hat{\rho}_h^2, \quad p = \mathbb{P}(\|\hat{C}_P\|^2\chi^2(H) > \hat{Q}_{n,H}). \tag{3.10}$$

In all examples below μ was estimated using the estimator described in Gervini (2008).

We now turn to presenting the results of several simulation experiments. Below we refer to the functional autocorrelation function of Mestre et al. (2021) as the *fACF*. All of the computations are performed using `R`, version 4.1.2 (Team, 2021).

3.3.1 White noise processes

We simulated several strong white noise (WN) processes, and examined their fACF and fSACF. The examples that we considered were

1. Standard Brownian motion (BM): $X_i(t) = W_i(t)$, where $\{W_i(t),\ t \in [0,1]\}_{i\in\mathbb{Z}}$ are i.i.d. standard Brownian motions.

2. Brownian bridge (BB): $X_i(t) = B_i(t)$, where $B_i(t) = W_i(t) - tW_i(1)$.

3. Fourier-Cauchy process (F): $X_i(t) = F_i(t)$, where $F_i(t) = Z_{1,i} + \sum_{k=1}^{3}[Z_{2k,i}\cos(2\pi kt) + Z_{2k+1,i}\sin(2\pi kt)]$ and $Z_{j,i}$ are a family of independent and identically distributed standard Cauchy random variables.

4. B-spline Exponential process (BS): $X_i(t) = Bs_i(t)$ where $Bs_i(t) = \sum_{k=1}\varepsilon_{i,k}B_k(t)$ in which $\varepsilon_{i,k}$ is a family of independent and identically distributed exponential random variables with mean 1, and $B_k(t)$ is the k-th orthogonal cubic B-spline basis function.

The top panel of Figure 3.1 displays 50 i.i.d. realizations of each WN process along with the estimated spatial median. The estimated fSACF for each process is displayed in the bottom row of Figure 3.1, along with blue lines indicating the confidence intervals in (3.9).

In order to evaluate the coverage properties of the proposed confidence intervals for the fSACF, each WN process was independently generated 1000 times for sample sizes $N \in \{100, 250, 500, 1000, 2000\}$. Table 3.1 shows the proportion of estimators $\hat{\rho}_h$ and $\tilde{\rho}_h$ that were not contained in the $1-\alpha$ confidence intervals (3.9) and (3.6), respectively, for $\alpha \in \{10\%, 5\%, 1\%\}$, and for several values of h. These results suggest that, for the sample sizes and processes considered, the confidence intervals derived from Theorems 3.2.1 and 3.2.2 have approximately correct coverage. We notably also observed this for the Fourier-Cauchy process, as well as the exponential process, which lends some evidence to the statement that the fSACF is robust against low-order moments/outliers and skewness in the data in practice. For smaller n and larger values of h, we noticed that the intervals tended to have somewhat larger than nominal coverage, which we think may be attributed to the fact that the estimators $\hat{\rho}_h$ and $\tilde{\rho}_h$ are normalized by $1/N$ rather than $1/(N-h)$. In unreported simulations when we normalized by $1/(N-h)$ instead we noticed the opposite effect. Additionally, we observed that the effect of estimating the spatial median appeared to be negligible in all cases.

3.3.2 Functional time series models

We now turn our attention to some examples of serially dependent FTS. We consider several models here, the first two in this section for the purpose of performing simulation experiments, and the second two presented in Section 3.4 are used for the purpose of time series modelling and residual analyses in two real data examples.

The time series models that we considered in our simulation experiments are from the general class of functional autoregressive models of order p (FAR(p)), which take the form

$$X_i(t) = \sum_{j=1}^{p} \Phi_j(X_{i-p})(t) + \epsilon_i(t) = \sum_{j=1}^{p} \int \phi_j(t,s) X_{i-j}(s) ds + \epsilon_i(t),$$

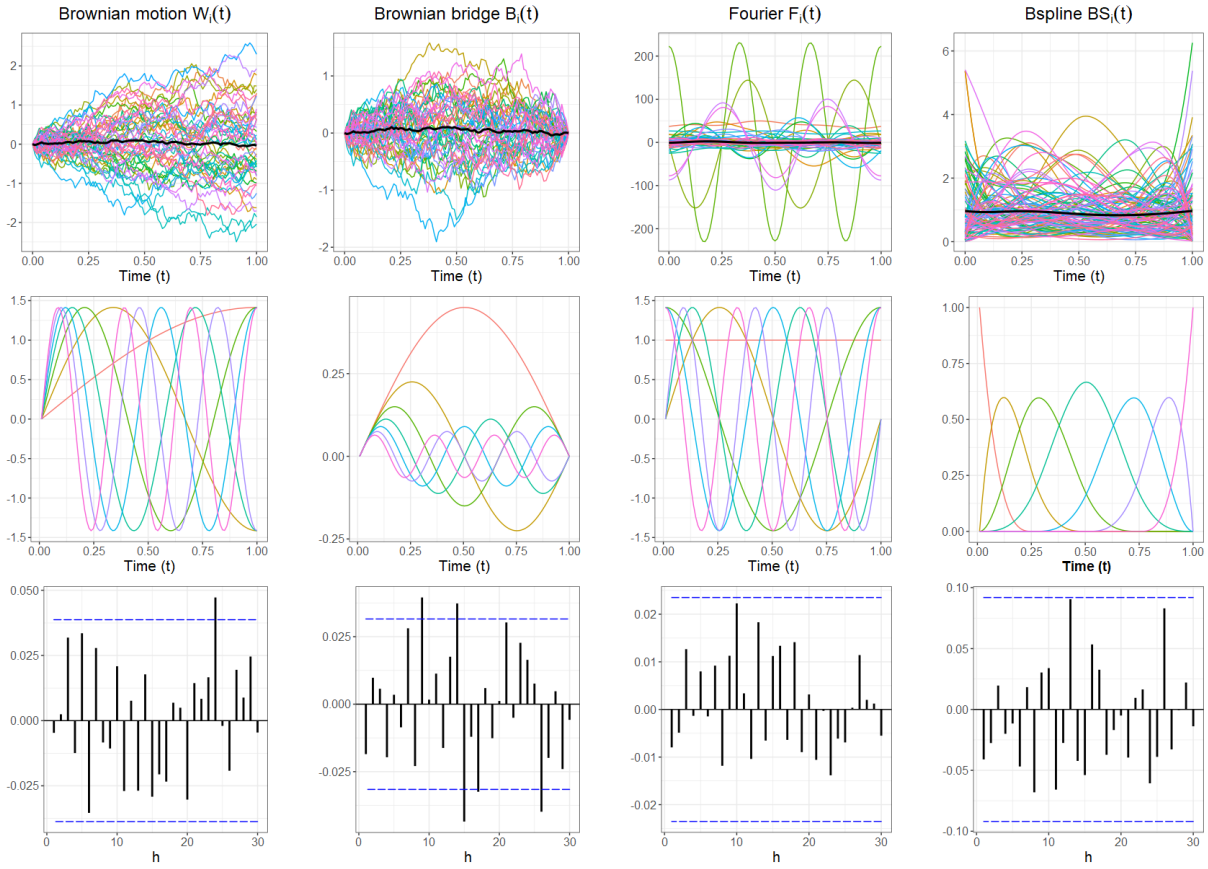


Figure 3.1: **Top**: 50 i.i.d. curves of each strong noise process with the spatial median estimate from Gervini (2008) (in black); **Middle**: The leading 7 basis functions in the Karhunen-Loéve expansion of each process; **Bottom**: The estimated fSACF of each series for lags $h \in \{1, \ldots, 30\}$ with 95% confidence bounds as in (3.9) (in blue).

where $\phi_j(t,s)$ are the kernels of the integral operators Φ_j, and ϵ_i is a sequence of i.i.d. mean zero functional processes. In our examples, we use the Gaussian kernel

$$\phi(t,s) = c \exp\left(-\frac{t^2+s^2}{2}\right),$$

where the constant c is chosen in order that $\|\phi\| = |S| \in [0,1]$, where $S = \text{sign}(c)\,\|\phi\|$. As such the parameter S describes both the magnitude and "sign" of the kernel ϕ. We considered the processes

Table 3.1: The empirical rate among 1000 independent simulations of four WN processes that the $1-\alpha$ confidence intervals (3.6) and (3.9) do not contain the estimators $\tilde{\rho}_h$ and $\hat{\rho}_h$ for $\alpha \in \{10\%, 5\%, 1\%\}$, $N \in \{100, 250, 500, 1000, 2000\}$ and several values of h.

	N	h	Significance level α							N	h	Significance level α					
			10%		5%		1%					10%		5%		1%	
Statistic			$\tilde{\rho}_h$	$\hat{\rho}_h$	$\tilde{\rho}_h$	$\hat{\rho}_h$	$\tilde{\rho}_h$	$\hat{\rho}_h$				$\tilde{\rho}_h$	$\hat{\rho}_h$	$\tilde{\rho}_h$	$\hat{\rho}_h$	$\tilde{\rho}_h$	$\hat{\rho}_h$
BM	100	1	0.104	0.108	0.052	0.055	0.011	0.013	BB	100	1	0.086	0.093	0.048	0.05	0.008	0.013
		5	0.116	0.112	0.053	0.062	0.011	0.010			5	0.087	0.080	0.040	0.043	0.007	0.005
		10	0.081	0.082	0.033	0.030	0.006	0.004			10	0.070	0.069	0.038	0.038	0.007	0.006
		15	0.074	0.079	0.034	0.039	0.003	0.005			15	0.070	0.078	0.022	0.027	0.003	0.002
	250	1	0.118	0.118	0.053	0.056	0.008	0.010		250	1	0.097	0.096	0.046	0.049	0.010	0.011
		5	0.092	0.096	0.049	0.049	0.010	0.013			5	0.110	0.104	0.057	0.059	0.011	0.015
		10	0.078	0.082	0.038	0.039	0.007	0.007			10	0.066	0.076	0.030	0.029	0.004	0.004
		15	0.088	0.089	0.034	0.035	0.005	0.006			15	0.094	0.097	0.050	0.050	0.002	0.002
	500	1	0.114	0.123	0.059	0.055	0.005	0.008		500	1	0.091	0.094	0.048	0.053	0.012	0.011
		5	0.087	0.094	0.04	0.038	0.004	0.004			5	0.097	0.101	0.051	0.052	0.01	0.013
		10	0.106	0.106	0.053	0.054	0.013	0.01			10	0.086	0.088	0.036	0.037	0.005	0.003
		15	0.100	0.099	0.058	0.054	0.011	0.012			15	0.101	0.105	0.048	0.044	0.011	0.007
	1000	1	0.104	0.108	0.043	0.045	0.010	0.011		1000	1	0.115	0.114	0.056	0.054	0.014	0.016
		10	0.105	0.107	0.059	0.060	0.008	0.008			10	0.094	0.094	0.044	0.041	0.010	0.009
		20	0.104	0.107	0.055	0.057	0.009	0.010			20	0.098	0.099	0.046	0.045	0.009	0.008
		30	0.112	0.104	0.056	0.058	0.015	0.014			30	0.091	0.086	0.042	0.041	0.008	0.01
	2000	1	0.098	0.096	0.043	0.045	0.004	0.005		2000	1	0.107	0.113	0.066	0.063	0.016	0.016
		10	0.096	0.096	0.058	0.058	0.009	0.009			10	0.084	0.084	0.045	0.043	0.007	0.007
		20	0.098	0.098	0.056	0.057	0.013	0.012			20	0.093	0.093	0.045	0.046	0.006	0.005
		30	0.107	0.107	0.053	0.055	0.009	0.008			30	0.096	0.098	0.041	0.044	0.006	0.008
F	100	1	0.86	0.108	0.051	0.055	0.012	0.016	BS	100	1	0.095	0.102	0.05	0.062	0.01	0.013
		5	0.089	0.102	0.052	0.053	0.007	0.009			5	0.08	0.081	0.036	0.029	0.005	0.005
		10	0.07	0.069	0.03	0.032	0.007	0.010			10	0.075	0.075	0.029	0.036	0.003	0.004
		15	0.062	0.074	0.028	0.026	0.003	0.003			15	0.07	0.077	0.026	0.033	0.006	0.009
	250	1	0.09	0.101	0.052	0.052	0.005	0.006		250	1	0.099	0.095	0.044	0.05	0.005	0.008
		5	0.074	0.08	0.031	0.032	0.007	0.008			5	0.084	0.086	0.047	0.046	0.012	0.012
		10	0.088	0.083	0.038	0.04	0.003	0.003			10	0.103	0.107	0.05	0.056	0.012	0.01
		15	0.078	0.087	0.039	0.038	0.006	0.007			15	0.08	0.083	0.044	0.042	0.006	0.005
	500	1	0.099	0.096	0.048	0.046	0.012	0.01		500	1	0.12	0.117	0.057	0.055	0.008	0.01
		5	0.102	0.097	0.051	0.048	0.01	0.012			5	0.105	0.113	0.051	0.052	0.012	0.012
		10	0.088	0.093	0.045	0.04	0.008	0.01			10	0.101	0.116	0.043	0.043	0.007	0.008
		15	0.093	0.092	0.041	0.043	0.009	0.005			15	0.101	0.101	0.036	0.042	0.003	0.005
	1000	1	0.117	0.113	0.062	0.062	0.019	0.019		1000	1	0.1	0.098	0.047	0.047	0.011	0.011
		10	0.097	0.09	0.045	0.046	0.011	0.01			10	0.09	0.088	0.035	0.038	0.005	0.008
		20	0.110	0.109	0.054	0.051	0.01	0.012			20	0.097	0.101	0.052	0.05	0.013	0.013
		30	0.085	0.081	0.047	0.045	0.007	0.006			30	0.096	0.096	0.039	0.039	0.006	0.007
	2000	1	0.115	0.11	0.071	0.07	0.016	0.016		2000	1	0.098	0.103	0.044	0.049	0.013	0.011
		10	0.104	0.108	0.063	0.06	0.011	0.014			10	0.087	0.087	0.043	0.046	0.011	0.01
		20	0.1	0.095	0.047	0.046	0.004	0.005			20	0.08	0.075	0.035	0.037	0.01	0.009
		30	0.1	0.099	0.047	0.046	0.006	0.004			30	0.101	0.103	0.048	0.046	0.01	0.011

- FAR(1, S): $X_i(t) = \int \phi_1(t,s) X_{i-1}(s) ds + \varepsilon_i(t)$, where $\varepsilon_i(t)$ follows the Brownian bridge WN process (BB).

- FAR(2, S_1, S_2): $X_i(t) = \int \phi_1(t,s)X_{i-1}(s)ds + \int \phi_2(t,s)X_{i-2}(s)ds + \varepsilon_i(t)$, where again $\varepsilon_i(t)$ follows the Brownian bridge WN process (BB).

Figure 3.2 illustrates simulated series of length 1000 from the FAR(1,S) process for various values of the parameter S, along with corresponding plots of the fACF of Mestre et al. (2021) and the fSACF. This figure highlights one of the advantages of the fSACF compared to the fACF, in that when the parameter S is negative, the fACF displays a similar behaviour as if the parameter were positive, whereas the fSACF in this case shows an alternating pattern.

To study the empirical power of the proposed confidence interval and portmanteau test in (3.9) and (3.10), we applied our tests to 1000 independently generated FAR(1, S) processes of various lengths $N \in \{100, 250, 500, 1000\}$ and for increasing values of S. Figure 3.3 shows the empirical rejection rate of the hypotheses $\mathcal{H}'_{0,H}$ for $H = 1, 10$ that we observed among the 1000 independent simulations with significance level set at $\alpha = 0.05$ in terms of power curves that are a function of S. We note that the rejection rate of $\mathcal{H}'_{0,1}$ at level 0.05 is equivalent to the rate at which the confidence interval (3.9) does not contain $\hat{\rho}_1$. We observed that the tests appeared to have empirical size close to the nominal size, and that expectedly as S increased to 1, the power of each test also increased to 1. The power was observed to increase more slowly for larger H, which we attribute to the fact that the FAR(1,S) process has decreasing correlation as a function of the lag. It was also clear that the empirical power of the test increased as the sample size n increased even when S was small, and the empirical power reached one when the dependence is around 0.45 regardless of the sample size.

We also performed some simulations to highlight the use of the fSACF in evaluating the goodness-of-fit of FTS models through residual analysis. Given a model capable of producing fitted/forecasted values of the series $\hat{X}_i(t)$, the residuals are calculated simply as

$$r_i(t) = X_i(t) - \hat{X}_i(t). \tag{3.11}$$

We considered performing residual analyses of FAR models using an estimator derived from FPCA. For the FAR(1) model, the kernel ϕ may be estimated using a least squares

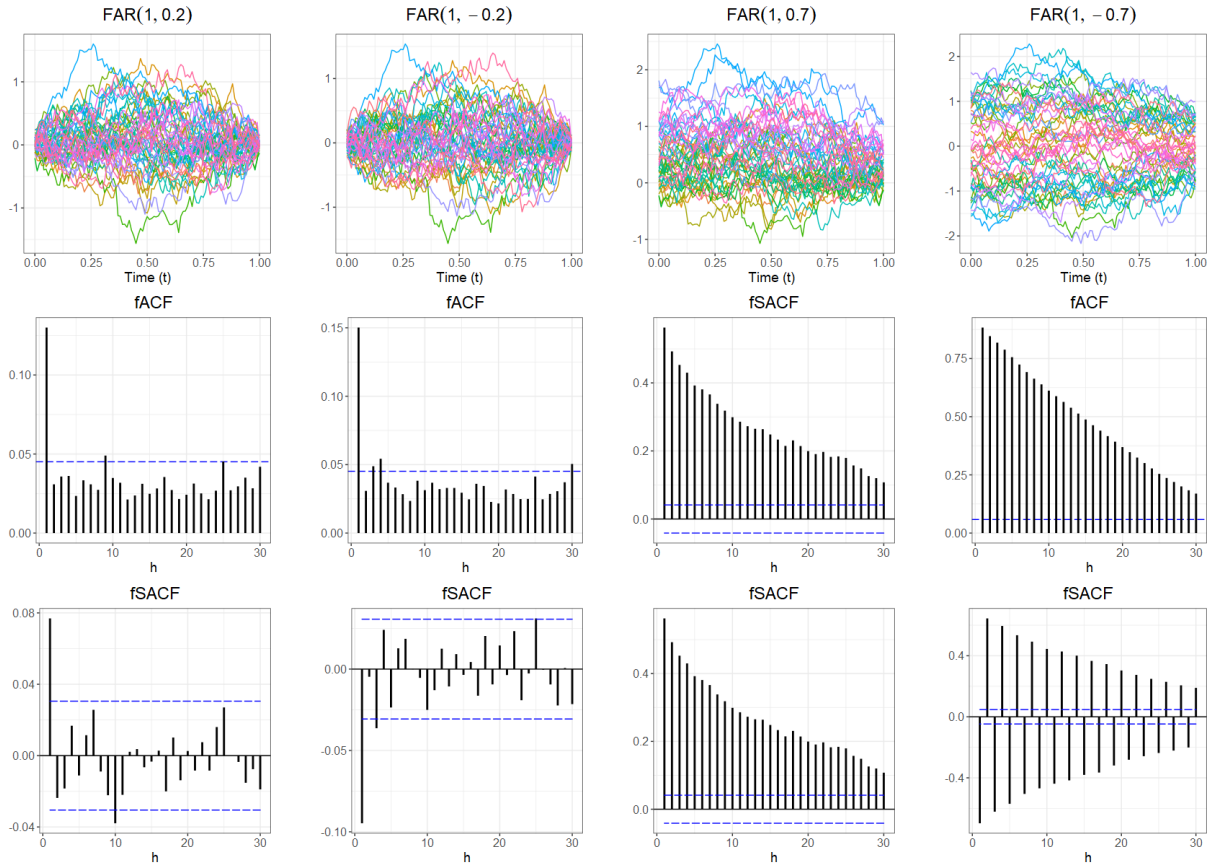


Figure 3.2: 1000 curves simulated from FAR$(1, S)$ models with different scaling parameter S. **Top**: 50 i.i.d. curves simulated based on each model; **Middle**: The fACF of each of the processes with 95% confidence bound (in blue) **Bottom**: The fSACF of each of the processes with 95% confidence bounds (3.9) (in blue).

principle by the first J FPCs:

$$\hat{\phi}_1(t,s) = \frac{1}{N-1}\sum_{k=1}^{N-1}\sum_{j=1}^{J}\sum_{i=1}^{J}\hat{\lambda}_j^{-1}\left\langle X_k, \hat{v}_j\right\rangle\left\langle X_{k+1}, \hat{v}_i\right\rangle \hat{v}_j(s)\hat{v}_i(t), \tag{3.12}$$

where the $\hat{\lambda}_j$ and $\hat{v}_j(\cdot)$ are the estimated eigenvalues and eigenfunctions of the covariance operator of the observations; see e.g. Shang (2014). Such an estimator may also be constructed for the kernels in general FAR(p) models, and we provide the details in Appendix B.3. The fitted values are then calculated in this case as

$$\hat{X}_i(t) = \int \hat{\phi}_1(t,s)X_{i-1}(s)ds.$$

There are a number of ways to determine the number of FPCs J to use, including information criteria (Shibata, 1981; Yao, 2007), resampling methods (Bathia et al., 2010; Hall and Vial, 2006) and cross-validation (Leng and Müller, 2006). Here, we use the CPV

$$\text{CPV}(J) = \frac{\sum_{k=1}^{J}\hat{\lambda}_k}{\sum_{k=1}^{N}\hat{\lambda}_k},$$

and choose the smallest J for which $\text{CPV}(J) > 90\%$. We used the **fda** package to carry out FPCA; see Ramsay et al. (2009).

Figure 3.4 shows three fSACF's: the left most is computed from simulated FAR(2,S_1,S_2) processes, whereas the middle and right figures are computed from the residuals of an FAR(2) and FAR(1) model fit, respectively. It is clear based on these plots that in each case there remains significant autocorrelation in the residuals based on the FAR(1) model, as expected, while the residuals of the FAR(2) model appear to be reasonably white.

3.4 Applications

In this section, we demonstrate the proposed methodology in two data applications. The first is concerned with fitting forecasting models and evaluating their goodness-of-fit with Spanish electricity price curves. We now introduce the two forecasting models that we consider.

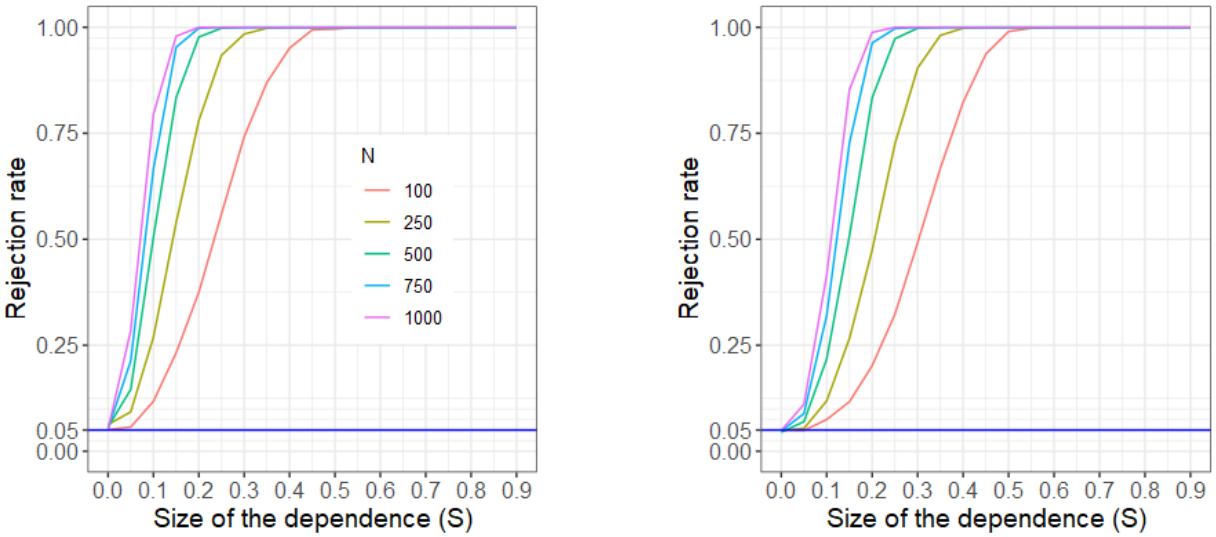


Figure 3.3: Empirical rejection rate of the portmanteau test under $\mathcal{H}'_{0,H}$ of data generated according to the FAR(1,S) model based on 1000 independent simulations. **Left**: H = 1; **Right**: H = 10. The nominal level α is set to be 0.05.

1) Functional seasonal autoregressive model (FSAR) (Horváth et al., 2020): The model is given as

$$X_i(t) = \int \phi_{\ell_1}(t,s)X_{i-\ell_1}(s)ds + \cdots + \int \phi_{\ell_p}(t,s)X_{i-\ell_p}(s)ds + \varepsilon_i(t),$$

where $\ell_1, \ldots, \ell_p$ are user selected lags. It is clear that the FSAR models are a special cases of general FAR models. We show in Appendix B.3 how estimators for the kernels in this model may be constructed based on FPCA and a least-squares principle akin to (3.12).

2) The forecasting model that we consider is the Hyndman-Ullah method (Hyndman and Ullah, 2007), which we abbreviate as the "HU" model: The HU(J) is a model that is potentially nonlinear and non-stationary that makes use of a truncated Karhunen–Loève expansion. It supposes that the underlying processes may be well-approximated by its projection onto the first J eigenfunctions of the covariance operator

$$X_i^{(J)}(t) = \hat{m}(t) + \sum_{j=1}^{J} \xi_{i,j}\hat{v}_j(t),$$

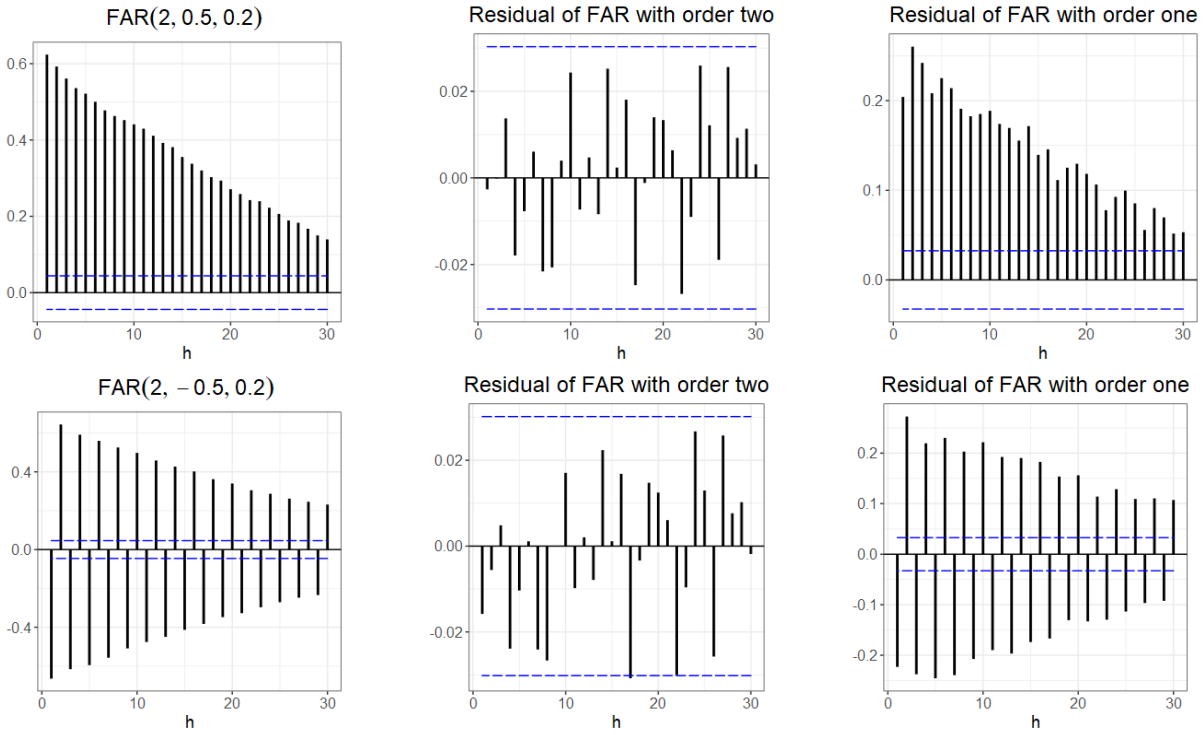


Figure 3.4: 1000 simulated curves from FAR$(2, S_1, S_2)$ models with different scaling parameters S_1 and S_2 with the error process being a Brownian bridge. **Left**: The fSACF of observations simulated from the FAR$(2, S_1, S_2)$ models with different coefficients of dependence S_1 and S_2. **Middle**: The fSACF of the residuals r_i from an FAR(2) model. **Right**: The fSACF of the residuals r_i from an FAR(1) model, which in this case is evidently under specified.

where the $\hat{m}$ is the standard sample mean function, $\xi_{i,j}$ is referred to as the PC score, or simply the 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 $X_i(t) \approx \hat{m}(t) + \sum_{j=1}^{J} \xi_{i,j}\hat{v}_j(t)$. J may be selected according to a number of criteria, e.g. general cross validation, information criterion, CPV, or goodness-of-fit testing.

2. To each of the FPC series, $\xi_{i,j}, i = 1, \ldots, N$, $j = 1, \ldots, J$, we fit a, potentially non-

stationary, scalar time series model. Notably the models can be distinct for different score series. Here, we fit SARIMA models, see e.g. Shumway and Stoffer (2017). These models may be used to compute forecasts and fitted values of the score series $\hat{\xi}_{ij}$.

3. Forecasts or fitted values are constructed as $\hat{X}_i(t) = \hat{m}(t) + \sum_{j=1}^{J} \hat{\xi}_{i,j}\hat{v}_j(t)$, and residual curves may be computed as $r_i(t) = X_i(t) - \hat{X}_i(t) = X_i(t) - \sum_{i=1}^{J} \hat{\xi}_{ij}\hat{v}_j(t)$.

3.4.1 Spanish electricity data

The first application we consider is Spanish electricity price data observed hourly from January 1, 2014, to December 31, 2014. The data was originally provided by the Spanish electricity market operator, and are also available in the R **fdaACF** package (Mestre et al., 2020). González et al. (2017) studied fitting and forecasting these data using an autoregressive–moving-average model with exogenous inputs (ARMAX), and Mestre et al. (2021) uses the same dataset to illustrate the properties of the fACF. We use $\{Y_i(t),\ i = 1, 2, \ldots, 365,\ t \in [1, 24]\}$ to denote the linearly interpolated Spanish electricity price curves, and these are displayed in the top left panel of Figure 3.5. Although each curve generally evolves according to a standard daily pattern of low prices in the early morning hours, and higher prices during the times of peak demand, we also observe that there is a strong weekly and seasonal trend to the series over the course of the year. The corresponding fACF and fSACF presented in the middle and right of the top row of Figure 3.5 both indicate such a pattern. To remove the trend component, we consider the pointwise differenced series $D_i(t) = Y_i(t) - Y_{i-1}(t)$. The bottom row of Figure 3.5 displays the differenced series along with the corrsponding fACF and fSACF plots. The differenced series appears roughly mean stationary, and has prominent autocorrelation corresponding to the weekly lag $h = 7$. In particular, the fSACF indicates that the correlation at lags one and two are "negative", while the correlation at weekly lags is "positive", which is to be expected. In contrast, the fACF does not capture this pattern.

We then fit FSAR and HU(J) models to the differenced series D_i, and examined their goodness-of-fit using a residual analysis based on the fSACF. For the FSAR model, we

chose lags 1 and 7, as we see the data exhibited both lag 1 and weekly autocorrelation, and selected the model dimension using the CPV criterion explaining 90% of the total variation. We also considered HU(J) models for $J \in \{1, \ldots, 5\}$, and to each resulting scalar score series we fit a SARIMA model using the automated model selection method in Hyndman and Khandakar (2008). In particular, this lead to SARIMA models with orders $(1,0,2) \times (2,0,0)_7$, $(1,0,2) \times (2,0,0)_7$, $(2,0,2) \times (2,0,0)_7$, $(2,0,1) \times (2,0,1)_7$ and $(2,0,2) \times (2,0,0)_7$ for the first five univariate score series, respectively, where $(a,b,c) \times (A,B,C)_s$ describes a SARIMA model with seasonal lag s, standard and seasonal differencing b and B, standard and seasonal autoregressive orders a and A, and standard and seasonal moving average orders of c and C. Figure 3.6 shows the fSACF of the residuals of these models. We observed that the FSAR model does not fit the data well, as strong autocorrelation is still present in the residuals at low lags as well as in the weekly lag. This did not improve by increasing the CPV value up to 99%. As for the HU(J) model with $J \in \{3,4,5\}$, most of the fSACF values fall within the 95% confidence bounds coloured in blue. We additionally computed the p-values of tests of $\mathcal{H}'_{0,H}$ as a function of H based on the statistic $\hat{Q}_{364,H}$ applied to the residuals of each model in Figure 3.6. These suggest again that the FSAR fit and HU(1) fit are poor. As for HU(3), these p-values for small lags were slightly above 0.05, while for HU(5) the p-values were for most lags above the 0.05 threshold. This suggests that, although there may still be some weak, lingering autocorrelation in the residuals of the HU(5) model at the weekly lag, this model generally appears to fit well.

3.4.2 Densely observed intraday asset price data

The second application that we considered was to the densely observed intraday asset price data used in Kokoszka et al. (2017). This dataset consists of financial asset prices, including commodity and currency exchange rates, obtained from the Chicago Mercer Exchange (*cmegroup.com*), S&P 500 index, and Apple stock price obtained from Nasdaq (*nasdaq.com*). Each series that we consider is tabulated in Table 3.2. There are 78 daily observations spanning 249 trading days within the time span from Jan 02, 2014, to Dec 31, 2014.

Let $P_i(t)$ denote the intraday asset price of an asset on the i-th day at time t, which

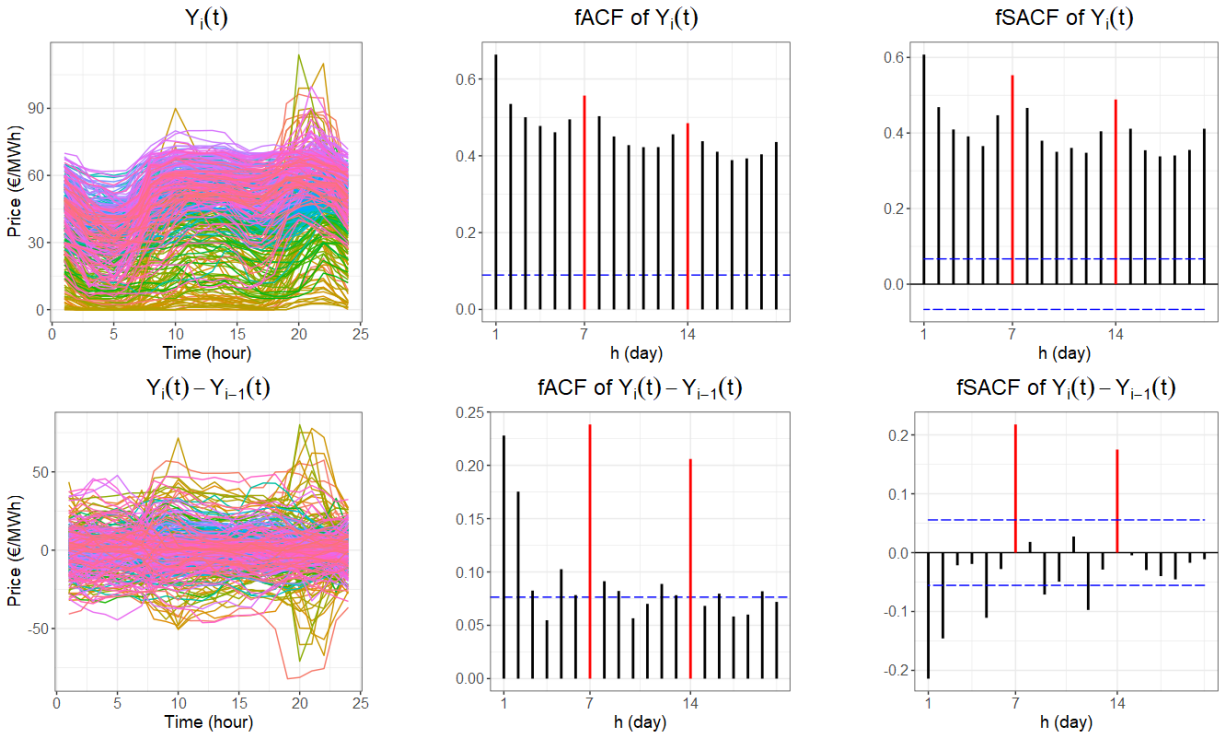


Figure 3.5: **Left**: Top is the hourly observed annual Spanish electricity data $\{Y_i(t), i = 1, \ldots, 365, t = 1, \ldots, 24\}$ from January 1st 2014 to December 31 sgemented by day. Bottom is the differenced curves $\{Y_i(t) - Y_{i-1}(t), t \in [0, 24]\}$; **Middle**: The fACF of $Y_i(t)$ (top) and $Y_i(t) - Y_{i-1}(t)$ (bottom), respectively; **Right**: The fSACF of $Y_i(t)$ (top) and $Y_i(t) - Y_{i-1}(t)$ (bottom), respectively. The blue lines in the middle and right panels are the 95% confidence interval under the null hypothesis $\mathcal{H}_{0,h}$, for $h = 1, \ldots, 20$.

we obtained from the raw data using linear interpolation. On each trading day, the first return is recorded at 9:35, and the last one is at 16:00. We consider four FTS derived from these data: 1) five minute log return sequence

$$R_i(t) = \ln P_i(t) - \ln P_i(t-5),$$

2) the cumulative intraday returns (CIDRs)

$$C_i(t) = \ln P_i(t) - \ln P_i(0),$$

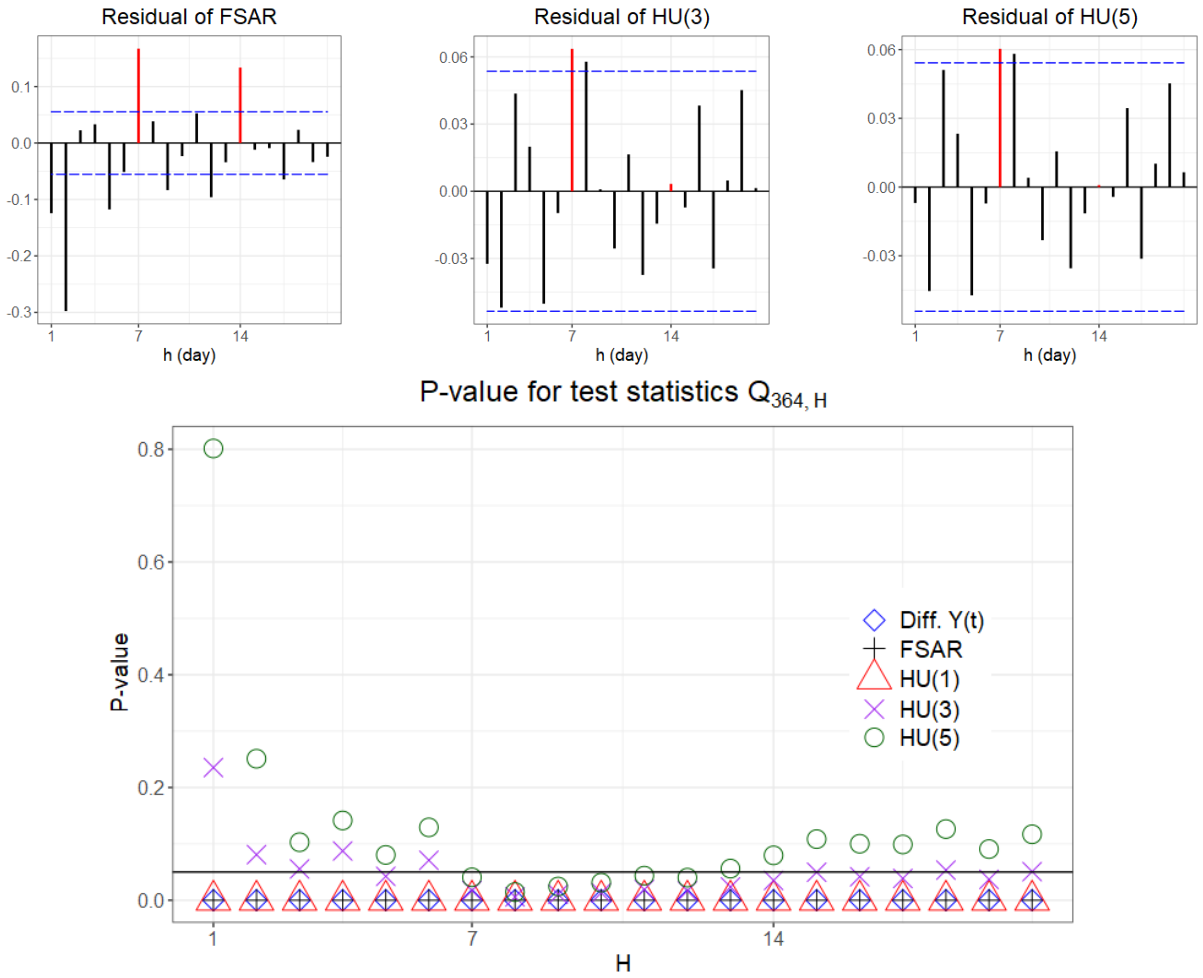


Figure 3.6: **Left**: The fSACF of the residual r_i after fitting by the FSAR model; using the ; **Middle**: The fSACF of the residual r_i after fitting the data using the HU model with three FPCs; **Right**: The fSACF of the residual r_i after fitting the data using the HU model with three FPCs; **Bottom** The p-values of test statistics $\hat{Q}_{364,H}$ under $\mathcal{H}'_{0,H}$ of residual using FSAR model, residual using HU model with three and five FPCs, and the differenced series $Y_i(t) - Y_{i-1}(t)$, respectively.

3) the squared sequence of five minute log return curve $R_i^2(t)$, and 4) squared CIDR sequence $C_i^2(t)$. The CIDR may be thought of as a normalized, smoothed curve that

Table 3.2: Assets used in the study spanned from Jan 02, 2014, to Dec 31, 2014.

Class	Notation	Description
Index	S&P 500	Standard and Poor's 500
Currency exchange	EC	Euro to US Dollar
Commodity futures	CL	Light sweet crude oil futures
Technology	AAPL	Apple Inc.

interpolates the standard log-returns with intraday data, which can potentially have high volatility; see, Ji et al. (2022), Kokoszka et al. (2019) and Rice et al. (2020).

Figure 3.7 displays the S&P 500 five minute log return $R_i(t)$, squared S&P 500 log return $R_i^2(t)$, the CIDR $C_i(t)$, and the squared CIDR $C_i^2(t)$ in the first four days, respectively. It appears that both the log return curves $R_i(u)$ and CIDRs $C_i(t)$ are roughly stationary. We then calculate the fSACF of each series, which is displayed in the case of light sweet crude oil futures (CL) in the the bottom row of Figure 3.7. We observed that both the five minute log-return series and CIDR series appear based on their fSACFs to be white noise processes. As for the squared curves, we observed that most series exhibited significant, although weak, spherical autocorrelation, especially in early lags. The cumulative significance of the observed spherical autocorrelations was assessed using the portmanteau test of $\mathcal{H}'_{0,10}$, the p-values of which are reported in Table 3.3. The appearance of white noise of the original series coexisting with correlation in the squared series is the hallmark of GARCH type behaviour or volatility. Confirming this observation with the fSACF, which is more robust to outliers and lower order moments in the data generating process, helps confirm that the patterns in the autocorrealtion observed previously in these series is not simply the consequence of outliers.

Table 3.3: Value of the test statistics $\hat{Q}_{249,10}$ and associated p-values computed from (3.10) for: 1) five minute log returns $R_i(t)$, 2) squared five minute log returns $R_i^2(t)$, 3) CIDR $C_i(t)$, and 4) squared CIDR $C_i^2(t)$, for each asset considered.

Type	$R_i(t)$		$R_i^2(t)$		$C_i(t)$		$C_i^2(t)$	
	$\hat{Q}_{249,10}$	p-value	$\hat{Q}_{249,10}$	p-value	$\hat{Q}_{249,10}$	p-value	$\hat{Q}_{249,10}$	p-value
S&P 500	3.0144	0.9811	31.0026	0.0006	6.2736	0.7918	12.9304	0.2276
EC	7.2702	0.6997	227.4773	0.0000	4.4191	0.9265	166.4873	0.0000
CL	6.6572	0.7574	390.0302	0.0000	4.4574	0.9244	210.533	0.0000
AAPL	5.8559	0.8272	24.0211	0.0080	6.8985	0.7350	7.7158	0.6566

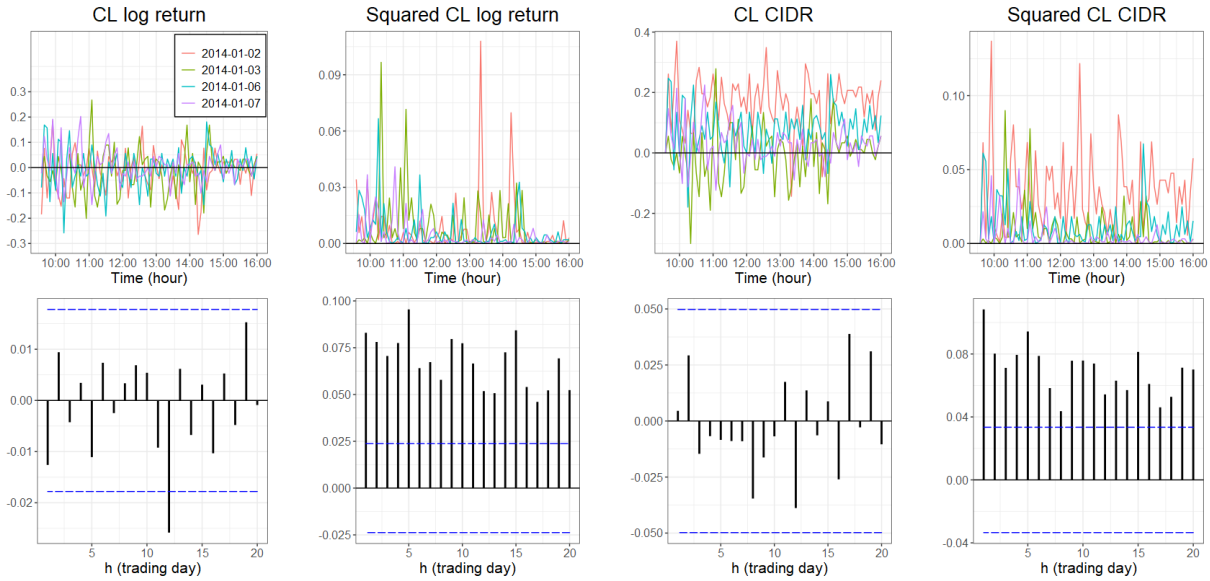


Figure 3.7: **Top**: samples of four curves from each of the four FTS considered, constructed from the intraday light sweet crude oil futures prices (CL). **Bottom**: the empirical fSACF of each to these series along with the 95% confidence interval 3.9.

3.5 Conclusion

We have put forward a new autocorrelation measure for FTS: the functional spherical autocorrelation function (fSACF). Unlike existing methods in the literature that rely on measuring the size of the empirical autocovariance operators, the approach we propose is based on measuring the inner product between projections of lagged pairs of the series onto the unit sphere. Some advantages conferred by this approach are its ability to capture the direction or "sign" of the serial dependence in the series at a given lag, and it is more robust to outliers/lower order moments of the data generating process.

A central limit theorem for the fSACF was established under the assumption that the underlying process is a strong white noise, and mild additional conditions were also put forward under which this result remains true when the spatial median defining the fSACF is replaced with an estimator. These large sample results were used to derive confidence intervals and portmanteau tests for the fSACF. Through Monte Carlo simulations, we showed that these confidence intervals and tests work well in finite samples, highlighting that they also appear to be robust against outliers/low moments in the data generating mechanism, and can be useful in performing model selection with FTS models. Two real data applications illustrated the use of the fSACF for model selection, and also helped confirm with a more robust approach the finding that intraday return curves derived from dense asset prices series appear to evolve as a weak white noise exhibiting conditional heteroscedasticity. Note that these densely observed intraday price data may not align with the assumption of strong white noise; instead, they may be more consistent with the assumption of weak white noise. Developing a version of the functional spherical ACF for this case is a potential avenue for future work.

Chapter 4

Factor Copula Models for Spatially Correlated Functional Data

4.1 Introduction

In functional data analysis (FDA), it is commonly assumed that the observed data are corrupted by noise and represent realizations of a smooth random function $X(t)$, where $t \in \mathcal{T}$ and $\mathcal{T}$ is a compact interval of $\mathbb{R}$. The mean function and covariance function of X are denoted by $\mu(t) = \mathbb{E}X(t)$ and $C(s,t) = \mathbb{C}\text{ov}(X(s), X(t))$, respectively. Over the past two decades, considerable attention has been given to the estimation of mean and covariance functions in functional data (FD). Relevant studies can be found in Li and Hsing (2010) and Zhang and Wang (2016). Among these studies, functional principal component analysis (FPCA) has gained significant popularity as a means of exploring the variational modes of FD. FPCA extends the concept of principal component analysis (PCA) to multivariate data and utilizes a linear combination of a finite number of functional principal components (FPCs) to approximate the infinite-dimensional true random function, thereby achieving dimension reduction.

Modelling FD using principal components (PCs) has long been a popular method, known as reduced-rank mixed-effects models in the FDA literature. Specifically, a ran-

dom function is expressed as the sum of its mean and the random deviation from the mean function. The mean function is modelled using a linear combination of fixed basis functions, such as B-spline basis functions, while the random deviation is represented as a linear combination of FPCs with random coefficients. Multivariate normality is commonly assumed for the joint distribution of the vector of random coefficients to facilitate model fitting. This model is widely employed to study the dependence between multiple curves. For example, Zhou et al. (2008) applied this reduced-rank mixed-effects model to investigate the dependence of paired sparse FD. Furthermore, Zhou et al. (2010) proposed reduced-rank mixed-effects models for spatially correlated hierarchical FD, while Li et al. (2014) extended this model to accommodate paired longitudinal data with simultaneous measurements of continuous and binary outcomes. Liu et al. (2017) utilized this model to conduct FPC analysis for spatially correlated FD. However, it is important to note that the use of FPCs may lead to computational complexity as they need to be estimated from the observed data. Recently, Zhang and Li (2022) proposed unified model of spatial functional data (SFD) through FPCA.

Moreover, a lack of justification for the assumption of multivariate normality is another concern regarding this reduced-rank model. Additionally, a bivariate Gaussian distribution is symmetric, meaning that positive and negative extreme events have the same probability distribution. Such symmetry does not always appear to hold in practice. For example, the return distribution of stocks is typically not empirically symmetric. Although extreme events have low probabilities, their occurrence can have severe consequences. Therefore, using a Gaussian distribution when asymmetric tail distributions are desired is inappropriate. Furthermore, a bivariate Gaussian distribution cannot accommodate tail dependence. According to page 208 in McNeil et al. (2015), the coefficient of upper tail dependence between a pair of random variables (X, Y) is defined as $\lim_{u\to\infty} \mathbb{P}(Y > u \mid X > u)$ when the limit exists. The coefficient of lower tail dependence can be similarly defined. For a bivariate Gaussian copula, this coefficient must be zero unless the two components are perfectly correlated (Embrechts et al., 2001). This implies that under a bivariate Gaussian copula, the components behave in the extreme tails as if they were either independent or perfectly correlated. Staicu et al. (2012) developed models and inferential methods for the analysis of functional skewed data, where the marginal distribution of observations

made at each time point is not Gaussian, but their joint dependence is characterized by a multivariate Gaussian copula. Cao et al. (2019) extended the model proposed by Staicu et al. (2012) to investigate the dependence structure of multiple functional outcomes. In these articles, the Gaussian copula is employed to separate the dependence structure from the marginal distributions. Skewed distributions such as the skew-t or the skew-normal distributions (Azzalini and Capitanio, 2003; Azzalini and Valle, 1996) are allowed for the marginal distributions, while the dependence structure is still characterized by the Gaussian copula. Therefore, the models developed in these papers still cannot accommodate tail asymmetry or tail dependence.

In this chapter, we are concerned about spatio-temporal dependence within FD. Our main contribution is that we develop a factor copula model which will partially abandon the normality assumption. We first represent the FDs that are collected from a number of different locations where the set of time points is smooth. We then express these functional observations by basis functions. In place of the traditional white noise term, as in Krupskii and Genton (2017), we assume that there is a common factor which has the same effect on all the spatial locations. This common factor plays an important role in that it allows the model to capture the dependence structure without the normality assumption on the noise.

The rest of the chapter is given as follows. Section 4.2 presents the factor copula model for SFD. Section 4.3 describes the estimation procedure, and a way to get a closed form expression for the likelihood estimation. We provide simulation studies to evaluate the performance using the closed-form expression in Section 4.4, as well as certain misspecified cases. We close the chapter with concluding remarks, followed by areas of future work, in Section 4.5.

4.2 Model

4.2.1 Copula models for spatial data

In a recent study, Krupskii et al. (2018) presented a copula model that offers an alternative approach to modelling replicated spatial data. Unlike the multivariate Gaussian copula, Krupskii's proposed copula model incorporates a common factor that is present across all locations in the spatial process. Specifically, consider a (scalar) random field denoted by $\{\boldsymbol{X}(\boldsymbol{s}),\ \boldsymbol{s} \in \mathbb{R}^d\}$, the common factor copula model is expressed as follows:

$$\boldsymbol{X}(\boldsymbol{s}) = \boldsymbol{Y}(\boldsymbol{s}) + V_0, \tag{4.1}$$

where $\boldsymbol{Y}(\boldsymbol{s})$ is a Gaussian random field and V_0 denotes the common factor. In addition, assuming that $Y(\boldsymbol{s})$ is a stationary Gaussian process, its corresponding correlation matrix can be parameterized using either an exponential or Matérn function, leading to a reduction in the number of parameters required to specify the model. Notably, Proposition 1 in Krupskii et al. (2018) suggests that the proposed copula can effectively model tail dependence and tail asymmetry of $\boldsymbol{X}(\boldsymbol{s})$ if V_0 meets certain mild conditions. For situations involving multiple variables of interest, Krupskii and Genton (2019) introduced a non-Gaussian copula capable of modelling the dependence structure of replicated multivariate spatial data. Moreover, to address spatio-temporal data, Krupskii and Genton (2017) proposed an extension of model (4.1), accommodating repeated observations measured over time across distinct spatial locations.

In this setting, there exists a common factor that affects the measurements of a spatio-temporal process. The proposed copula can still enjoy the ability that it may model the data with tail dependence and asymmetry. They justified that a simple form for the likelihood is available and therefore maximum likelihood estimates are immediately available.

Introducing a common factor in the aforementioned work eliminates the need for multivariate normality assumptions, resulting in the ability to accommodate tail dependence and asymmetry under appropriate conditions. Motivated by these findings, our objective is to extend this model to the realm of SFD analysis.

4.2.2 Factor copula models for spatial functional data

To the best of our knowledge, using a common factor copula to model dependence of multiple random functions to achieve tail asymmetry or dependence has not been investigated in the literature. As illustrated in Section 4.1, the random vector of the coefficients in a reduced rank model is usually assumed to follow a multivariate Gaussian distribution. Although a multivariate normality assumption is quite common in practice, sometimes it may fail to fit the data sufficiently well, especially when investigating tail asymmetry and tail dependence is of primary interest. Next, we drop the multivariate normality assumption, and accommodate the common factor copula model for SFD.

To characterize spatial correlation of random functions observed at different locations, we borrow the idea of the reduced rank model while assuming a spatial correlation structure for the multivariate normally distributed random coefficient vector. To accommodate tail asymmetry and tail dependence, we introduce a common factor in the representation of SFD. As SFD are actually spatial data that are observed repeatedly in time at each location, the common factor in model (4.1) is indeed time-dependent in the new setting. To fix ideas, let $X(\boldsymbol{s}_k; t)$ denote a discrete observation at time t on a random curve sampled at location $\boldsymbol{s}_k$, $k = 1, \ldots, K$. We assume that $X(\boldsymbol{s}_k; t)$ (as a function of time t) can be represented as

$$X(\boldsymbol{s}_k; t) = Y(\boldsymbol{s}_k; t) + V_0(t), \tag{4.2}$$

where $Y(\boldsymbol{s}_k; t)$ is a spatio-temporal process defined on $(\boldsymbol{s}_k, t) \in \mathcal{D} \times \mathcal{T}$, and $V_0(t)$ is a common process that can affect all locations equally at time t but it does not dependent on the spatial location $\boldsymbol{s}_k$. Note that in model (4.2), there is no error term that usually represents the noise or measurement error in a typical FDA model. We note that measurment error could also be a factor in obtaining observations of the form (4.2), although, as in Krupskii and Genton (2017) and Krupskii et al. (2018), we do not consider that problem here. Instead, we let the common process $V_0(t)$ replace the traditional error term $\epsilon(t)$. A remarkable difference between the common process $V_0(t)$ and the error term $\epsilon(t)$ is that observed functions at all locations are affected by the same process $V_0(t)$. Note that $V_0(t)$ does not need to be Gaussian, whereas $\epsilon(t)$ is usually assumed to be a Gaussian process or independent for different t's.

Assuming that spaital dependence is second-order stationary and isotropic. When the times of two observations are fixed, the isotropy condition states that the correlation between any two observations depends solely on distance and not on their relative orientation. This helps us infer the spatial dependence structure from a single replicate of the spatio-temporal process. The general covariance function of $Y(\boldsymbol{s}_k; t)$ can be expressed as

$$R(\|\boldsymbol{s}_{k_1} - \boldsymbol{s}_{k_2}\|, t_1, t_2) = \mathbb{C}\text{ov}(Y(\boldsymbol{s}_{k_1}; t_1), Y(\boldsymbol{s}_{k_2}; t_2)),$$

for any $(\boldsymbol{s}_{k_1}, t_1), (\boldsymbol{s}_{k_2}, t_2) \in \mathcal{D} \times \mathcal{T}$. In addition, consider $Y(\boldsymbol{s}_k; t)$ as spatial replicates of a temporal process with a standard Karhunen-Loève expansion

$$Y(\boldsymbol{s}_k; t) = \mu(t) + \sum_{\ell=1}^{\infty} \xi_{k\ell} f_\ell(t), \tag{4.3}$$

where $\mu(t) = \mathbb{E}Y(\boldsymbol{s}; t)$ is the mean function taken over all locations, $f_\ell(\cdot)$'s are orthonormal functions referred to as the FPCs, and FPC scores

$$\xi_{k\ell} = \xi_\ell(\boldsymbol{s}_k) = \int_{\mathcal{T}} \{Y(\boldsymbol{s}_k; t) - \mu(t)\} f_\ell(t) dt,$$

is the loading of $Y(\boldsymbol{s}_k; t)$ on the ℓ-th principal component. We assume that $\{\xi_{k\ell}\}$ are zero-mean, second-order stationary and isotropic random fields that are uncorrelated across ℓ. Spatial dependene among the functional data is induced by dependence with each $\xi_{k\ell}$. Denote the spatial covariance function of $\xi_{k\ell}$ as

$$C_\ell(\|\boldsymbol{s}_{k_1} - \boldsymbol{s}_{k_2}\|) = \mathbb{C}\text{ov}(\xi_\ell(\boldsymbol{s}_{k_1}), \xi_\ell(\boldsymbol{s}_{k_2})),$$

for any $(\boldsymbol{s}_{k_1}, t_1), (\boldsymbol{s}_{k_2}, t_2) \in \mathcal{D} \times \mathcal{T}$. Then the covariance function for $Y(\boldsymbol{s}_k; t)$ can be written as

$$\begin{aligned} R(\|\boldsymbol{s}_{k_1} - \boldsymbol{s}_{k_2}\|, t_1, t_2) &= \mathbb{C}\text{ov}\left(\sum_{k=1}^{\infty} \xi_\ell(\boldsymbol{s}_{k_1}) f_\ell(t_1), \sum_{k=1}^{\infty} \xi_\ell(\boldsymbol{s}_{k_2}) f_\ell(t_2)\right) \\ &= \sum_{k=1}^{\infty} C_\ell(\|\boldsymbol{s}_{k_1} - \boldsymbol{s}_{k_2}\|) f_\ell(t_1) f_\ell(t_2). \end{aligned}$$

To reduce the complexity of (4.3), we employ the reduced rank model proposed by James et al. (2000) to approximate the random function $Y(\boldsymbol{s};\cdot)$ by its first L FPCs $f = (f_1(t), f_2(t), \ldots, f_L(t))^\top$. Then $Y(\boldsymbol{s}_k; t)$ can be approximately expressed as

$$Y(\boldsymbol{s}_k; t) = \mu(t) + \sum_{\ell=1}^{L} \xi_{k\ell} f_\ell(t). \tag{4.4}$$

Following James et al. (2000), we approximate the mean curve $\mu(t)$ with the regression spline. Specifically, $\mu(t)$ is written as a linear combination of B-spline basis functions $(b_1(t), \ldots, b_J(t))^\top$. To represent each FPC, we employ the same set of B-spline basis functions. Then $Y(\boldsymbol{s}_k; t)$ in model (4.4) can be rewritten as

$$Y(\boldsymbol{s}_k; t) = \sum_{i=1}^{J} v_i b_i(t) + \sum_{i=1}^{J} \alpha_{ki} b_i(t),$$

where v_i, $i = 1, \ldots, J$, are fixed coefficients that are independent of locations, and α_{ki} is a linear combination of $\xi_{k1}, \ldots, \xi_{kL}$. Let $\boldsymbol{\alpha}_k = (\alpha_{k1}, \ldots, \alpha_{kJ})^\top, k = 1, \ldots, K$. We assume that the joint distribution of $(\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_K^\top)^\top$ follows a multivariate Gaussian with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_\alpha$. As spatial dependence among the random function $Y(\boldsymbol{s}_k; t)$ is induced by the dependence within each of the random coefficients ξ_k, we assume the Σ_α is parameterized by a powered exponential spatial correlation structure. More specifically, we assume that

$$\mathbb{C}\mathrm{ov}\left(\alpha_{k_1 i}, \alpha_{k_2 i}\right) = \exp\left\{-\left(\frac{d(\boldsymbol{s}_{k_1}, \boldsymbol{s}_{k_2})}{\theta_i}\right)^{\gamma_i}\right\} \sigma_i^2, \tag{4.5}$$

where $\theta_i > 0$ is a scale parameter, $0 < \gamma_i \leq 2$ is a shape parameter, $i = 1, \ldots, J$ and $d(x, x')$ denotes a distance measure between locations x and x'. In this chapter, we employ the Euclidean distance metric. However, alternative distance measures such as the Manhattan distance and Minkowski distance can also be utilized. We choose to use the Euclidean distance due to its intuitive geometric interpretation and widespread usage as a default metric in many applications. However, it is important to note that it may not always be the most optimal choice in certain scenarios. For further discussion on distance metrics, we refer to Shahid et al. (2009). For the coefficients of different B-spline bases (i.e., $i \neq j$), we assume that $\mathbb{C}\mathrm{ov}(\alpha_{k_1 i}, \alpha_{k_2 j}) = 0$, regardless of the locations k_1 and k_2. Consequently,

for any two time points t_1, t_2, we have

$$\mathbb{C}\text{ov}(Y(\boldsymbol{s}_k;t_1),Y(\boldsymbol{s}_k;t_2)) = \sum_{i=1}^{J} b_i(t_1)b_i(t_2)\,\mathbb{C}\text{ov}(\alpha_{k_1 i},\alpha_{k_2 i}). \tag{4.6}$$

The common process $V_0(t)$ is the key component that distinguishes our model from conventional FD models. Following Krupskii and Genton (2017), we adopt a common process with the form

$$V_0(t) = \beta(t)\mathcal{E}_{N(t)}, \tag{4.7}$$

where $\beta(t)$ is a non-random time dependent coefficient function, and $N(t)$ is a Poisson process with the rate $\lambda(t)$ controlling the dependence of model (4.2) in the time domain. If the Poisson process does not have a jump between t_i and t_{i+1}, then $\mathcal{E}_{N(t_i)} = \mathcal{E}_{N(t_{i+1})}$. However, when there is indeed a jump between t_i and t_{i+1}, these two random variables are assumed to be independent with the same distribution as $\mathcal{E}_0$. According to the property of a Poisson process, $\mathbb{P}\left(N(t_i) = N(t_{i+1})\right) = \exp\{-\int_{t_i}^{t_{i+1}} \lambda(t)dt\}$. Therefore, if the interval between two time points is sufficiently large, the common process evaluated at these two points is less likely to have the same common factor, which implies lower dependence. Similarly, when the average rate $\lambda(t)$ between two time points is large, it is also less likely to have the same common factor, therefore giving rise to lower dependence.

As shown in Krupskii et al. (2018), different designs for the common factor $\mathcal{E}_0$ can lead to different dependence structures. For example, when the common factor $\mathcal{E}_0$ follows a Weibull distribution, the tail property of the copula depends on the shape parameter; when the common factor $\mathcal{E}_0$ is the difference between two exponential factors with different means, the resulting copula is asymmetric and there is a closed form for the upper tail dependence coefficient; when the common factor $\mathcal{E}_0$ follows a Pareto distribution, the resulting copula is asymmetric with perfect upper tail dependence. In this chapter, we further study model (4.2) with $Y(\boldsymbol{s}_k;\cdot)$ specified through (4.4) and V_0 through (4.7) where the factor $\mathcal{E}_0$ follows an exponential distribution or a Pareto distribution.

Let $\mathcal{T} = \{t_1,\ldots,t_T\}$ denote a (common) set of time points when the FD are observed. This assumption can be relaxed to allow for different observational time points for random functions at different locations. Let $F_{\boldsymbol{X}}(\cdot)$ and $f_{\boldsymbol{X}}(\cdot)$ denote the joint cumulative

distribution function (cdf) and probability density function (pdf) of the vector

$$\boldsymbol{X} = (X(\boldsymbol{s}_1; t_1), \ldots, X(\boldsymbol{s}_K; t_1), \ldots, X(\boldsymbol{s}_1; t_T), \ldots, X(\boldsymbol{s}_K; t_T))^\top,$$

respectively. Note that we rearrange the random variables $X(\boldsymbol{s}_k, t_j)$'s in this order such that observations that are made at the same time points are gathered together. Let $F_{kj}(x)$ and $f_{kj}(x)$ denote the marginal cdf and pdf of $X(\boldsymbol{s}_k, t_j)$, respectively. Then the copula $C(\boldsymbol{u})$ of $\boldsymbol{X}$ and its density $c(\boldsymbol{u})$ can be written as:

$$C_{\boldsymbol{X}}(\boldsymbol{u}) = F\left(F_{11}^{-1}(u_{11}), \ldots, F_{KT}^{-1}(u_{KT})\right) \tag{4.8}$$

and

$$c_{\boldsymbol{X}}(\boldsymbol{u}) = \frac{f\left(F_{11}^{-1}(u_{11}), \ldots, F_{KT}^{-1}(u_{KT})\right)}{f_{11}(F_{11}^{-1}(u_{11})) \ldots f_{KT}(F_{KT}^{-1}(u_{KT}))}, \tag{4.9}$$

respectively, where $\boldsymbol{u} = (u_{11}, \ldots, u_{KT})^\top \in [0,1]^{KT}$.

4.3 Estimation

In this section, we show that the likelihood function may be written in a closed form when the common factor $\mathcal{E}_0$ follows an exponential distribution, and then estimate parameters through the likelihood function. This allows for fast estimation even when the number of spatial locations is fairly large.

Based on model (4.2), one needs to know $Y(\boldsymbol{s}_k; t_j)$ and $V_0(t_j)$ for $k = 1, \ldots, K$ and $j = 1, \ldots, T$ in order to obtain the exact likelihood function. However, neither of them are directly observable, so we employ a conditioning technique to find the likelihood function. Let $\boldsymbol{N} := (N_1, \ldots, N_T)^\top$ denote the collection random observations of the Poisson process $N(t)$ at each time in the set $\mathcal{T}$, and thus $\boldsymbol{N}$ forms a T-dimensional Poisson random vector. Let us denote a realization of the random vector $\boldsymbol{N}$ by $\overline{\boldsymbol{N}} = (\overline{N}_1, \ldots, \overline{N}_T)^\top$ and the collection of all possible realizations of $\boldsymbol{N}$ by $\mathscr{H}$. By conditioning on $N(t_1) = 1$, there are $T-1$ points remaining random each representing a specific time instance that may or may not have associated events. At each of these points, there are two possible outcomes: a

jump or no jump. Consequently, the cardinality of the resulting set $\mathscr{H}$ is 2^{T-1}. Then, we may calculate the joint pdf of $\boldsymbol{X}$ using the total law of probability as

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \sum_{\overline{\boldsymbol{N}} \in \mathscr{H}} f_{\boldsymbol{X}|\overline{\boldsymbol{N}}}(\boldsymbol{x} \mid \boldsymbol{N} = \overline{\boldsymbol{N}}) \mathbb{P}(\boldsymbol{N} = \overline{\boldsymbol{N}}). \tag{4.10}$$

Without loss of generality, we assume $N_1 = 1$, and the jump size is at most one between any adjacent time points. As a result, $\mathbb{P}(N_j = N_{j+1}) = \exp\{-\int_{t_j}^{t_{j+1}} \lambda(t)dt\}$, and $\mathbb{P}(N_j \neq N_{j+1}) = 1 - \exp\{-\int_{t_j}^{t_{j+1}} \lambda(t)dt\}$. For every realization $\overline{\boldsymbol{N}} \in \mathscr{H}$, we have

$$\begin{aligned}
\mathbb{P}(\overline{\boldsymbol{N}} = l) &= \mathbb{P}\left(N(t_2) = N_2, \cdots, N(t_T) = N_T \mid N(t_1) = 1\right) \\
&= \prod_{j: N_j = N_{j+1}} \exp\left\{-\int_{t_j}^{t_{j+1}} \lambda(t)dt\right\} \prod_{j: N_j \neq N_{j+1}} \left(1 - \exp\left\{-\int_{t_j}^{t_{j+1}} \lambda(t)dt\right\}\right).
\end{aligned}$$

Let $\Phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\cdot)$ denote the cdf of the multivariate normal distribution with covariance matrix $\boldsymbol{\Sigma}$ and mean vector $\boldsymbol{\mu}$. Let $\boldsymbol{x}_t = (x(s_1;t), \ldots, x(s_K;t))^\top$. Given the history information of $N(t)$ and assuming that $\mathcal{E}_0 \sim \exp(1)$, the conditional distribution of $\boldsymbol{X}$ is given by

$$\begin{aligned}
&f_{\boldsymbol{X}|\overline{\boldsymbol{N}}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \mid \overline{\boldsymbol{N}}) \\
&= \int_{\mathbb{R}_+^m} \Phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}_1 - \xi_1\boldsymbol{\beta}_1, \ldots, \boldsymbol{x}_T - \xi_T\boldsymbol{\beta}_T) \exp\left(-\sum_{i=1}^m \xi_i\right) d\xi_1 \cdots d\xi_m,
\end{aligned}$$

where ξ_j is the value of the j-th common factor $\mathcal{E}_j$, $\boldsymbol{\beta}_j = \beta(t_j) \cdot \mathbb{1}_K$ (i.e., a vector of ones of length K) and $m = \max \overline{\boldsymbol{N}}$. To simplify the notation, since we know that $m \leq T$, some of the terms $\xi_1, \ldots, \xi_T$ have the same value. We combine the terms that have the same differential of the variable together and obtain

$$\begin{aligned}
&f_{\boldsymbol{X}|\overline{\boldsymbol{N}}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T \mid \overline{\boldsymbol{N}}) \\
&= \int_{\mathbb{R}_+^m} \Phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}_1 - \xi_1\boldsymbol{\beta}_1, \ldots, \boldsymbol{x}_m - \xi_m\boldsymbol{\beta}_m) \exp\left(-\sum_{i=1}^m \xi_i\right) d\xi_1 \cdots d\xi_m \\
&= (2\pi)^{-TK/2} |\boldsymbol{\Sigma}|^{-1/2} \int_{\mathbb{R}_+^m} \exp\left\{h(\xi_1, \ldots, \xi_m)\right\} d\xi_1 \cdots d\xi_m,
\end{aligned} \tag{4.11}$$

where

$$
\begin{aligned}
&h(\xi_1, \ldots, \xi_m) \\
&= -\frac{1}{2}\left((\widetilde{\boldsymbol{x}}_1 - \xi_1\widetilde{\boldsymbol{\beta}}_1)^\top \cdots (\widetilde{\boldsymbol{x}}_m - \xi_m\widetilde{\boldsymbol{\beta}}_m)^\top\right)^\top \boldsymbol{\Sigma}^{-1}\left((\widetilde{\boldsymbol{x}}_1 - \xi_1\widetilde{\boldsymbol{\beta}}_1)\cdots(\widetilde{\boldsymbol{x}}_m - \xi_m\widetilde{\boldsymbol{\beta}}_m)^\top\right) \\
&- \sum_{i=1}^{m}\xi_i \\
&= C_0 - \frac{1}{2}(\xi_1 - \xi_1^* \cdots \xi_m - \xi_m^*)^\top \Sigma_*^{-1}(\xi_1 - \xi_1^* \cdots \xi_m - \xi_m^*).
\end{aligned} \tag{4.12}
$$

In the above formula, we let $\widetilde{\boldsymbol{x}}_k$ denote the subvector of $\boldsymbol{w} = \boldsymbol{x} - \boldsymbol{\mu} = ((\boldsymbol{x}_1 - \boldsymbol{\mu}_1)^\top, \ldots, (\boldsymbol{x}_m - \boldsymbol{\mu}_m)^\top)^\top = (\boldsymbol{w}_1^\top, \cdots, \boldsymbol{w}_m^\top)^\top$, in which $\widetilde{\boldsymbol{w}}_k$ includes all $\boldsymbol{w}_j$'s such that $N_j = k$, $k = 1, \ldots, m$, and $\widetilde{\boldsymbol{\beta}}_k$'s are defined similarly. Note that equation (4.12) contains two quadratic functions of $\xi_1, \ldots, \xi_m$. By equating the coefficients of $\xi_1, \ldots, \xi_m$ in the two functions, we have

$$
\begin{aligned}
\Sigma_*^{-1} = \boldsymbol{B}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{B}, \qquad \boldsymbol{\xi}^* = \boldsymbol{\Sigma}_*\left(-\mathbb{I}_m + \boldsymbol{B}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)\mathbb{1}_m, \\
C_0 = -\frac{1}{2}\left((\boldsymbol{X}\mathbb{1}_m)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\mathbb{1}_m - (\boldsymbol{B}\boldsymbol{\xi}^*)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{\xi}^*\right),
\end{aligned}
$$

where $\boldsymbol{B}$ and $\boldsymbol{X}$ are $KT \times m$ matrices such that the kth column is equal to $\boldsymbol{\beta}_k^*$ ($\boldsymbol{w}_k^*$, respectively) that can be obtained by replacing all element except for $\boldsymbol{\beta}_k$ ($\boldsymbol{w}_k$) in $\boldsymbol{\beta}$ ($\boldsymbol{w}$) by zeros. That is

$$
\boldsymbol{\beta}_k^* = (\mathbf{0}_1^\top, \mathbf{0}_2^\top, \cdots, \widetilde{\boldsymbol{\beta}}_k^\top, \cdots, \mathbf{0}_T^\top)^\top
$$

and

$$
\boldsymbol{B} = \begin{pmatrix} \boldsymbol{\beta}_1^* & \boldsymbol{\beta}_2^* & \cdots & \boldsymbol{\beta}_m^* \end{pmatrix} = \begin{pmatrix} \widetilde{\boldsymbol{\beta}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \widetilde{\boldsymbol{\beta}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \widetilde{\boldsymbol{\beta}}_m \end{pmatrix}.
$$

Furthermore, I_m denotes an m-dimensional identity matrix. The details may be found in the Appendix C.1. With the notation, we simplify the integral in (4.11) as

$$
f_{\boldsymbol{X}|\overline{\boldsymbol{N}}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \mid \overline{\boldsymbol{N}}) = (2\pi)^{(m-KT)/2}|\boldsymbol{\Sigma}|^{-1/2}|\boldsymbol{\Sigma}_*|^{1/2}\exp(C_0)\Phi_{\boldsymbol{\Sigma}_*}(\xi_1^*, \cdots, \xi_m^*), \tag{4.13}
$$

where ξ_i^*, $i \in \{1, \ldots, m\}$ represents the corresponding index of the i-th entry in the $\boldsymbol{\xi}^*$ vector.

Since the number of time points in the time interval $\mathcal{T}$ is finite in practice, the number of possible outcomes of the Poisson process $N(t)$ evaluated at times points should also be finite. Having the above expressions, we may rewrite (4.10) as

$$f_{\boldsymbol{X}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) = \sum_{l \in \mathscr{H}} f_{\boldsymbol{X}|\overline{\boldsymbol{N}}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \mid \overline{\boldsymbol{N}} = l)\mathbb{P}(\overline{\boldsymbol{N}} = l).$$

The marginal distribution of $X(\boldsymbol{s}_k; t_i)$ can also be computed explicitly when the common factor $\mathcal{E}_0$ follows an exponential distribution. Generally, it is assumed that $Y(\boldsymbol{s}_k; t_j) \sim \mathcal{N}(\mu(t_j), \tau(t_j))$, and the coefficient of the common factor is β. To simplify the notation, we denote $\mu_j := \mu(t_j)$ and $\tau_j := \tau(t_j)$. We may then express the marginal distribution of $Y(\boldsymbol{s}_k; t_j)$ as

$$\begin{aligned} &F_{kj}(x \mid \mu_j, \tau_j, \beta_j) := F(x_k(t_j) \mid \mu_j, \tau_j, \beta_j) \\ &= \Phi\left(\frac{x_k(t_j) - \mu_j}{\tau_j}\right) - \exp\left(\frac{\tau^2}{2\beta_j^2} - \frac{x_k(t_j) - \mu_j}{\beta_j}\right)\Phi\left(\frac{x_k(t_j) - \mu_j}{\tau_j} - \frac{\tau_j}{\beta_j}\right), \end{aligned} \tag{4.14}$$

where $\Phi(\cdot)$ is the cdf of the univariate standard normal distribution. Consequently, the probability density function (PDF) is

$$\begin{aligned} f_{kj}(x \mid \mu_j, \tau_j, \beta_j) &= \frac{1}{\tau_j}\phi\left(\frac{x_k(t_j) - \mu_j}{\tau_j}\right) \\ &+ \frac{1}{\beta_j}\exp\left(\frac{\tau_j^2}{2\beta_j^2} - \frac{x(t_j) - \mu_j}{\beta_j}\right)\Phi\left(\frac{x_k(t_j) - \mu_j}{\tau} - \frac{\tau_j}{\beta_j}\right) \\ &- \exp\left(\frac{\tau_j^2}{2\beta_j^2} - \frac{x(t_j) - \mu_j}{\beta_j}\right)\phi\left(\frac{x(t_j) - \mu_j}{\tau_j} - \frac{\tau_j}{\beta_j}\right)\frac{1}{\tau_j}. \end{aligned} \tag{4.15}$$

If we have N i.i.d. replicates, $\boldsymbol{x}_i = (x_{11,i}, x_{12,i}, \ldots, x_{1T,i}, x_{21,i}, \ldots, x_{KT,i})$, where $x_{kj,i} \coloneqq x(s_k, t_j)$ and $\boldsymbol{x} = (\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \ldots, \boldsymbol{x}_N^\top)^\top$. Then, the log-likelihood function in terms of the copula density can be expressed as

$$l(\boldsymbol{u} \mid \beta, \lambda, \gamma, \theta) = \sum_{i=1}^{N} \log\left(\sum_{l \in \mathscr{H}} f_{\boldsymbol{X}|\overline{\boldsymbol{N}}}(\boldsymbol{x}_{1,i}, \ldots, \boldsymbol{x}_{m,i} \mid \overline{\boldsymbol{N}} = l)\mathbb{P}(\overline{\boldsymbol{N}} = l)\right) \tag{4.16}$$

$$-\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{\top}\log\left(f_{kj}(x_{kj,i})\right). \tag{4.17}$$

To employ the copula density to define the above log-likelihood function, the input should be a vector $\boldsymbol{u} = (u_{11}, \ldots, u_{KT})^\top$ with $u_{jk} \in [0,1]$, instead of the raw observed data $\ell_{11}, \ldots, \ell_{KT}$. One may use the empirical CDF to do the inversion as $u_{kj} = \hat{F}(\ell_{kj})$. However, this require the estimate empirical distribution function $\hat{F}(\cdot)$. When having N replicates, an alternative way is to use the nonparametric score function as follows:

$$u_{kj}^{(n)} := \frac{\text{Rank}(\ell_{jk}^{(n)}) - 0.5}{N}, \quad n = 1, \ldots, N.$$

Then, the observations $X_{kj}^{(n)}$ may be obtained by taking the CDF transformation $X_{kj}^{(n)} = (F_{ij})^{-1}(u_{kj}^{(n)})$.

4.4 Simulation Studies

In this section we present the results of several simulation studies that serve the purpose of evaluating the parameter estimation of the model described in Sections 4.2 and 4.3. We first introduce the data generating process as follows:

Step 1: Specify the number of locations K and number of time points T. Then, simulate K locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K$ over a two dimensional lattice $[0,1] \times [0,1]$.

Step 2: Calculate the covariance matrix Σ_Y. We assume that it is separable so that $\Sigma_Y = \Sigma_{loc} \times \Sigma_{time}$. We note here that we assume that the SFD are fully observed and so we do not take into consideration the presence of measurement errors.

Step 3: Simulate the Gaussian component Y where $Y \sim \mathcal{N}(\mathbf{0}_{K\times T}, \Sigma_Y)$.

Step 4: Simulate the common factor using the exponential distribution with parameter value 1.

Step 5: We then calculate the probability of no jump using the formula $\exp\{-\lambda * \delta t\}$.

Step 6: Simulate a uniform random variable ℓ where $\ell \sim \text{Unif}(0, 1)$.

In Step 4, we are able to compute the true covariance/correlation matrix Σ_{loc} parameterized by the separable correlation structure presented in (4.5). Next, with the T number of points, we equally discretize the time points as $\mathcal{T} = (1, \ldots, T)^\top$. Alternatively, we may parameterize those points to a unit length $[0, 1]$. Along with the pre-specified B-spline basis functions for both mean and random components $\boldsymbol{\xi}(t_i) = (b_1(t_i), \ldots, b_L(t_i))^\top$, we may calculate the correlation matrix using (4.6). Using this correlation matrix, we simulate the Gaussian part of the data (i.e. Y) N times.

In Step 4, to simulate the common factor, we need to first simulate a Poisson process $\overline{\boldsymbol{N}}$. Under the assumption that there could only be one jump between any two adjacent time points, the event that there is a jump between any two adjacent time points has a probability $p = 1 - \exp\{-\int_{t_i}^{t_{i+1}} \lambda(t)dt\}$. This helps us to simplify the simulation of Poisson process to the simulation of $4 \times N$ i.i.d. Bernoulli samples with probability p. Following this, we generate $(T-1)$ i.i.d. samples following an exponential distribution with parameter 1 for each time point from 2 to T. If the probability is low, there is no jump and the value of the common factor at current time t_i is the same as the value at previous time point t_{i-1}. If the simulated probability is high, it correspondes to a jump, and then we simulate a new value from exponential 1. We repeat this procedure for a total $Nsim$ times.

There are many values that need to be pre-specified, namely 1) the number of B-spline basis functions for the mean curve $\mu(t)$ and the random noise around the mean curve. We set J to be 5. 2) The coefficient vector for the mean curve $\boldsymbol{\xi}$ is set to be $(0.7, -0.5, 0.6, 0.5, -0.3)^\top$. As a consequence, the covariance matrix $\boldsymbol{\Sigma}$ has dimension 25×25. 3) The number of locations K we choose to be 10, which are lying in a subspace of $\mathbb{R}^2$. 4) The number of time points T is set to be 5 and $\mathcal{T} = \{t_1, \ldots, t_5\}$. The coefficients for the exponential common factor are $\beta(t)$ which controls the magnitude, and the intensity of the jump in $p = \exp\left(-\int_{t_i}^{t_{i-1}} \lambda(t)\right)$. We set the magnitude parameter $\beta(t)$ to be constant β, i.e., it is time invariant. We set the parameter vectors as $\Omega = (\sigma, \theta, \gamma, \beta, \lambda)^\top = (1.1, 0.6, 0.3, 1.2, 0.15)^\top$ and basis coefficient vector $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3, \xi_4, \xi_5)^\top = (0.7, -0.5, 0.6, 0.4, -0.3)^\top$. We first look at the estimates for the dependence vector Ω.

In Figures 4.1 and 4.2, we see that when the tail dependence is generated from the exponential distribution with shape parameter 1 (i.e. exp(1)), the estimates are very accurate for a small sample size of $N = 50$. When the sample size is large $N = 200$, as the boxplot indicates that of all dependence parameters are shrinking which indicates that the variation decreases. This is expected as we obtain the closed form expression of the likelihood function when the common factor is generated from exp(1). We also perform an additional simulation in the case when $\beta(t)$ is time-varying. In that case, all of the parameter estiamtes are still very close to the true values.

(a) $N = 50$ (b) $N = 200$

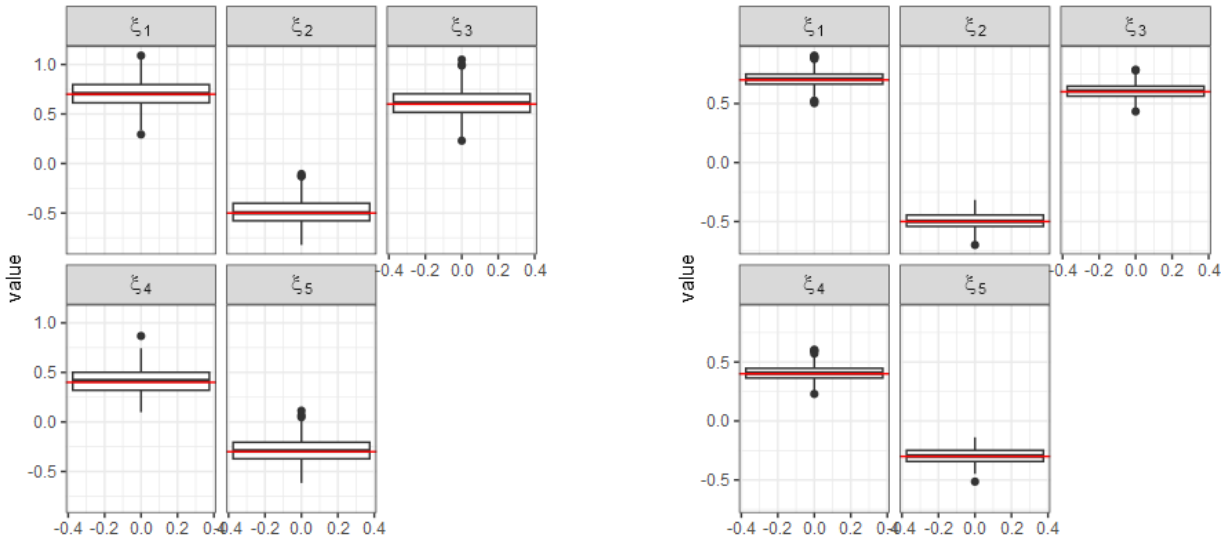


Figure 4.1: Boxplots for the the parameter estimates coefficients of the mean curve ξ_1, ξ_2, ξ_3, ξ_4 and ξ_5 with different sample sizes of N in 500 simulation runs when the jump process is generating from an exponential distribution with parameter.

We then take a look at the case when the model is misspecified. We use two different distributions to generate a common factor, namely the Gaussian and Pareto distributions. In particular, the Pareto distribution we use is with a scale parameter 1 and a shape parameter 2, denoted $\mathcal{E}_0 \sim \text{Pareto}(1, 2)$. In addition, we include the case where the common factor is generated from a standard Gaussian distribution $\mathcal{N}(0, 1)$ when there is no tail

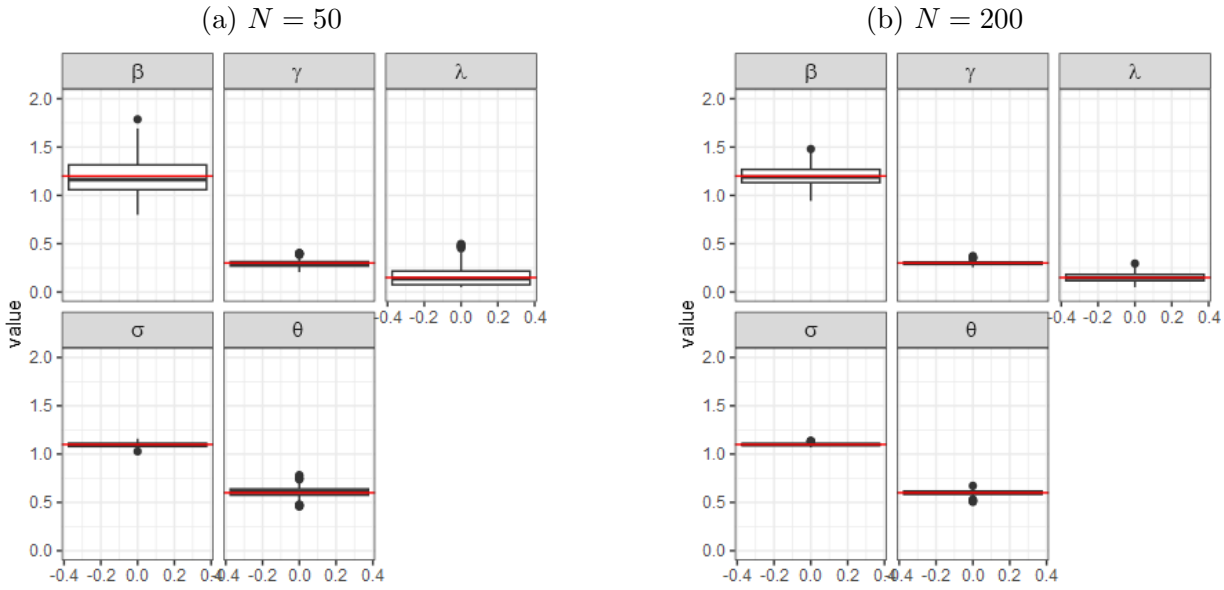


Figure 4.2: This figure presents the parameter estimates for σ, θ, γ, λ and β with different sample sizes of N in 500 simulation runs when the jump process is generating from an exponential distribution. The red lines denote the true parameter values.

distortion.

In Figure 4.3, we see that when the tail dependence is generated from different distributions, Pareto$(1, 2)$ or $\mathcal{N}(0, 1)$, the parameter estimates seem a bit off from the true values. In particular, for the mean vector, it seems that all of the estimates for $\boldsymbol{\xi}$ are overestimated for the Pareto, and underestimated for the Gaussian case. It becomes more apparent in Figure 4.4 that the mean estimates when the tail is generated using the exponential distribution result in a perfect fit, while the cases using the Pareto distribution or Gaussian distribution are inaccurate, despite that the shape of the mean curve is well captured and they differ from an affine transformation. As for the dependence parameters, γ, λ, σ and θ are each accurately estimated as shown in Figure 4.5. Notably, with a small sample size of $N = 50$, the estimates for those four parameters perform well, with only a few outliers that fall outside the interquartile range. When the sample size increases, for instance, to $N = 200$, this deviation is much improved and the true value is almost the same as the

estimates. This is not the case for β, however. Similar to the basis estimates, it seems that the model tends to overestimate the parameter β, and it does not change when the sample size increases. This suggests that the model is robust to the spatial correlation structure and model misspecification, and it captures the overall shape of the trend presented in the data. However, additional (future) work is necessary for finding some robust ways to estimate the magnitude of the trend.

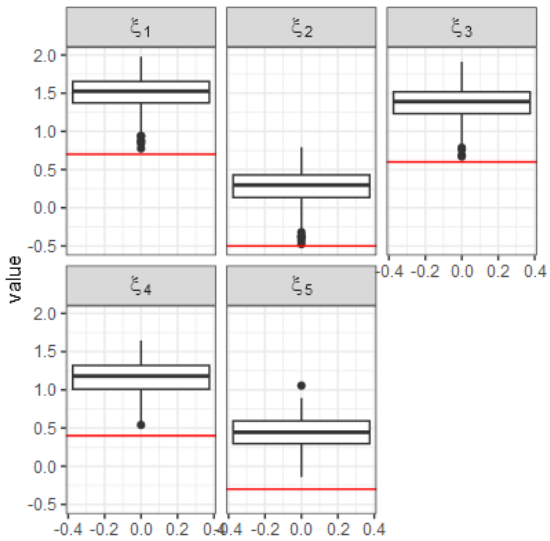
4.5 Concluding Remarks and Future works

This chapter introduced a model for FDA that challenges the traditional assumption of normality consistent with the copula paradigm. The proposed model builds upon FPCA and specifically targets SFD observed across multiple locations. The underlying function comprises a mean function, a variance (or covariance) function, and a common factor. The mean and variance functions are modeled using B-spline basis functions. Unlike previous models, the proposed model accommodates a common factor that affects all locations uniformly and does not require adherence to a Gaussian distribution. We anticipate that this model will effectively capture spatio-temporal dependence in FD. To examine the dependence structure, a new copula model is introduced, which exhibits desirable properties of tail dependence and asymmetry when tailored with specific common factor distributions. Furthermore, we derive a closed-form likelihood function, enabling parameter estimation through maximum likelihood estimation. Throughout simulation studies, we show that this model is able to capture the spatial correlation presented in the data, while capturing the shape of the temporal trend. The proposed model can accommodate various dependence structures by tuning parameters and modifying the distribution of common factors. In summary, this model is versatile for FD with spatio-temporal dependence and can be used to obtain desirable properties that are unattainable using the normality assumption by precluing the existence of heavy tails. Moreover, the parameter estimate for spatial dependence parameters converges fast to the true values even with a small sample size N.

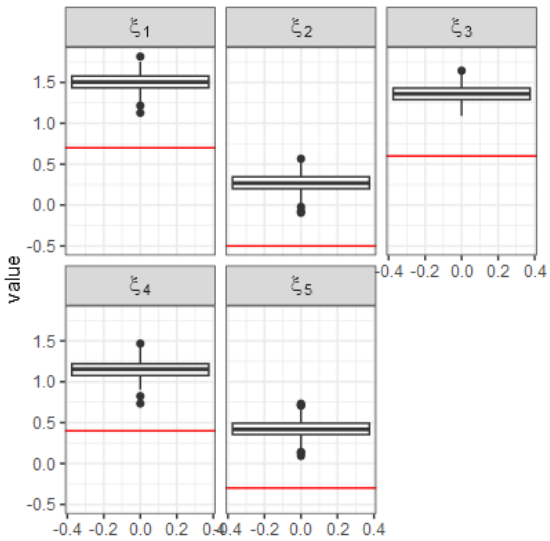
Regarding future work, we have several ideas. Firstly, we would like to incorporate learning or optimization techniques into the problem, such as mini-batch gradient descent.

These techniques can potentially improve the convergence speed of the optimization procedure while maintaining accuracy. Furthermore, they can allow practitioners to fine-tune the model by incorporating prior or domain knowledge, for example, by specifying the learning rate for each parameter separately. Additionally, it is desirable to explore the tail dependence of the copula. Specifically, deriving closed-form expressions for the upper and lower tail dependence coefficients, as well as investigating other distributions, would be valuable. Although the parameter estimation appears satisfactory, there is still potential to enhance the accuracy and efficiency of the algorithm. The number of B-spline basis functions can also be modified based on exploratory data analysis. Lastly, we aim to apply our method to real-life datasets, such as housing price data used in Zhang and Li (2022), climate datasets similar to those provided by the North American CORDEX program as utilized in French and Kokoszka (2021). By doing so, we hope to assist decision-makers in various applied fields and beyond.

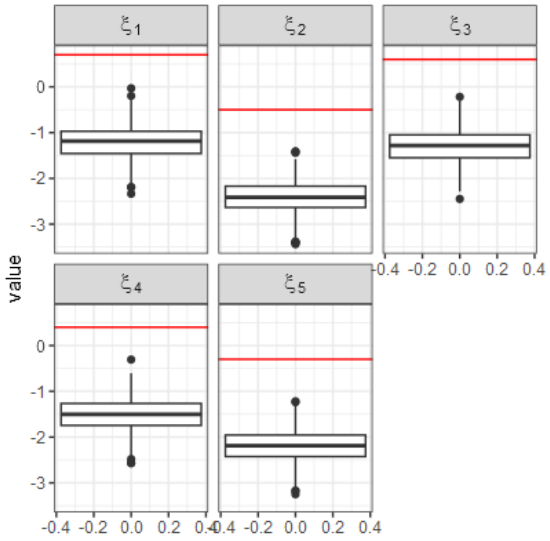
(a) Pareto, $N = 50$



(b) Pareto, $N = 200$



(c) Gaussian, $N = 50$



(d) Gaussian, $N = 200$

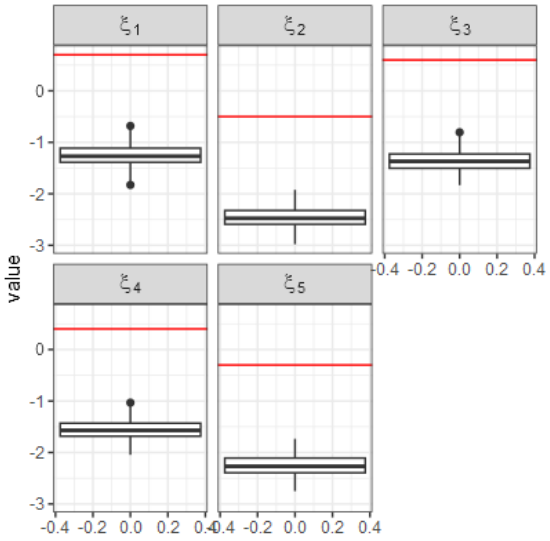


Figure 4.3: Boxplots for the the parameter estimates coefficients of the mean curve ξ_1, ξ_2, ξ_3, ξ_4 and ξ_5 with different sample sizes of N in 500 simulation runs when the jumping process is generating from **Top row**: the Pareto distribution and **Bottom row**: standard Gaussian distribution. The red lines denote the true parameter values.

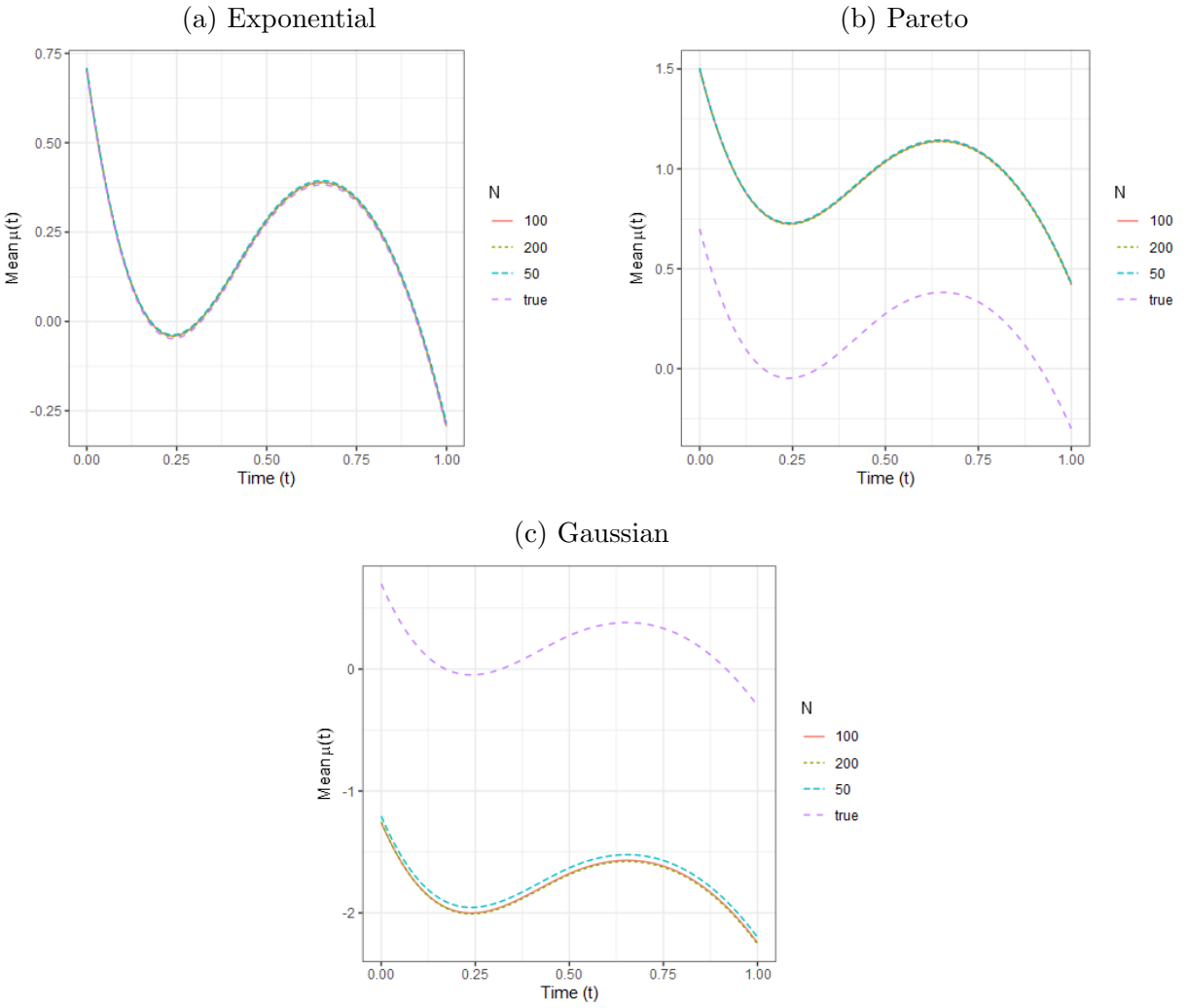
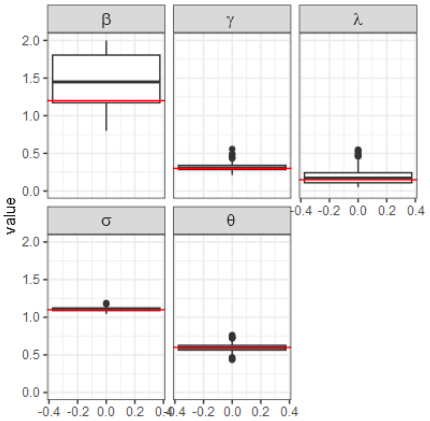
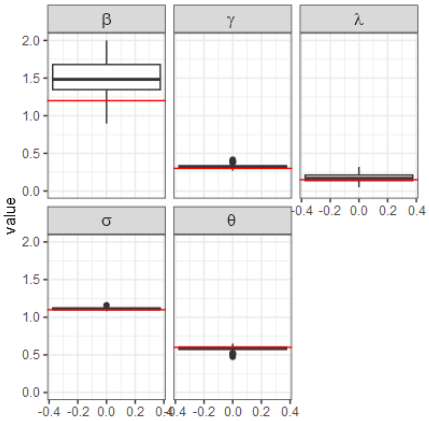


Figure 4.4: The estimated mean function versus the true mean function with different sample sizes of N for the case the tails are simulated from: **Left**: Exponential distribution, **Right**: Pareto distribution and **Bottom**: Gaussian distribution, respectively.

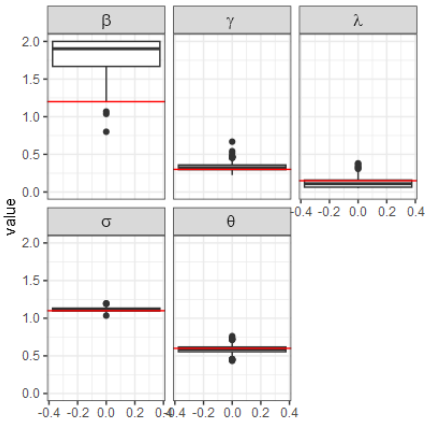
(a) Pareto, $N = 50$



(b) Pareto, $N = 200$



(c) Gaussian, $N = 50$



(d) Gaussian, $N = 200$

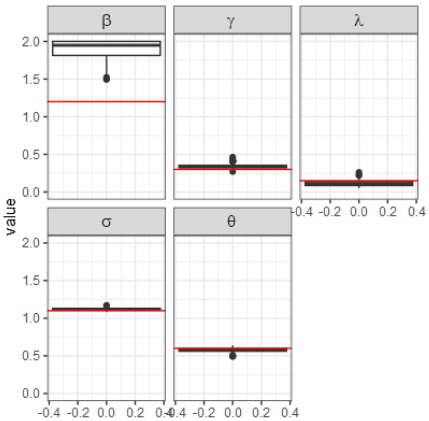


Figure 4.5: This figure presents the parameter estimates for σ, θ, γ, λ and β with different sample sizes of N in 500 simulation runs when the jump process is generating from **Top row**: Pareto distribution and **Bottom row**: Gaussian distribution. The red lines denote the true parameter values.

Chapter 5

Concluding Remarks and Future work

In this thesis, we have investigated several important research problems in functional data analysis (FDA) on the following topics: 1) Introducing a new significance test for assessing the predictability of competing real-time updated probabilistic forecasts; 2) Developing diagnostic and inferential tools to quantify the correlation between paired random processes, and 3) presenting a method capable of modelling tail asymmetric behaviour and heavy tails in functional data objects. In this chapter, we provide a summary of each of the preceding chapters along with discussions. Furthermore, we explore potential extensions and future research directions before concluding the thesis.

In Chapter 2, we developed a significance test for comparing continuously-updated probabilistic forecasts based on the L^2 norm. The proposed method was applied to data from the National Basketball Association (NBA), in which we calculated the aggregated L^2 difference between competing probabilistic forecasts over time. The methodology was adapted from the framework introduced in Lai et al. (2011) and extended to handle continuously-updated predictions as new information became available. Moreover, we devised a graphical representation that facilitates the visualization of the performance between competing methods at any given time point. To assess the finite sample properties of the test, we conducted a simulation study. The results demonstrated that the proposed methods effec-

tively distinguished competing models with different amounts of information. Lastly, we applied our methods to real NBA game data. The results revealed that, in general, the probabilistic forecasts generated by a large companies' ensemble machine learning model were well-calibrated and outperformed the naive models. However, the performance of this model was similar to that of a simple logistic regression model that we specified, which only included two covariates: pre-game relative strength and score difference during the game. Hence, this approach may still have plenty of room to improve.

This project focused on developing a statistical significance test on the independent observations with equal length. However, in reality, this might be a bit restrictive as the length or the duration of the independent observations may be different, for example, the game with overtime. Moreover, the difference in length may contain extra information. For instance, in a medical study the patients may have different durations in an intensive care unit, some sports games may have overtime past regulation time, and the sleeping times of individuals are different, in which the duration may or may not be informative. Our method does not account for these types of scenarios.

Next, in Chapter 2, we compared the predictive performance aggregated throughout the entire game. Alternatively, there may be a curve-dependent cutoff point, the outcome is determined and the predictions beyond that point may be disregarded. For instance, if the home team is leading by 50 points in a basketball game with three minutes left, it is reasonable to conclude that the home team is definitely going to win. Finding these cutoff points by disregarding the probabilistic forecasts beyond may be helpful in extending the method we have developed. The other concern is that the number of bins in the calibration plot may be subjective. Although in the appendix we compare using ten bins and find that the result seems adequate and comparable with the automated binning selecting algorithm from Dimitriadis et al. (2021), it may be helpful to develop a tailored binning selection algorithm for any given application.

In Chapter 3, we proposed a method called functional spherical autocorrelation which may be used to quantify the dependence between paired FTS objects. This is achieved based on measuring the average angle between lagged pairs of series after having been projected onto the unit sphere. We then derived the limiting distribution of estimators under the assumption that the underlying structure is similar to white noise processes. The

novelty of the proposed method was that it is able to capture the "sign" or "direction" of the correlation. Unlike the usual case, the estimator we used to centre the observed data was the geometric median instead of the mean function. This is because the standardization in our proposed method is highly nonlinear and using the mean function may bring in unnecessary bias that is associated with the skewness of the underlying FTS objects. We also showed that under mild conditions that the estimating effects of the centre of the data using the geometric median is asymptotically negligible as the sample size increases. As a by-product of using the geometric median, our method is also robust against the outlier curves, which the conventional method suffers in these scenarios. We also developed a portmanteau type of test in which we test the dependence at multiple lags simultaneously. Its asymptotic distributions were given as well.

The simulation studies showed that the proposed methods have a similar rejection rate to the existing methods, specifically fACF from Mestre et al. (2021), while accurately capturing the direction of the serial dependence. Additionally, the proposed methods exhibit good performance in finite samples. In addition, the power of the test is adequate and type one error remains at the desire level. We further showed that our method is able to access the "whiteness" of the data, and is helpful in the model selection step. Lastly, we applied our methods on two real datasets, the Spanish electricity data and the financial time series data. In the first application, we found that the autocorrelation of certain lags were negative while the seasonal effects had positive correlation with the current observation. In the second application, we demonstrate the GARCH type of model is more suitable for the series as the square of the differenced series seem auto-corrleated while the original series looked similar to a white noise sequence.

In Chapter 3, we projected the pair of stochastic processes on the unit sphere without using dimensional reduction methods such as FPCA. It has a good performance when the data generating process follows from the FAR(p) family. However, when the data generating process is from a functional moving average model with order q, denoted as FMA(q), our method cannot handle this case. On the other hand, the conventional methods given in Mestre et al. (2021) do have a version of the functional partial autocorrelation function that is able to handle the observations coming from a FMA(q) process. Thus, using our method alone may have a difficult time to distinguish between white noise and FMA pro-

cesses. Moreover, our method does not consider missing data or partially observed time series. Furthermore, we assume that the FTS objects have a strict stationarity throughout the entire domain, and we do not consider any change point or structural break. It is advisable to first perform a stationarity test proposed by Horváth et al. (2014) before applying our method. In addition, the forecasting method assumes mean stationartiy, although the Spanish electricity curves do not appear to be stationary. Thus, other types of pre-processing steps may be necessary before fitting the FTS models.

In Chapter 4, we proposed a new factor copula model to deal with functional data with spatial dependence, adapted from the factor copula model proposed by Krupskii and Genton (2017) and Krupskii et al. (2018). This model is capable to handle scenarios involving heavy tail dependence and tail asymmetry, which some basic models in the existing literature are unable to accommodate. The key idea behind the model is the assumption of the existing a common process that visits every location simultaneously in replacing the conventional random error term. This common process can be generated by various distributions, including Gaussian and exponential distributions. Specifically, when the generator of the common factor follows an exponential distribution, we are able to obtain a closed-form expression, which significantly reduces the computational cost.

Through simulation studies, we demonstrated that parameter estimation using the derived closed-form expression performs exceptionally well, even in situations where the model is misspecified and the generator is from different distributions. Notably, our model is robust against the model misspecification in capturing both the spatial dependence structure and the shape of the trend function.

Moving forward, our future investigations will focus on extending our model to cases where the common factor is generated from other distributions. Additionally, we plan to apply our methods to real datasets, including those provided by the North American CORDEX program, as well as housing price data used in other related works in studying SFD.

5.1 Future Work

In this section, we briefly go through some possible future directions based on the work presented in this thesis.

Alternative normed approaches for evaluating probabilistic forecasts: In Chapter 2, we used the L^2 norm to quantify the aggregated difference between two competing forecasters. However, the theory is general enough to allow for extensions to other norms such as L^1 or L^∞. Each norm may have its own interpretation. For instance, if we want to assess the maximum difference between two predictive methods, using the L^∞ norm may be a more reasonable choice.

Functional spherical partial autocorrelation function: As mentioned in the previous section, one drawback of our method is that we only have the ACF counterpart. When the data generating process is FMA(q), our method will suffer from the lack of distinguishability between them and the white noise processes. Thus, it is desirable to develop the PACF analogue to the FTS while retaining the direction of the dependence which was absent in method proposed in Mestre et al. (2021).

Functional spherical cross-correlation: In Chapter 3, we investigated the dependency on two time indexes of one long FTS object. An extension of this is to consider the spherical cross-correlation in the multivariate FTS setting. This is useful to investigate whether two FTS have lag effect or not. Such a scenario is common in the case of environmental and medical studies. For instance, Kowal et al. (2017) considered the local fields potential data in rats in which the prefrontal cortex and the posterior parietal cortex are correlated. Our method may help in investigating the lagged effects between the bivariate case and determine if it is practically important.

Spherical autocorrelation on panel data: In Chapter 3, we considered the case of dependency among different time windows of a univariate FTS object that is sufficiently long. On the contrary, it is common for the observed series to be short but have multiple independent curves occurring simultaneously. This type of data is known as *panel data*, in which the number of subjects can be large, but each subject may have a short time span resulting in a smaller sample size. Many real-life scenarios consist of this type of data, for

instance, women's menstrual cycles from Gold et al. (1995); Ji and Dubin (2020), where there are hundreds of study participants, each with only three cycles.

Multi-source spherical autocorrelation: In Chapter 3, we assumed that time is the only source of dependency. Alternatively, it is possible that the correlation comes from multiple sources, such as spatial-temporal data. In such cases, it is necessary to include the dependency from both sources and summarize the overall correlation. Additionally, it might be of interest to calculate the correlation from each of the sources separately to observe the effects of each factor.

Asymptotic theory and fast numerical methods for the factor copula model: In Chapter 4, we followed the work of Krupskii and Genton (2017), in which the likelihood function may be written in an alternative form when the error follows a certain distributions, enabling numerical optimization. However, we, as well as the authors of Krupskii and Genton (2017) and Krupskii et al. (2018), used `R` built-in functions, which can be slow. It is of interest to analyze the runtime and develop fast numerical implementations for performing parameter estimation of such models.

The list above is not meant to be exhaustive, as the methodologies presented in this thesis are general enough to have many more potential extensions. Overall, this thesis considers problems in FDA ranging from independent to temporal and, finally, incorporating both time and spatial dependencies. We envision that it will shed light and serve as a foundational building block for future research problems.

References

Abdi, H. and L. J. Williams (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics 2*, 433–459.

Arnone, E., A. Clemente, L. M. Sangalli, E. Lila, J. Ramsay, and L. Formaggia (2023). *fdaPDE: Physics-Informed Spatial and Functional Data Analysis*. R package version 1.1-16.

Aue, A., L. Horváth, and D. F. Pellatt (2017). Functional generalized autoregressive conditional heteroskedasticity. *Journal of Time Series Analysis 38*, 3–21.

Aue, A., D. D. Norinho, and S. Hörmann (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association 110*, 378–392.

Aue, A., G. Rice, and O. Sönmez (2020). Structural break analysis for spectrum and trace of covariance operators. *Environmetrics 31*, e2617.

Azzalini, A. and A. Capitanio (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society, Series B 65*, 367–389.

Azzalini, A. and A. D. Valle (1996). The multivariate skew-normal distribution. *Biometrika 83*, 715–726.

Bathia, N., Q. Yao, and F. Ziegelmann (2010). Identifying the finite dimensionality of curve time series. *Annals of Statistics 38*, 3352–3386.

Becker, C., R. Fried, and S. Kuhnt (2014). *Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather*. Springer, London.

Bernardo, J. M. (1979). Expected information as expected utility. *Annals of Statistics 7*, 686–690.

Bickel, J. E. (2007). Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Analysis 4*, 49–65.

Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley & Sons, New York.

Billingsley, P. (1995). *Probability and Measures* (3rd ed.). Wiley, New York.

Bloomfield, P. (2004). *Fourier Analysis of Time Series: an Introduction* (2nd ed.). Wiley, Toronto.

Boente, G. and N. L. Kudraszow (2022). Robust smoothed canonical correlation analysis for functional data. *Statistica Sinica 51*, 4942–4956.

Boente, G., D. Rodriguez, and M. Sued (2019). The spatial sign covariance operator: Asymptotic results and applications. *Journal of Multivariate Analysis 170*, 115–128.

Bosq, D. (2000). *Linear Processes in Function Spaces*. Springer, New York.

Bosq, D. (2014). Computing the best linear predictor in a hilbert space. applications to general armah processes. *Journal of Multivariate Analysis 124*, 436–450.

Box, G. and G. Jenkins (1970). *Time Series Analysis Forecasting and Control*. Holden-Day, San Francisco.

Box, G. E. (1953). Non-normality and tests on variances. *Biometrika 40*, 318–335.

Box, G. E. (1980). Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A 143*, 383–404.

Box, G. E. and D. A. Pierce (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American statistical Association 65*, 1509–1526.

Bremnes, J. B. (2004). Probabilistic forecasts of precipitation in terms of quantiles using nwp model output. *Monthly Weather Review 132*, 338–347.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78*, 1–3.

Brillinger, D. R. (2001). *Time Series: Data Analysis and Theory*. Society for Industrial and Applied Mathematics, Philadelphia.

Brockwell, P. J. and R. A. Davis (1991). *Time Series: Theory and Methods*. Springer, New York.

Brown, B. M. (1971). Martingale central limit theorems. *Annals of Mathematical Statistics 42*, 59–66.

Brown, L. D., T. T. Cai, and A. DasGupta (2001). Interval estimation for a binomial proportion. *Statistical Science 16*, 101–117.

Cao, J., K. Soiaporn, R. J. Carroll, and D. Ruppert (2019). Modeling and prediction of multiple correlated functional outcomes. *Journal of Agricultural, Biological and Environmental Statistics 24*, 112–129.

Cardot, H., P. Cénac, and P. Zitt (2013). Efficient and fast estimation of the geometric median in hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli 19*, 18–43.

Chen, T. and Q. Fan (2018). A functional data approach to model score difference process in professional basketball games. *Journal of Applied Statistics 45*, 112–127.

Clark, T. E. and M. W. McCracken (2015). Nested forecast model comparisons: A new approach to testing equal accuracy. *Journal of Econometrics 186*, 160–177.

Commandeur, J. J. and S. J. Koopman (2007). *An Introduction to State Space Time Series Analysis*. Oxford University Press, New York.

Council, N. R. (2006). *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*. The National Academies Press, Washington.

Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons, New York.

Cryer, J. D. and K.-S. Chan (2008). *Time Series Analysis: with Applications in R* (2 ed.). Springer, New York.

Cuevas, A., M. Febrero, and R. Fraiman (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics 22*, 481–496.

Czado, C. and T. Nagler (2022). Vine copula based modeling. *Annual Review of Statistics and Its Application 9*, 453–477.

Dawid, A. P. (1986). *Probability Forecasting*, Volume 8, pp. 210–218. New York: John Wiley & Sons.

Diebold, F. X. and R. S. Mariano (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics 20*, 134–144.

Dimitriadis, T., T. Gneiting, and A. I. Jordan (2021). Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences 118*, e2016191118.

Dubin, J. A. and H.-G. Müller (2005). Dynamical correlation for multivariate longitudinal data. *Journal of the American Statistical Association 100*, 872–881.

Duchon, C. and R. Hale (2012). *Time Series Analysis in Meteorology and Climatology: an Introduction*. Wiley, Hoboken.

Durbin, J. and S. J. Koopman (2012). *Time Series Analysis by State Space Methods* (2nd ed.). Oxford University Press, Croydon.

Elo, A. E. (1978). *The Rating of Chessplayers, Past and Present*. Arco Pubulishing, New York.

Embrechts, P., A. McNeil, and D. Straumann (2001). Correlation and dependence in risk management: properties and pitfalls. *Risk Management: Value at Risk and Beyond 1*, 176–223.

Erikson, R. S. and C. Wlezien (2012). Markets vs. polls as election predictors: An historical assessment. *Electoral Studies 31*, 532–539.

ESPN (2020). Espn, espn internet ventures, national basketball association teams, scores, stats, news, standings, rumors. www.espn.com/nba/. Accessed on 2020-05-01.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, New York.

Fraiman, R. and G. Muniz (2001). Trimmed means for functional data. *Test 10*, 419–440.

Francq, C. and J.-M. Zakoian (2019). *GARCH Models: Structure, Statistical Inference and Financial Applications* (2nd ed.). Wiley, New Jersey.

French, J. P. and P. S. Kokoszka (2021). A sandwich smoother for spatio-temporal functional data. *Spatial Statistics 42*, 100413.

Fritz, H., P. Filzmoser, and C. Croux (2012). A comparison of algorithms for the multivariate l 1-median. *Computational Statistics 27*, 393–410.

Gabel, A. and S. Redner (2012). Random walk picture of basketball scoring. *Journal of Quantitative Analysis in Sports 8*.

Gabrys, R. and P. Kokoszka (2007). Portmanteau test of independence for functional observations. *Journal of the American Statistical Association 102*, 1338–1348.

Genest, C. and A.-C. Favre (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering 12*, 347–368.

Gervini, D. (2008). Robust functional estimation using the median and spherical principal components. *Biometrika 95*, 587–600.

Gijbels, I. and S. Nagy (2017). On a general definition of depth for functional data. *Statistical Science 32*, 630–639.

Giraldo, R., P. Delicado, and J. Mateu (2020). *geofd: Spatial Prediction for Function Value Data*. R package version 2.0.

Gnanadesikan, R. (2011). *Methods for Statistical Data Analysis of Multivariate Observations* (2nd ed.). John Wiley & Sons, New York.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association 106*, 746–762.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society, Series B 69*, 243–268.

Gneiting, T. and M. Katzfuss (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application 1*, 125–151.

Gneiting, T., L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test 17*, 211.

Gold, E. B., B. Eskenazi, B. L. Lasley, S. J. Samuels, M. O'Neill Rasor, J. W. Overstreet, and M. B. Schenker (1995). Epidemiologic methods for prospective assessment of menstrual cycle and reproductive characteristics in female semiconductor workers. *American Journal of Industrial Medicine 28*, 783–797.

Gomberg, J. (2015). Earthquake forewarning in the cascadia region. *U.S. Geological Survey Open-File Report*.

González, J. P., A. M. S. M. San Roque, and E. A. Perez (2017). Forecasting functional time series with a new hilbertian armax model: Application to electricity price forecasting. *IEEE Transactions on Power Systems 33*, 545–556.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B 14*, 107–114.

Górecki, J., M. Hofert, and M. Holena (2020). Hierarchical archimedean copulas for matlab and octave: The hacopula toolbox. *Journal of Statistical Software 93*, 1–36.

Hall, P. and C. Vial (2006). Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society, Series B 68*, 689–705.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press, New Jersey.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting 15*, 559–570.

Hilden, J. and T. A. Gerds (2014). A note on the evaluation of novel biomarkers: do not rely on integrated discrimination improvement and net reclassification index. *Statistics in Medicine 33*, 3405–3414.

Hofert, M., I. Kojadinovic, M. Mächler, and J. Yan (2018). *Elements of Copula Modeling with R*. Springer, New York.

Hofert, M., I. Kojadinovic, M. Maechler, J. Yan, J. G. Nešlehová, and R. Morger (2023). *copula: Multivariate Dependence with Copulas*. R package version 1.1-2.

Hörmann, S., L. Horváth, and R. Reeder (2013). A functional version of the arch model. *Econometric Theory 29*, 267–288.

Hörmann, S. and Ł. Kidziński (2017). *freqdom.fda: Functional Time Series: Dynamic Functional Principal Components*. R package version 0.9.1.

Hörmann, S., Ł. Kidziński, and M. Hallin (2015). Dynamic functional principal components. *Journal of the Royal Statistical Society, Series B 77*, 319–348.

Hörmann, S. and P. Kokoszka (2010). Weakly dependent functional data. *Annals of Statistics 38*, 1845–1884.

Hörmann, S. and P. Kokoszka (2012). Functional time series. In T. S. Rao, S. S. Rao, and C. R. Rao (Eds.), *Time Series Analysis: Methods and Applications*, pp. 157–186. Elsevier, New York.

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*. Springer, New York.

Horváth, L., P. Kokoszka, and G. Rice (2014). Testing stationarity of functional time series. *Journal of Econometrics 179*, 66–82.

Horváth, L., Z. Liu, G. Rice, and S. Wang (2020). A functional time series analysis of forward curves derived from commodity futures. *International Journal of Forecasting 36*, 646–665.

Hristopulos, D. T. (2020). *Random Fields for Spatial Data Modeling*. Springer, New York.

Hsing, T. and R. Eubank (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley, Chichester, UK.

Huber, P. J. (1972). The 1972 wald lecture robust statistics: A review. *Annals of Mathematical Statistics 43*, 1041–1067.

Huber, P. J. (2011). Robust statistics. In *International Encyclopedia of Statistical Science*, pp. 1248–1251. Springer, New York.

Hull, J. C. and A. D. White (2004). Valuation of a cdo and an n-th to default cds without monte carlo simulation. *The Journal of Derivatives 12*, 8–23.

Hyndman, R. J. and Y. Khandakar (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software 27*, 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*, 29–45.

Hyndman, R. J. and H. L. Shang (2022). *ftsa: An R package for analyzing functional time series*. R package version 6.1.

Hyndman, R. J. and M. S. Ullah (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis 51*, 4942–4956.

Ignaccolo, R., J. Mateu, and R. Giraldo (2014). Kriging with external drift for functional data for air quality monitoring. *Stochastic Environmental Research and Risk Assessment 28*, 1171–1186.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika 48*, 419–426.

James, G. M., T. J. Hastie, and C. A. Sugar (2000). Principal component models for sparse functional data. *Biometrika 87*, 587–602.

Ji, K. and J. A. Dubin (2020). A semiparametric stochastic mixed effects model for bivariate cyclic longitudinal data. *Canadian Journal of Statistics 48*, 471–498.

Ji, Q., D. Zhang, and Y. Zhao (2022). Intra-day co-movements of crude oil futures: China and the international benchmarks. *Annals of Operations Research 313*, 77–103.

Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. CRC press, London.

Joe, H. (2014). *Dependence Modeling with Copulas*. CRC Press, Boca Raton.

Jolliffe, I. T. (1986). *Principal Component Analysis*. Springer, New York.

Jolliffe, I. T. and J. Cadima (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society, Series A 374*, 20150202.

Karhunen, K. (1946). Zur spektraltheorie stochastischer prozesse. *Annales Academiae Scientiarum Fennicae Mathematica 34*.

Kemperman, J. H. B. (1987). The median of a finite measure on a banach space. In Y. Dodge (Ed.), *Statistical Data Analysis Based on the L1-norm and Related Methods*, pp. 217–230. North-Holland, Amsterdam.

Kim, M., P. Kokoszka, and G. Rice (2023). White noise testing for functional time series. *Statistic Surveys 17*, 119–168.

Kim, M. and D. Petoukhov (2022). *wwntests: Hypothesis Tests for Functional Time Series*. R package version 1.0.2.

Kokoszka, P., H. Miao, S. Stoev, and B. Zheng (2019). Risk analysis of cumulative intraday return curves. *Journal of Time Series Econometrics 11*.

Kokoszka, P. and M. Reimherr (2017). *Introduction to Functional Data Analysis*. CRC Press, Boca Raton.

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.

Koner, S. and A.-M. Staicu (2023). Second-generation functional data. *Annual Review of Statistics and its Application 10*, 547–572.

Kowal, D. R., D. S. Matteson, and D. Ruppert (2017). A bayesian multivariate functional dynamic linear model. *Journal of the American Statistical Association 112*, 733–744.

Krupskii, P. (2014). *Structured Factor Copulas and Tail Inference*. Ph. D. thesis, University of British Columbia.

Krupskii, P. and M. G. Genton (2017). Factor copula models for data with spatio-temporal dependence. *Spatial Statistics 22*, 180–195.

Krupskii, P. and M. G. Genton (2019). A copula model for non-Gaussian multivariate spatial data. *Journal of Multivariate Analysis 169*, 264–277.

Krupskii, P., R. Huser, and M. G. Genton (2018). Factor copula models for replicated spatial data. *Journal of the American Statistical Association 113*, 467–479.

Krupskii, P. and H. Joe (2013). Factor copula models for multivariate data. *Journal of Multivariate Analysis 120*, 85–101.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics 22*, 79–86.

Lai, T. L., S. T. Gross, and D. B. Shen (2011). Evaluating probability forecasts. *Annals of Statistics 39*, 2356–2382.

Lane, B. (2002). *Statistical Methods in Hydrology*. Iowa State Press, Iowa.

Leng, X. and H.-G. Müller (2006). Classification using functional data analysis for temporal gene expression data. *Bioinformatics 22*, 68–76.

Leurgans, S. E., R. A. Moyeed, and B. W. Silverman (1993). Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society, Series B 55*, 725–740.

Li, H., J. Staudenmayer, and R. J. Carroll (2014). Hierarchical functional data with mixed continuous and binary measurements. *Biometrics 70*, 802–811.

Li, W. and Q.-M. Shao (2001). Gaussian processes: Inequalities, small ball probabilities and applications. In D. N. Shanbhag and C. R. Rao (Eds.), *Stochastic Processes: Theory and Methods*, pp. 533–598. Elsevier, London.

Li, Y. and T. Hsing (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Annals of Statistics 38*, 3321–3351.

Liu, C., S. Ray, and G. Hooker (2017). Functional principal component analysis of spatially correlated data. *Statistics and Computing 27*, 1639–1654.

Ljung, G. M. and G. E. Box (1978). On a measure of lack of fit in time series models. *Biometrika 65*, 297–303.

Loève, M. (1946). Fonctions aléatoires à décomposition orthogonale exponentielle. *La Revue Scientifique 84*, 159–162.

Magnus, J. R. (1986). The exact moments of a ratio of quadratic forms in normal variables. *Annales d'Economie et de Statistique 4*, 95–109.

Mardia, K. V., J. T. Kent, and J. M. Bibby (1979). *Multivariate Analysis*. Academic Press, San Diego.

Maronna, R. A., R. D. Martin, V. J. Yohai, and M. Salibián-Barrera (2019). *Robust Statistics: Theory and Methods (with R)* (2nd ed.). Wiley, New York.

Matheson, J. E. and R. L. Winkler (1976). Scoring rules for continuous probability distributions. *Management Science 22*, 1087–1096.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). CRC Press, New York.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behaviour. In P. Zarembka (Ed.), *Frontiers in Econometrics*, pp. 105–142. New York: Academic Press.

McNeil, A. J., R. Frey, and P. Embrechts (2015). *Quantitative Risk Management: Concepts, Techniques and Tools-revised edition*. Princeton University Press, New Jersey.

Mercer, J. (1909). Functions of positive and negative type, and their connection the theory of integral equations. *Philosophicol Trinsdictions of the Rogyal Society of London, Series A 209*, 4–446.

Mestre, G., J. P. González, G. Rice, and E. A. Antonio Muñoz San Muñoz, Perez (2020). *fdaACF: Autocorrelation Function for Functional Time Series*. R package version 1.0.0.

Mestre, G., J. Portela, G. Rice, A. M. 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.

Murphy, A. H. (1972). Scalar and vector partitions of the probability score: Part i. two-state situation. *Journal of Applied Meteorology 11*, 273–282.

Murphy, A. H. (1998). The early history of probability forecasts: Some extensions and clarifications. *Weather and Forecasting 13*, 5–15.

Murphy, A. H. and R. L. Winkler (1987). A general framework for forecast verification. *Monthly Weather Review 115*, 1330–1338.

Murphy, A. H. and R. L. Winkler (1992). Diagnostic verification of probability forecasts. *Interonal Journal of Forecasting 7*, 435–455.

Nazarov, A. (2009). Exact l2-small ball asymptotics of gaussian processes and the spectrum of boundary-value problems. *Journal of Theoretical Probability 22*, 640–665.

Nelsen, R. B. (1999). *An Introduction to Copulas*. Springer, New York.

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica 59*, 1161–1167.

Oh, D. H. and A. J. Patton (2017). Modeling dependence in high dimensions with factor copulas. *Journal of Business & Economic Statistics 35*, 139–154.

Panaretos, V. M. and S. Tavakoli (2013). Fourier analysis of stationary time series in function space. *Annals of Statistics 41*, 568–603.

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika 23*, 114–133.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine 2*, 559–572.

Ramsay, J., G. Hooker, and S. Grave (2009). *Functional Data Analysis with R and MATLAB*. Springer, New York.

Ramsay, J. O. and C. Dalzell (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society, Series B 53*, 539–561.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis* (2 ed.). Springer, New York.

Ramsay, J. O. and B. W. Silverman (2007). *Applied Functional Data Analysis: Methods and Case Studies*. Springer, New York.

Ranjan, R. and T. Gneiting (2010). Combining probability forecasts. *Journal of the Royal Statistical Society, Series B 72*, 71–91.

Rice, G., T. Wirjanto, and Y. Zhao (2020). Tests for conditional heteroscedasticity of functional data. *Journal of Time Series Analysis 41*, 733–758.

Rowe, T. and S. Beard (2018). Probabilities, methodologies and the evidence base in existential risk assessments. *Centre for the Study of Existential Risk, University of Cambridge*.

Salas, J. D., J. W. Delleur, V. Yevjevich, and W. L. Lane (1980). *Applied Modeling of Hydrologic Time Series*. Water Resources Publications, Chelsea.

Salibian-Barrera, M., V. Yohai, R. Maronna, D. Martin, G. Brownson, K. Konis, K. Konis, C. Croux, G. Haesbroeck, M. Maechler, and M. Koller (2023). *RobStatTM: Robust Statistics: Theory and Methods*. R package version 1.0.7.

Sang, P., L. Wang, and J. Cao (2019). Weighted empirical likelihood inference for dynamical correlations. *Computational Statistics & Data Analysis 131*, 194–206.

Seillier-Moiseiwitsch, F. and A. P. Dawid (1993). On testing the validity of sequential probability forecasts. *Journal of the American Statistical Association 88*, 355–359.

Serfling, R. and U. Wijesuriya (2017). Depth-based nonparametric description of functional data, with emphasis on use of spatial depth. *Computational Statistics & Data Analysis 105*, 24–45.

Shahid, R., S. Bertazzon, M. L. Knudtson, and W. A. Ghali (2009). Comparison of distance measures in spatial analytical modeling for health service planning. *BMC health services research 9*, 1–14.

Shang, H. (2014). A survey of functional principal component analysis. *AStA Advances in Statistical Analysis 98*, 121–142.

Shang, H. L. (2013). ftsa: An R package for analyzing functional time series. *The R Journal 5*, 64–72.

Shang, H. L. (2017). Functional time series forecasting with dynamic updating: An application to intraday particulate matter concentration. *Econometrics and Statistics 1*, 184–200.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika 68*, 45–54.

Shumway, R. H. and D. S. Stoffer (2017). *Time Series Analysis and Its Applications: with R Examples*. Springer, New York.

Silver, N. (2014). Introducing nfl elo ratings. https://fivethirtyeight.com/features/introducing-nfl-elo-ratings/. Accessed on 2020-03-18.

Silver, N. (2020). 2020 election forecast. https://projects.fivethirtyeight.com/2020-election-forecast/. Accessed: 2020-09-30.

Silver, N., J. Boice, and N. Paine (2019). How our nba predictions work. https://fivethirtyeight.com/methodology/how-our-nba-predictions-work/. Accessed on 2020-03-18.

Silver, N. and R. Fischer-Baum (2015). How we calculate nba elo ratings. https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/. Accessed on 2020-03-11.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris 8*, 229–231.

Small, C. G. (1990). A survey of multidimensional medians. *International Statistical Review 58*, 263–277.

Sørensen, H., J. Goldsmith, and L. M. Sangalli (2013). An introduction with medical applications to functional data analysis. *Statistics in Medicine 32*, 5222–5240.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology 15*, 72–101.

Spiegelhalter, D. J. (1986). Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine 5*, 421–433.

Staicu, A.-M., C. M. Crainiceanu, D. S. Reich, and D. Ruppert (2012). Modeling functional data with spatially heterogeneous shape characteristics. *Biometrics 68*, 331–343.

Stern, H. S. (1994). A brownian motion model for the progress of sports scores. *Journal of the American Statistical Association 89*, 1128–1134.

Team, R. C. (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Tsay, R. S. (2005). *Analysis of Financial Time Series* (3rd ed.). Wiley, New Jersey.

Ullah, S. and C. F. Finch (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology 13*, 43.

Wang, J.-L., J.-M. Chiou, and H.-G. Müller (2016). Functional data analysis. *Annual Review of Statistics and its Application 3*, 257–295.

Whittle, P. (1951). *Hypothesis Testing in Time Series Analysis*. Uppsala: Almqvist & Wiksells.

Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: with Engineering Applications*. MIT Press, Cambridge.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association 22*, 209–212.

Wold, S., K. Esbensen, and P. Geladi (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems 2*, 37–52.

Yan, J. (2007). Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software 21*, 1–21.

Yao, F. (2007). Functional principal component analysis for longitudinal and survival data. *Statistica Sinica 17*, 965–983.

Yeh, C.-K., G. Rice, and J. A. Dubin (2022). Evaluating real-time probabilistic forecasts with application to national basketball association outcome prediction. *The American Statistician 76*, 214–223.

Yeh, C.-K., G. Rice, and J. A. Dubin (2023). Functional spherical autocorrelation: A robust estimate of the autocorrelation of a functional time series. *Electronic Journal of Statistics 17*, 650–687.

Zhang, H. and Y. Li (2022). Unified principal component analysis for sparse and dense functional data under spatial dependency. *Journal of Business & Economic Statistics 40*, 1523–1537.

Zhang, X. and J.-L. Wang (2016). From sparse to dense functional data and beyond. *Annals of Statistics 44*, 2281–2321.

Zhou, L., J. Z. Huang, and R. J. Carroll (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika 95*, 601–619.

Zhou, L., J. Z. Huang, J. G. Martinez, A. Maity, V. Baladandayuthapani, and R. J. Carroll (2010). Reduced rank mixed effects models for spatially correlated hierarchical functional data. *Journal of the American Statistical Association 105*, 390–400.

APPENDICES

Appendix A

Technical Details of Chapter 2

A.1 Definition of $p_i(t)$ and Proof of Theorem 2.4.1

In order to formally define $p_i(t)$, we first define two collections of sigma-algebras: $\mathcal{F}_i$, which is the information available up to and including the $i'th$ game, and $\mathcal{F}_{i-1,t}$ which is the information in all games up to and including the $i-1$'st game, along with the information up to time t in the $i'th$ game. Clearly then for all $t \in [0,1]$, $\mathcal{F}_{i-1,t} \subset \mathcal{F}_i \subset \mathcal{F}_{i,t}$. Let Y_i denote the zero-one variables encoding wins for the home team. We define $p_i(t) = \mathbb{E}[Y_i \mid \mathcal{F}_{i-1,t}]$ to be the probability that the home team wins in the i'th game given the information up to time t in that game (and previous games), which we are aiming to forecast.

Suppose that we have two methods $\hat{p}_i^A(t)$, $\hat{p}_i^B(t)$ for forecasting the in-game probabilities of the home team winning in game i at time t.

Assumption A.1.1 *$\hat{p}_i^A(t)$, $\hat{p}_i^B(t)$ are both measurable with respect to $\mathcal{F}_{i-1,t}$. In other words, only the information from all previous games as well as the current game up to time t are used to produce the forecasts $\hat{p}_i^A(t)$, $\hat{p}_i^B(t)$.*

Let $L^2[0,1]$ denote the space of square integrable functions defined on $[0,1]$, which has canonical inner-product defined for $f, g \in L^2[0,1]$ as $\langle f, g \rangle = \int_0^1 f(t)g(t)dt$. We assume

here that $L(a,b) = (a-b)^2$ is the Brier loss, but the following results generalize to any loss functions with a linear equivalent (Lai et al., 2011). Define

$$\begin{aligned} Z_N(t) &= \sqrt{N}[\hat{\Delta}_N(t) - \Delta_N(t)] \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left[L(Y_i, \hat{p}_i^A(t)) - L(Y_i, \hat{p}_i^B(t)) - [L(p_i(t), \hat{p}_i^A(t)) - L(p_i(t), \hat{p}_i^B(t))] \right]. \end{aligned}$$

Note that under the null hypothesis $H_0 : \|\Delta_N\|^2 = 0$ of equivalent forecasting skill between the models A and B, $Z_N(t) = \sqrt{N}\hat{\Delta}_N(t)$ in L^2 sense. In order to formally state and prove Theorem 2.4.1, we also make the following assumptions. A sequence of covariance kernels that arises in calculating the distribution of $Z_N(t)$ is

$$\begin{aligned} C_N(t,s) = \frac{4}{N} \sum_{i=1}^{N} \big(&\mathbb{E}[(\hat{p}_i^A(s) - \hat{p}_i^B(s))(\hat{p}_i^A(t) - \hat{p}_i^B(t)) p_i(s)(1-p_i(s)) \mid \mathcal{F}_{i-1}] \mathbb{1}_{s \geq t} \\ &+ \mathbb{E}[(\hat{p}_i^A(s) - \hat{p}_i^B(s))(\hat{p}_i^A(t) - \hat{p}_i^B(t)) p_i(t)(1-p_i(t)) \mid \mathcal{F}_{i-1}] \mathbb{1}_{t > s} \big), \end{aligned} \tag{A.1}$$

where $\mathbb{1}$ is the indicator function. As we will show, C_N is essentially the average covariance function of the (centred) difference between the forecasts $\hat{p}_i^A(t) - \hat{p}_i^B(t)$. This kernel is evidently symmetric for each N, and we show below that it is also positive (semi-)definite, which is to say that for all functions $f \in L^2[0,1]$,

$$\int_0^1 \int_0^1 C_N(t,s) f(t) f(s) dt ds \geq 0.$$

Assumption A.1.2 *There exists a symmetric, positive definite kernel C satisfying $\int_0^1 C(t,t)dt < \infty$ such that*

$$\int_0^1 \int_0^1 [C_N(t,s) - C(t,s)]^2 dt ds = o_P(1). \tag{A.2}$$

Assumption A.1.2 basically entails that there is a degree of "stationarity" or "ergodicity" between subsequent forecasts. This would be implied if, for example, subsequent forecasts in each game were independent of each other and exhibit an approximately stable

distribution, although independence could be relaxed a great deal here with (A.2) still holding. This condition is analogous to condition (a) in Theorem 1 of Seillier-Moiseiwitsch and Dawid (1993), the validity of which in probabilistic forecast evaluation is discussed there, and this condition is also central to the results of Lai et al. (2011).

The kernel C defined in Assumption A.1.2 may be used to define a symmetric, positive definite linear operator $\mathcal{C} : L^2[0,1] \mapsto L^2[0,1]$ via

$$\mathcal{C}(f)(t) = \int_0^1 C(t,s)f(s)ds, \tag{A.3}$$

which by Mercer's Theorem (Bosq, 2000) must have associated to it a decreasing sequence of eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$ and corresponding orthonormal eigenfunctions $\phi_1, \phi_2, \cdots \in L^2[0,1]$ satisfying $\mathcal{C}(\phi_i) = \lambda_i \phi_i$. A similar collection of eigenvalues and $\lambda_{1,N} \geq \lambda_{2,N} \geq \cdots$ may be defined the operator $\mathcal{C}_N$ obtained by replacing C by C_N in the definition (A.3).

Assumption A.1.3 *The sequence of variables Z_N are uniformly tight in $L^2[0,1]$ (see pg. 46 of Bosq, 2000).*

Uniform tightness is used in order to extend asymptotic Gaussianity of the finite dimensional distributions of the process Z_N to the entire process. Intuitively assuming tightness in this context assumes a degree of continuity or "smoothness" to the forecasts $\hat{p}_i^A(t)$ and $\hat{p}_i^B(t)$ as a function of t. A sufficient condition to imply Assumption A.1.3 is

$$\sup_{N \geq 1} \sum_{i=1}^{\infty} \lambda_{i,N} < \infty;$$

see the proof on page 52 of Bosq (2000).

Restatement of Theorem 2.4.1 Under H_0 and Assumptions A.1.1-A.1.3,

$$\|Z_N\|^2 \xrightarrow{D} \sum_{i=1}^{\infty} \lambda_i \chi_i^2(1),$$

where the $\chi_i^2(1)$, $i = 1, 2, \ldots$ are i.i.d. χ^2 random variables with one degree of freedom and the constants $\{\lambda_i,\ i \geq 1\}$ are defined in (A.3). These constants may be asymptotically conservatively estimated by the eigenvalues of

$$\hat{C}_{cons}(t,s) = \frac{1}{N} \sum_{i=1}^{N} [\hat{p}_i^A(t) - \hat{p}_i^B(t)][\hat{p}_i^A(s) - \hat{p}_i^B(s)].$$

Proof A.1.1 *For all test functions $v \in L^2[0,1]$, and using the definition of the Brier score, we have that*

$$\langle Z_N, v\rangle/2 = \sqrt{N}[\hat{\Delta}_N(t) - \Delta_N(t)]/2 = \frac{1}{2\sqrt{N}}\sum_{i=1}^{N}\langle L(Y_i,\hat{p}_i^A) - L(Y_i,\hat{p}_i^B) - [L(p_i,\hat{p}_i^A) - L(p_i,\hat{p}_i^B)], v\rangle$$
$$= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\langle(\hat{p}_i^A - \hat{p}_i^B)(p_i - Y_i), v\rangle =: \frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_{i,v}.$$

$X_{i,v}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_i$. To see this, notice by the tower property of conditional expectation and Fubini's theorem that

$$\mathbb{E}[X_{i,v} \mid \mathcal{F}_{i-1}] = \mathbb{E}[\mathbb{E}[X_{i,v}|\mathcal{F}_{i-1,t}]\mathcal{F}_{i-1}] = \mathbb{E}[\langle\mathbb{E}[(\hat{p}_i^A - \hat{p}_i^B)(p_i - Y_i) \mid \mathcal{F}_{i-1,t}], v\rangle \mid \mathcal{F}_{i-1}] = 0,$$

since $\mathbb{E}[(\hat{p}_i^A(t) - \hat{p}_i^B(t))(p_i(t) - Y_i) \mid \mathcal{F}_{i-1,t}] = (\hat{p}_i^A(t) - \hat{p}_i^B(t))\mathbb{E}[(p_i(t) - Y_i) \mid \mathcal{F}_{i-1,t}] = 0$, a.s. for all t, using Assumption A.1.1. Further,

$$\mathbb{E}[X_{i,v}^2 \mid \mathcal{F}_{i-1}] = \int_0^1\int_0^1 \mathbb{E}[(\hat{p}_i^A(t) - \hat{p}_i^B(t))(p_i(t) - Y_i)(\hat{p}_i^A(s) - \hat{p}_i^B(s))(p_i(s) - Y_i)v(t)v(s) \mid \mathcal{F}_{i-1}]dtds \tag{A.4}$$

Suppose $s \geq t$, then using the tower property

$$\mathbb{E}[(\hat{p}_i^A(s) - \hat{p}_i^B(s))(p_i(s) - Y_i)(\hat{p}_i^A(t) - \hat{p}_i^B(t))(p_i(t) - Y_i)v(t)v(s) \mid \mathcal{F}_{i-1}]$$
$$= \mathbb{E}[\mathbb{E}[(\hat{p}_i^A(s) - \hat{p}_i^B(s))(p_i(s) - Y_i)(\hat{p}_i^A(t) - \hat{p}_i^B(t))(p_i(t) - Y_i)v(t)v(s) \mid \mathcal{F}_{i-1,s}] \mid \mathcal{F}_{i-1}]$$
$$= \mathbb{E}[(\hat{p}_i^A(s) - \hat{p}_i^B(s))(\hat{p}_i^A(t) - \hat{p}_i^B(t))v(t)v(s)\mathbb{E}[(p_i(s) - Y_i)(p_i(t) - Y_i) \mid \mathcal{F}_{i-1,s}] \mid \mathcal{F}_{i-1}].$$

Note that $Y_i \mid \mathcal{F}_{i-1,s}$ is a Bernoulli($p_i(s)$) random variable, thus it follows that $\mathbb{E}[(p_i(s) - Y_i)(p_i(t) - Y_i) \mid \mathcal{F}_{i-1,s}] = p_i(s)(1 - p_i(s))$. From this we obtain that for $s \geq t$,

$$\mathbb{E}[(\hat{p}_i^A(s) - \hat{p}_i^B(s))(p_i(s) - Y_i)(\hat{p}_i^A(t) - \hat{p}_i^B(t))(p_i(t) - Y_i)v(t)v(s) \mid \mathcal{F}_{i-1}]$$
$$= \mathbb{E}[(\hat{p}_i^A(s) - \hat{p}_i^B(s))(\hat{p}_i^A(t) - \hat{p}_i^B(t))p_i(s)(1 - p_i(s)) \mid \mathcal{F}_{i-1}]v(t)v(s).$$

Plugging into (A.4) and applying the same calculation when $t > s$ gives that

$$\mathbb{E}[X_{i,v}^2 \mid \mathcal{F}_{i-1}] = \int_0^1\int_0^1 \{\mathbb{E}[(\hat{p}_i^A(s) - \hat{p}_i^B(s))(\hat{p}_i^A(t) - \hat{p}_i^B(t))p_i(s)(1 - p_i(s)) \mid \mathcal{F}_{i-1}]\mathbb{1}_{s\geq t}$$

$$+ \mathbb{E}[(\hat{p}_i^A(s) - \hat{p}_i^B(s))(\hat{p}_i^A(t) - \hat{p}_i^B(t))p_i(t)(1-p_i(t)) \mid \mathcal{F}_{i-1}]\mathbb{1}_{t>s}\}v(t)v(s)dtds.$$

It follows that

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[X_{i,v}^2 \mid \mathcal{F}_{i-1}] = \int_0^1\int_0^1 \frac{1}{4}C_N(t,s)v(t)v(s)dtds.$$

Hence, using Assumption A.1.2,

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[X_{i,v}^2 \mid \mathcal{F}_{i-1}] \xrightarrow{P} \int_0^1\int_0^1 \frac{1}{4}C(t,s)v(t)v(s)dtds = \langle \mathcal{C}(v), v\rangle =: \sigma_v^2. \tag{A.5}$$

We consider the cases of $\sigma_v^2 = 0$ and $\sigma_v^2 > 0$ separately. In the case where $\sigma_v^2 = 0$, we first note that, using the Cauchy-Schwarz inequality, $\mathbb{E}[X_{i,v}^2 \mid \mathcal{F}_{i-1}] \leq \|v\|^2\mathbb{E}[\|(\hat{p}_i^A(\cdot) - \hat{p}_i^B(\cdot))(p_i(\cdot)-Y_i)\|^2 \mid \mathcal{F}_{i-1}]$, which is uniformly bounded, and hence $(1/N)\sum_{i=1}^N \mathbb{E}[X_{i,v}^2 \mid \mathcal{F}_{i-1}]$ is uniformly integrable. It follows then from (A.5) that $\sigma_v^2 = 0$ implies

$$\mathbb{V}\mathrm{ar}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}X_{i,v}\right) = \mathbb{E}\left\{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[X_{i,v}^2 \mid \mathcal{F}_{i-1}]\right\} \to 0.$$

When $\sigma_v^2 > 0$, let $S_{N,v}^2 = \mathbb{V}\mathrm{ar}(\sum_{i=1}^N X_{i,v}) = \sum_{i=1}^N \mathbb{E}[X_{i,v}^2]$. Using (A.5), it follows that there exists a constant $D > 0$ so that for all N sufficiently large, $S_{N,v}^2 \geq DN$. Thus, for all $\epsilon > 0$ and N sufficiently large, we have due to the uniform boundedness of $X_{i,v}^2$ that

$$\frac{1}{S_{N,v}^2}\sum_{i=1}^{N}\mathbb{E}X_{i,v}^2\mathbb{1}_{\{|X_{i,v}|\geq\epsilon S_{N,v}\}} = 0.$$

Hence both the Lindeberg condition and the asymptotic constancy of the conditional variance condition of the Martingale central limit theorem, c.f. Theorem 1 of Brown (1971), hold, giving that $\langle Z_N, v\rangle/2 \xrightarrow{D} \mathcal{N}(0,\sigma_v^2)$. This shows that the finite dimensional distributions of Z_N asymptotically coincide with those of a Gaussian process with covariance kernel C. This along with Assumption A.1.3 imply that Z_N converges weakly in $L^2[0,1]$ to a Gaussian process with covariance kernel C (see Prohorov's theorem, Theorem 2.6 in Bosq 2000). The form of the limiting distribution of $\|Z_N\|^2$ follows from the continuous mapping theorem and the Karhunen-Loéve representation theorem.

The fact that the eigenvalues of the kernel $\hat{C}_{cons}$ are uniformly larger than those of C_N follows since $\hat{C}_{cons}$ is obtained by replacing $p_i(t)(1-p_i(t))$ and $p_i(s)(1-p_i(s))$ with the upper bound $1/4$.

A.2 Selection of the Number of Bins using CORP Binning Method

Dimitriadis et al. (2021) develops a criterion for estimating the optimal number of bins in constructing reliability diagrams, which we also implemented and compared to. The optimal number of bins at each time point estimated using the method of Dimitriadis et al. (2021) and for the forecasting methods ScDnoInt, ESPN, and PgRSSCD are displayed in Figure A.1, from which we see that their method tends to recommend more than 10 bins at each time point, and tends to recommend closer to 20 bins in the middle of the game when the forecast probabilities tend to be more uniformly distributed. The corresponding real-time reliability diagrams for the methods ScDnoInt, ESPN, and PgRSSCD estimated using the bins defined in Dimitriadis et al. (2021) are shown in Figure 2.4. These lead to similar conclusions as when 10 bins are used in each case.

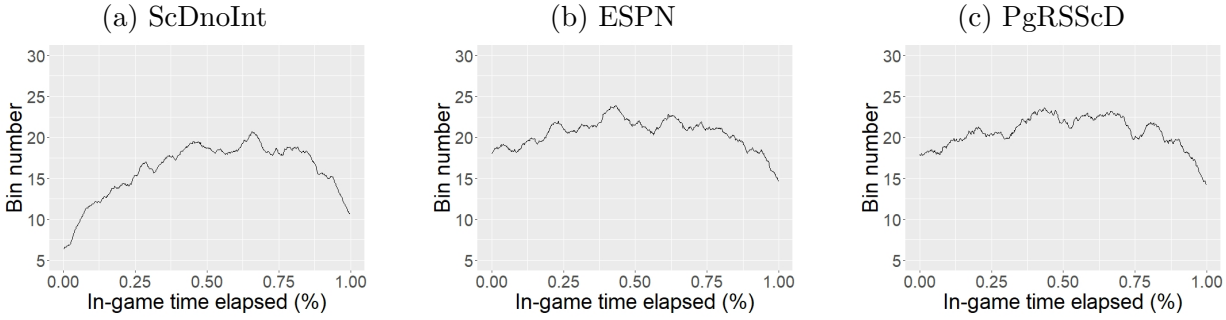


Figure A.1: The change in number of bins across time using the CORP method: Methods (a) logit model using Score difference without intercept (ScDnoInt); (b) **ESPN** forecasts; (c) logit model based on Pre-game relative strength and Score difference (**PgRSScD**).

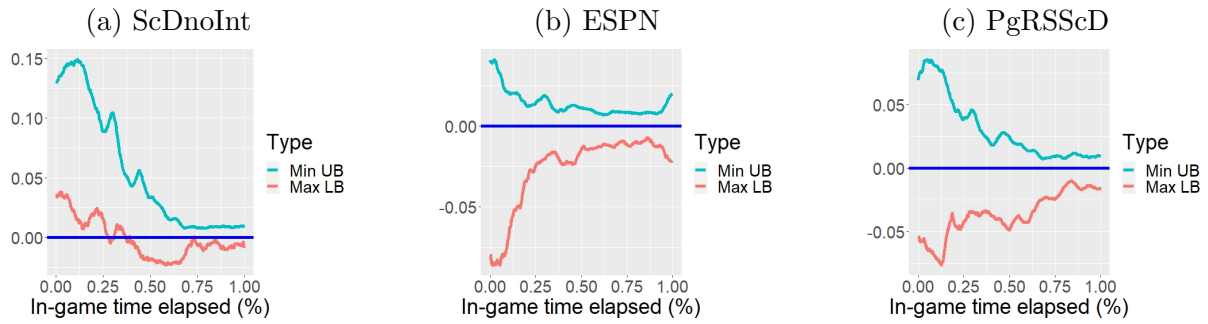


Figure A.2: Plots of $U_{0.95}^{min}(t)$ and $L_{0.95}^{max}(t)$ against t based on the bins obtained from CORP method: Methods (a) logit model using Score difference without intercept (**ScDnoInt**); (b) **ESPN** forecasts; (c) logit model based on Pre-game relative strength and Score difference (**PgRSScD**).

Appendix B

Technical Details of Chapter 3

B.1 Technical Details and Proofs

Throughout these proofs, we assume without loss of generality that the spatial median $\mu = 0$. We use C_i, $i \geq 1$ to denote unimportant positive numerical constants that may change from line to line.

B.1.1 Proof of Theorem 1

Proof B.1.1 *We use the notation*

$$\sqrt{N}\tilde{\rho}_h = \frac{1}{\sqrt{N}} \sum_{i=1}^{N-h} \langle S(X_i), S(X_{i+h}) \rangle =: \frac{1}{\sqrt{N}} \sum_{i=1}^{N-h} T_{i,h}. \tag{B.1}$$

$\{T_{i,h}\}_{i\in\mathbb{Z}}$ forms a strictly stationary and $h-$dependent sequence of scalar random variables. By equation (2) of Gervini (2008) and the independence of the X_i's, $\mathbb{E}T_{i,h} = 0$. Using Fubini's theorem and the independence of X_i's, we obtain that

$$\begin{aligned}
\mathbb{E}T_{i,h}^2 &= \mathbb{E}\left\langle S(X_i), S(X_{i+h})\right\rangle^2 \\
&= \iint \mathbb{E}S(X_i)(t)S(X_{i+h})(t)S(X_i)(s)S(X_{i+h})(s)dtds = \|C_P\|_2^2.
\end{aligned}$$

Once again by equation (2) of Gervini (2008) and the independence of the X_is, we obtain that for $i \neq j$, $\mathbb{E}T_{i,h}T_{j,h} = 0$. Hence by Theorem 6.4.2 of Brockwell and Davis (1991), $\sqrt{N}\tilde{\rho} \xrightarrow{D} \mathcal{N}(0, \|C_P\|_2^2)$. It follows by elementary arguments that for fixed H, $\sqrt{N}\|\tilde{R}_H - R_H\|_E \xrightarrow{P} 0$, where

$$R_H = \frac{1}{N}\sum_{i=1}^{N}(T_{i,1}, \ldots, T_{i,H})^\top =: \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{T}_{i,H}.$$

Similarly as above, the vector valued process $\{\boldsymbol{T}_{i,H}\}_{i\in\mathbb{Z}}$ is strictly stationary, mean zero, and H dependent, with $\mathbb{E}\boldsymbol{T}_{i,H}\boldsymbol{T}_{i,H}^\top = \|C_P\|^2\mathbb{I}_H$, and $\mathbb{E}\boldsymbol{T}_{i,H}\boldsymbol{T}_{j,H}^\top = 0$, if $i \neq j$. Hence by Proposition 11.2.2 of Brockwell and Davis (1991), $\sqrt{N}\tilde{R}_H \xrightarrow{D} \mathcal{N}_H(0, \|C_P\|^2\mathbb{I}_H)$. It follows from the continuous mapping theorem that $\|\sqrt{N}\tilde{R}_H\|_E^2 \xrightarrow{D} \|C_P\|^2\chi^2(H)$.

B.1.2 Proof of Theorem 3.2.2

Lemma B.1.1 *Under Assumption 3.2.2, if $\|u\| \leq C_1$, then there exists a constant C_3 so that*

$$\mathbb{E}\frac{\|X_i\|^2}{\|X_i - u\|^2} \leq C_3.$$

Proof B.1.2 *For all $M > 0$,*

$$\begin{aligned}\mathbb{E}\frac{\|X_i\|^2}{\|X_i - u\|^2} &= \mathbb{E}\frac{\|X_i\|^2}{\|X_i - u\|^2}\mathbb{1}\{\|X_i - u\| > M\} + \mathbb{E}\frac{\|X_i\|^2}{\|X_i - u\|^2}\mathbb{1}\{\|X_i - u\| \leq M\} \\ &\leq \mathbb{E}\frac{\|X_i\|^2}{\|X_i - u\|^2}\mathbb{1}\{\|X_i - u\| > M\} + [M + C_1]C_2,\end{aligned}$$

where in the second line we applied Assumption 3.2.2, and used that $\|X_i - u\| \leq M$ implies by the triangle inequality that $\|X_i\| \leq M + C_1$. Again by the triangle inequality $\|X_i\| - C_1 \leq \|X_i - u\|$, which implies that so long as $\|X_i\| > C_1$,

$$\frac{\|X_i\|}{\|X_i - u\|} \leq \frac{\|X_i\|}{\|X_i\| - C_1}.$$

Since $\|X_i - u\| \leq \|X_i\| + C_1$, we have if $M > 2C_1$, then $\|X_i - u\| > M$ implies that $\|X_i\| > C_1$. This implies then with $M > 2C_1$ that

$$\frac{\|X_i\|^2}{[\|X_i\| - C_1]^2}\mathbb{1}\{\|X_i - u\| > M\} \leq C_5$$

almost surely for a positive constant C_5. *Hence Lemma B.1.1 holds with* $C_3 = C_5 + [M + C_1]C_2$ *for any* $M > 2C_1$.

Theorem 3.2.2 follows immediately from Theorem 3.2.1 and the following additional lemma.

Lemma B.1.2 *Under Assumptions 3.2.1 and 3.2.2,* $\sqrt{N}(\hat{\rho}_h - \tilde{\rho}_h) = o_P(1)$.

Proof B.1.3 *Letting* $\hat{m}_{i,h} = \|X_i\|\|X_{i+h}\|\|X_i - \hat{\mu}\|\|X_{i+h} - \hat{\mu}\|$, *we have that*

$$
\begin{aligned}
&\sqrt{N}(\hat{\rho}_h - \tilde{\rho}_h) \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} [\langle S(X_i - \hat{\mu}), S(X_{i+h} - \hat{\mu})\rangle - \langle S(X_i), S(X_{i+h})\rangle] \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{1}{\hat{m}_{i,h}} [\langle (X_i - \hat{\mu})\|X_i\|, (X_{i+h} - \hat{\mu})\|X_{i+h}\|\rangle \\
&\quad - \langle X_i\|X_i - \hat{\mu}\|, X_{i+h}\|X_{i+h} - \hat{\mu}\|\rangle] \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{1}{\hat{m}_{i,h}} [\langle X_i\|X_i\|, X_{i+h}\|X_{i+h}\|\rangle - \langle X_i\|X_i - \hat{\mu}\|, X_{i+h}\|X_{i+h} - \hat{\mu}\|\rangle] \\
&\quad - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\langle \hat{\mu}\|X_i\|, X_{i+h}\|X_{i+h}\|\rangle}{\hat{m}_{i,h}} - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\langle X_i\|X_i\|, \hat{\mu}\|X_{i+h}\|\rangle}{\hat{m}_{i,h}} \\
&\quad + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\langle \hat{\mu}\|X_i\|, \hat{\mu}\|X_{i+h}\|\rangle}{\hat{m}_{i,h}} \\
&=: R_{1,N} - R_{2,N} - R_{3,N} + R_{4,N}.
\end{aligned}
$$

We now aim to show that $R_{j,N} = o_P(1)$, $j = 1, \ldots, 4$. *Starting with* $R_{2,N}$, *we have by the linearity of the inner product and the Cauchy-Schwarz inequality that*

$$
|R_{2,N}| = \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\langle \hat{\mu}\|X_i\|, X_{i+h}\|X_{i+h}\|\rangle}{\hat{m}_{i,h}} \right| = \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{\langle \hat{\mu}, X_{i+h}\rangle}{\|X_i - \hat{\mu}\|\|X_{i+h} - \hat{\mu}\|} \right| \tag{B.2}
$$

$$= \left| \left\langle \hat{\mu}, \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \frac{X_{i+h}}{\|X_i - \hat{\mu}\| \|X_{i+h} - \hat{\mu}\|} \right\rangle \right|$$
$$\leq \|\sqrt{N}\hat{\mu}\| \left\| \frac{1}{N} \sum_{i=1}^{N} \frac{X_{i+h}}{\|X_i - \hat{\mu}\| \|X_{i+h} - \hat{\mu}\|} \right\| =: \|\sqrt{N}\hat{\mu}\| \|R_{2,N}^*\|. \tag{B.3}$$

By Assumption 3.2.1, $\|\sqrt{N}\hat{\mu}\| = O_P(1)$. Let $\epsilon, \delta > 0$. Also by Assumption 3.2.1, there exists a compact set $K \subset \mathcal{H}$ so that

$$\mathbb{P}(\|R_{2,N}^*\| > \epsilon) \leq \frac{\delta}{2} + P(\{\|R_{2,N}^*\| > \epsilon\} \cap \{\sqrt{N}\hat{\mu} \in K\}). \tag{B.4}$$

Define the process $G_{2,N} : \mathcal{H} \to \mathcal{H}$ by

$$G_{2,N}(u) = \frac{1}{N} \sum_{i=1}^{N} \frac{X_{i+h}}{\|X_i - N^{-1/2}u\| \|X_{i+h} - N^{-1/2}u\|},$$

so that $R_{2,N}^* = G_{2,N}(\sqrt{N}\hat{\mu})$. Then

$$\mathbb{P}(\{\|R_{2,N}^*\| > \epsilon\} \cap \{\sqrt{N}\hat{\mu} \in K\}) \leq \mathbb{P}\left(\sup_{u \in K} \|G_{2,n}(u)\| > \epsilon \right). \tag{B.5}$$

Using that $\{X_i\}_{i \in \mathbb{Z}}$ are i.i.d., we get that

$$\mathbb{E}G_{2,N}(u) = \mathbb{E}\|X_0 - N^{-1/2}u\|^{-1} \mathbb{E} \frac{X_h}{\|X_h - N^{-1/2}u\|}.$$

Notice that for each $u \in K$,

$$Y_{h,N} := \left\| \frac{X_h}{\|X_h - N^{-1/2}u\|} - \frac{X_h}{\|X_h\|} \right\| \overset{a.s.}{\to} 0 \text{ as } N \to \infty.$$

Moreover, it follows from Lemma B.1.1 that $\{Y_{h,N}\}_{N \geq 1}$ are uniformly integrable. Hence by the Vitali convergence theorem (see e.g. Theorem 16.14 of Billingsley, 1995), $\mathbb{E}Y_{h,N} \to 0$ as $N \to \infty$. According to the contraction property of expectation in Hilbert space (see pg. 29 of Bosq, 2000), it follows that

$$\left\| \mathbb{E} \frac{X_h}{\|X_h - N^{-1/2}u\|} \right\| \to 0 \text{ as } N \to \infty. \tag{B.6}$$

Since $\mathbb{E}\|X_0 - N^{-1/2}u\|^{-1} < \infty$ *by Assumption 3.2.2, we then obtain that* $\mathbb{E}G_{2,N}(u) \to 0$ *as* $N \to \infty$ *for each* $u \in K$. *Also for* $u, u' \in K$, *we obtain using simple arithmetic with the reverse triangle inequality and again Assumption 3.2.2 that*

$$\begin{aligned}
&\|\mathbb{E}G_{2,N}(u) - \mathbb{E}G_{2,N}(u')\| \\
&= \mathbb{E}\frac{\|X_0 - N^{-1/2}u\| - \|X_0 - N^{-1/2}u'\|}{\|X_0 - N^{-1/2}u\|\|X_0 - N^{-1/2}u'\|}\left\|\mathbb{E}\frac{X_h}{\|X_h - N^{-1/2}u\|}\right\| \\
&\quad + \left\|\mathbb{E}\frac{X_h[\|X_h - N^{-1/2}u\| - \|X_h - N^{-1/2}u'\|]}{\|X_h - N^{-1/2}u\|\|X_h - N^{-1/2}u'\|}\right\|\mathbb{E}\|X_0 - N^{-1/2}u'\|^{-1} \\
&= C_6\|u - u'\|.
\end{aligned}$$

Hence $\mathbb{E}G_{2,N}(u)$ *is uniformly continuous, from which it follows in conjunction with* $\mathbb{E}G_{2,N}(u) = o(1)$ *that*

$$\sup_{u\in K}\|\mathbb{E}G_{2,N}(u)\| = o(1), \text{ as } N \to \infty. \tag{B.7}$$

Also, using that $\{X_i\}_{i\in\mathbb{Z}}$ *are i.i.d., we obtain that*

$$\begin{aligned}
&\mathbb{E}\|G_{2,N}(u) - \mathbb{E}G_{2,N}(u)\|^2 \\
&= \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\left[\frac{X_{i+h}}{\|X_i - N^{-1/2}u\|\|X_{i+h} - N^{-1/2}u\|} - \mathbb{E}G_{2,N}(u)\right]\right\|^2 \\
&= \frac{1}{N^2}\left[\sum_{i=1}^{N}\mathbb{E}\left\|\frac{X_{i+h}}{\|X_i - N^{-1/2}u\|\|X_{i+h} - N^{-1/2}u\|} - \mathbb{E}G_{2,N}(u)\right\|^2\right. \\
&+ \sum_{1\le i\ne j\le N}\mathbb{E}\left\langle\frac{X_{i+h}}{\|X_i - N^{-1/2}u\|\|X_{i+h} - N^{-1/2}u\|} - \mathbb{E}G_{2,N}(u),\right. \\
&\qquad\left.\left.\frac{X_{j+h}}{\|X_j - N^{-1/2}u\|\|X_{j+h} - N^{-1/2}u\|} - \mathbb{E}G_{2,N}(u)\right\rangle\right].
\end{aligned}$$

When $i \ne j$, *and* $|j - i| \ne h$, *it follows due to the assumption that the* $\{X_i\}_{i\in\mathbb{Z}}$ *are i.i.d. that*

$$\mathbb{E}\left\langle\frac{X_{i+h}}{\|X_i - N^{-1/2}u\|\|X_{i+h} - N^{-1/2}u\|} - \mathbb{E}G_{2,N}(u),\right.$$

$$\frac{X_{j+h}}{\|X_j - N^{-1/2}u\|\|X_{j+h} - N^{-1/2}u\|} - \mathbb{E}G_{2,N}(u)\Big\rangle = 0.$$

And when $|j-i| = h$, by the Cauchy-Schwarz inequality the same inner product term is bounded by

$$\mathbb{E}\left\|\frac{X_h}{\|X_0 - N^{-1/2}u\|\|X_h - N^{-1/2}u\|} - \mathbb{E}G_{2,N}(u)\right\|^2. \tag{B.8}$$

Hence

$$\mathbb{E}\|G_{2,N}(u) - \mathbb{E}G_{2,N}(u)\|^2 \leq \frac{C_7}{N}\mathbb{E}\left\|\frac{X_h}{\|X_0 - N^{-1/2}u\|\|X_h - N^{-1/2}u\|} - \mathbb{E}G_{2,N}(u)\right\|^2. \tag{B.9}$$

By Assumption 3.2.2 and Lemma B.1.1, for all n sufficiently large,

$$\mathbb{E}\left\|\frac{X_h}{\|X_0 - N^{-1/2}u\|\|X_h - N^{-1/2}u\|}\right\|^2 = \mathbb{E}\|X_0 - N^{-1/2}u\|^{-2}\mathbb{E}\left\|\frac{X_h}{\|X_h - N^{-1/2}u\|}\right\|^2 \leq C_8,$$

which implies the term in (B.8) is bounded for all n sufficiently large. This along with (B.9) gives that

$$G_{2,N}(u) \xrightarrow{P} \mathbb{E}G_{2,N}(u), \text{ for each } u \in K. \tag{B.10}$$

We now aim to show $G_{2,n}$ is stochastically Lipshitz. To that end, for $u, u' \in K$, simple arithmetic gives that

$$\begin{aligned}
&\|G_{2,N}(u) - G_{2,N}(u')\| \\
&= \left\|\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{\|X_i - N^{-1/2}u\|} - \frac{1}{\|X_i - N^{-1/2}u'\|}\right]\frac{X_{i+h}}{\|X_{i+h} - N^{-1/2}u\|}\right. \\
&\quad \left. + \frac{1}{N}\sum_{i=1}^{N}\left[\frac{X_{i+h}}{\|X_i - N^{-1/2}u\|} - \frac{X_{i+h}}{\|X_i - N^{-1/2}u'\|}\right]\right.
\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad \times \frac{1}{\|X_i - N^{-1/2}u'\|}\frac{X_{i+h}}{\|X_i - N^{-1/2}u\|}\Bigg\| \\
&\leq \left\| \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{\|X_i - N^{-1/2}u\|} - \frac{1}{\|X_i - N^{-1/2}u'\|}\right]\frac{X_{i+h}}{\|X_i - N^{-1/2}u\|}\right\| \\
&\quad + \left\| \frac{1}{N}\sum_{i=1}^{N}\left[\frac{X_{i+h}}{\|X_i - N^{-1/2}u\|} - \frac{X_{i+h}}{\|X_i - N^{-1/2}u'\|}\right]\right. \\
&\qquad\qquad \left. \times \frac{1}{\|X_i - N^{-1/2}u'\|}\frac{X_{i+h}}{\|X_i - N^{-1/2}u\|}\right\| \\
&=: \|L_{1,N}(u,u')\| + \|L_{2,N}(u,u')\|.
\end{aligned}
$$

Notice that by the triangle inequality,

$$
\begin{aligned}
&\|L_{1,N}(u,u')\| \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\frac{N^{-1/2}\|u-u'\|}{\|X_i - N^{-1/2}u\|\|X_i - N^{-1/2}u'\|}\left\|\frac{X_{i+h}}{\|X_{i+h} - N^{-1/2}u\|}\right\| \\
&= \|u-u'\|\frac{1}{N^{3/2}}\sum_{i=1}^{N}\frac{1}{\|X_i - N^{-1/2}u\|\|X_i - N^{-1/2}u'\|}\left\|\frac{X_{i+h}}{\|X_{i+h} - N^{-1/2}u\|}\right\| \\
&=: \|u-u'\|B_N,
\end{aligned}
$$

where by Assumption 3.2.2 and Lemma B.1.1, $\mathbb{E}B_n \leq C_9$. This implies that the process $\{G_{2,N}(u),\ u \in K\}$ satisfies the equicontinuity conditions of Assumption 3A of Newey (1991). This along with (B.10) imply according to Theorem 2.1 of Newey (1991) that

$$
\sup_{u\in K}\|G_{2,N} - \mathbb{E}G_{2,N}\| = o_P(1).
$$

In conjunction with (B.7), this shows that $\sup_{u\in K}\|G_{2,N}(u)\| = o_P(1)$. Due to (B.2), (B.4), and (B.5), it follows that $R_{2,N} = o_P(1)$. Following a nearly identical argument, it may be shown that $R_{3,N} = o_P(1)$ and $R_{4,N} = o_P(1)$. Although the argument is also similar to show that $R_{1,N} = o_P(1)$, it differs in a few places and so we think it is valuable to provide the details. Notice that we may write

$$
R_{1,N} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{\hat{m}_{i,h}}\left[\langle X_i\|X_i\|, X_{i+h}\|X_{i+h}\|\rangle\right.
$$

$$
\begin{aligned}
&\quad -\langle X_i\|X_i-\hat{\mu}\|, X_{i+h}\|X_{i+h}-\hat{\mu}\|\rangle] \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{\hat{m}_{i,h}}\langle X_i\|X_i\|, X_{i+h}[\|X_{i+h}\|-\|X_{i+h}-\hat{\mu}\|]\rangle \\
&\quad + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{1}{\hat{m}_{i,h}}\langle X_i[\|X_i\|-\|X_i-\hat{\mu}\|], X_{i+h}\|X_{i+h}-\hat{\mu}\|\rangle \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\langle X_i\|X_i\|, X_{i+h}[\|X_{i+h}\|-\|X_{i+h}-\hat{\mu}\|]\rangle}{\|X_{i+h}\|\|X_i-\hat{\mu}\|\|X_{i+h}-\hat{\mu}\|} \\
&\quad + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\langle X_i[\|X_i\|-\|X_i-\hat{\mu}\|], X_{i+h}\|X_{i+h}-\hat{\mu}\|\rangle}{\|X_{i+h}\|\|X_i-\hat{\mu}\|\|X_i\|} \\
&=: R_{1,N}^{(1)} + R_{1,N}^{(2)}.
\end{aligned}
$$

Define

$$
G_{1,N}^{(1)}(u) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\langle X_i\|X_i\|, X_{i+h}[\|X_{i+h}\|-\|X_{i+h}-N^{-1/2}u\|]\rangle}{\|X_{i+h}\|\|X_i-N^{-1/2}u\|\|X_{i+h}-N^{-1/2}u\|}
$$

so that $R_{1,N}^{(1)} = G_{1,N}^{(1)}(\sqrt{N}\hat{\mu})$. *Arguing as in* (B.5), *it is enough to show that for any compact* $K \subset \mathcal{H}$,

$$
\sup_{u\in K}|G_{1,N}^{(1)}(u)| = o_P(1).
$$

Due the independence of the $X_i's$ *and Fubini's Theorem,*

$$
\mathbb{E}G_{1,N}^{(1)}(u) = \sqrt{N}\left\langle \mathbb{E}\frac{X_0}{\|X_0-N^{-1/2}u\|}, \mathbb{E}\frac{X_h[\|X_h\|-\|X_h-N^{-1/2}u\|]}{\|X_h-N^{-1/2}u\|\|X_h\|}\right\rangle.
$$

By the Cauchy-Schwarz and reverse triangle inequality, (B.6), *and Assumption* 3.2.2, *it follows that*

$$
\begin{aligned}
\mathbb{E}G_{1,N}^{(1)}(u) &\le \sqrt{N}\left\|\mathbb{E}\frac{X_0}{\|X_0-N^{-1/2}u\|}\right\|\left\|\mathbb{E}\frac{X_h[\|X_h\|-\|X_h-N^{-1/2}u\|]}{\|X_h-N^{-1/2}u\|\|X_h\|}\right\| \\
&\le C_{10}\left\|\mathbb{E}\frac{X_0}{\|X_0-N^{-1/2}u\|}\right\|\mathbb{E}\|X_h-N^{-1/2}u\|^{-1} \to 0,
\end{aligned} \tag{B.11}
$$

as $N \to \infty$ for each $u \in K$. Moreover, noting that (X_i, X_{i+h}) is independent of (X_j, X_{j+h}) when $i \neq j$, $|i-j| \neq h$, we obtain along with the Cauchy-Schwarz inequality that

$$\begin{aligned} Var(G_{1,N}^{(1)}(u)) &= \frac{1}{N}\Bigg[\sum_{i=1}^{N} Var\left(\frac{\langle X_i\|X_i\|, X_{i+h}[\|X_{i+h}\| - \|X_{i+h}-\hat{\mu}\|]\rangle}{\|X_{i+h}\|\|X_i-\hat{\mu}\|\|X_{i+h}-\hat{\mu}\|}\right) \\ &\quad + 2\sum_{1\le i<j\le N} Cov\Bigg(\frac{\langle X_i\|X_i\|, X_{i+h}[\|X_{i+h}\| - \|X_{i+h}-\hat{\mu}\|]\rangle}{\|X_{i+h}\|\|X_i-\hat{\mu}\|\|X_{i+h}-\hat{\mu}\|}, \\ &\qquad \frac{\langle X_j\|X_j\|, X_{j+h}[\|X_{j+h}\| - \|X_{j+h}-\hat{\mu}\|]\rangle}{\|X_{j+h}\|\|X_j-\hat{\mu}\|\|X_{j+h}-\hat{\mu}\|}\Bigg)\Bigg] \\ &\le C_{11}\, Var\left(\frac{\langle X_i\|X_i\|, X_{i+h}[\|X_{i+h}\| - \|X_{i+h}-\hat{\mu}\|]\rangle}{\|X_{i+h}\|\|X_i-\hat{\mu}\|\|X_{i+h}-\hat{\mu}\|}\right). \end{aligned}$$

Again applying Assumption 3.2.2 and the reverse triangle inequality, we obtain that

$$Var\left(\frac{\langle X_i\|X_i\|, X_{i+h}[\|X_{i+h}\| - \|X_{i+h}-\hat{\mu}\|]\rangle}{\|X_{i+h}\|\|X_i-\hat{\mu}\|\|X_{i+h}-\hat{\mu}\|}\right) \le C_{12}N^{-1/2} \to 0 \text{ as } N \to \infty.$$

This implies by Chebyshev's inequality that $G_{1,n}^{(1)}(u) - \mathbb{E}G_{1,n}^{(1)}(u) = o_P(1)$. In order to establish the stochastic equicontinuity of $G_{1,N}^{(1)}$, we note that for $u, u' \in K$,

$$\begin{aligned} |G_{1,N}^{(1)}(u) - G_{1,N}^{(1)}(u')| &= \Bigg|\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \frac{\langle X_i\|X_i\|, X_{i+h}[\|X_{i+h}\| - \|X_{i+h}-N^{-1/2}u\|]\rangle}{\|X_{i+h}\|\|X_i-N^{-1/2}u\|\|X_{i+h}-N^{-1/2}u\|} \\ &\qquad - \frac{\langle X_i\|X_i\|, X_{i+h}[\|X_{i+h}\| - \|X_{i+h}-N^{-1/2}u'\|]\rangle}{\|X_{i+h}\|\|X_i-N^{-1/2}u'\|\|X_{i+h}-N^{-1/2}u'\|}\Bigg| \\ &\le \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Bigg[\Bigg|\frac{1}{\|X_i-N^{-1/2}u\|\|X_{i+h}-N^{-1/2}u\|} \\ &\qquad - \frac{1}{\|X_i-N^{-1/2}u'\|\|X_{i+h}-N^{-1/2}u'\|}\Bigg| \\ &\qquad \times \left|\left\langle X_i\|X_i\|, X_{i+h}\frac{[\|X_{i+h}-N^{-1/2}u\| - \|X_{i+h}-N^{-1/2}u'\|]}{\|X_{i+h}\|}\right\rangle\right|\Bigg] \\ &\quad + \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Bigg[\Bigg|\frac{1}{\|X_i-N^{-1/2}u\|\|X_{i+h}-N^{-1/2}u\|} \end{aligned}$$

$$
\left. - \frac{1}{\|X_i - N^{-1/2}u'\| \|X_{i+h} - N^{-1/2}u'\|} \right|
$$

$$
\times \left[\left| \left\langle X_i \|X_i\|, X_{i+h} \frac{[\|X_{i+h}\| - \|X_{i+h} - N^{-1/2}u\|]}{\|X_{i+h}\|} \right\rangle \right| \right]
$$

$$
=: L_{1,N}(u,u') + L_{2,N}(u,u').
$$

By Cauchy-Schwarz and the reverse triangle inequalities,

$$
L_{1,N}(u,u') \leq \|u-u'\| \frac{1}{N} \sum_{i=1}^{N} \frac{\|X_i\|}{\|X_i - N^{-1/2}u\| \|X_{i+h} - N^{-1/2}u'\|} =: \|u-u'\| B_{N,1}.
$$

By Assumption 3.2.2 and Lemma B.1.1, $EB_{1,n} < \infty$. It follows similarly that

$$
L_{2,N}(u,u') \leq \|u-u'\| B_{N,2}, \text{ with } \mathbb{E}B_{2,N} < \infty.
$$

Hence $|G^{(1)}_{1,N}(u) - G^{(1)}_{1,N}(u')| \leq \|u-u'\| B^*_N$, where $EB^*_N < \infty$. It follows once again by Theorem 2.1 of Newey (1991) that

$$
\sup_{u \in K} |G^{(1)}_{1,N}(u) - \mathbb{E}G^{(1)}_{1,N}(u)| = o_P(1).
$$

Owing then to (B.11), $\sup_{u \in K} |G^{(1)}_{1,N}(u)| = o_P(1)$. It follows similarly that $\sup_{u \in K} |G^{(2)}_{1,N}(u)| = o_P(1)$, and hence $R_{1,N} = o_P(1)$, giving the result.

B.1.3 Proof of Theorem 3.2.3

Proof B.1.4 *We first note that*

$$
\left| \frac{1}{N} \sum_{i=1}^{N-h} \langle S(X_i - \hat{\mu}), S(X_{i+h} - \hat{\mu}) \rangle - \frac{1}{N} \sum_{i=1}^{N} \langle S(X_i - \hat{\mu}), S(X_{i+h} - \hat{\mu}) \rangle \right| \leq \frac{h}{N} \overset{a.s.}{\to} 0.
$$

Thus the theorem follows upon showing that

$$
\left| \frac{1}{N} \sum_{i=1}^{N} \langle S(X_i - \hat{\mu}), S(X_{i+h} - \hat{\mu}) \rangle - \frac{1}{N} \sum_{i=1}^{N} \langle S(X_i), S(X_{i+h}) \rangle \right| \overset{a.s.}{\to} 0.
$$

Let $A_N = \{(x,y) \in \mathcal{H} \times \mathcal{H} : 2\|x - \hat{\mu}\| > \|x\| \text{ and } 2\|y - \hat{\mu}\| > \|y\|\}$. *Then*

$$\begin{aligned}
&\left| \frac{1}{N} \sum_{i=1}^{N} \langle S(X_i - \hat{\mu}), S(X_{i+h} - \hat{\mu}) \rangle - \frac{1}{N} \sum_{i=1}^{N} \langle S(X_i), S(X_{i+h}) \rangle \right| \\
&= \frac{1}{N} \sum_{i=1}^{N} |\langle S(X_i - \hat{\mu}), S(X_{i+h} - \hat{\mu}) \rangle - \langle S(X_i), S(X_{i+h}) \rangle| \\
&= \frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n\}} |\langle S(X_i - \hat{\mu}), S(X_{i+h} - \hat{\mu}) \rangle - \langle S(X_i), S(X_{i+h}) \rangle| \\
&+ \frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n^c\}} |\langle S(X_i - \hat{\mu}), S(X_{i+h} - \hat{\mu}) \rangle - \langle S(X_i), S(X_{i+h}) \rangle| \\
&=: S_{N,1} + S_{N,2}.
\end{aligned}$$

We aim to show $S_{N,1} \overset{a.s.}{\to} 0$ *and* $S_{N,2} \overset{a.s.}{\to} 0$. *Looking at the first term* $S_{N,1}$, *we have*

$$\begin{aligned}
&S_{N,1} \\
&= \frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n\}} \left| \left\langle \frac{X_i - \hat{\mu}}{\|X_i - \hat{\mu}\|}, \frac{X_{i+h} - \hat{\mu}}{\|X_{i+h} - \hat{\mu}\|} \right\rangle - \left\langle \frac{X_i}{\|X_i\|}, \frac{X_{i+h}}{\|X_{i+h}\|} \right\rangle \right| \\
&= \frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n\}} \left| \left\langle \frac{(X_i - \hat{\mu})\|X_i\|}{\|X_i - \hat{\mu}\|\,\|X_i\|}, \frac{(X_{i+h} - \hat{\mu})\|X_{i+h}\|}{\|X_{i+h} - \hat{\mu}\|\,\|X_{i+h}\|} \right\rangle \right. \\
&\qquad \left. - \left\langle \frac{X_i\|X_i - \hat{\mu}\|}{\|X_i\|\,\|X_i - \hat{\mu}\|}, \frac{X_{i+h}\|X_{i+h} - \hat{\mu}\|}{\|X_{i+h}\|\,\|X_{i+h} - \hat{\mu}\|} \right\rangle \right| \\
&= \frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n\}} \frac{1}{\hat{m}_{i,h}} \times \\
&|\langle (X_i - \hat{\mu})\|X_i\|, (X_{i+h} - \hat{\mu})\|X_{i+h}\| \rangle - \langle X_i(\|X_i - \hat{\mu}\|), X_{i+h}(\|X_{i+h} - \hat{\mu}\|) \rangle|,
\end{aligned}$$

where again $\hat{m}_{i,h} = \|X_i\|\,\|X_{i+h}\|\,\|X_i - \hat{\mu}\|\,\|X_{i+h} - \hat{\mu}\|$. *By expanding the summand and applying the triangle inequality, we get that*

$$\begin{aligned}
&\frac{1}{\hat{m}_{i,h}} |\langle (X_i - \hat{\mu})\|X_i\|, (X_{i+h} - \hat{\mu})\|X_{i+h}\| \rangle \\
&\quad - \langle X_i(\|X_i - \hat{\mu}\|), X_{i+h}(\|X_{i+h} - \hat{\mu}\|) \rangle|
\end{aligned}$$

$$= \frac{1}{\hat{m}_{i,h}} \left| \{ \langle X_i \left\| X_i \right\|, X_{i+h} \left\| X_{i+h} \right\| \rangle - \langle X_i (\|X_i - \hat{\mu}\|), X_{i+h} (\|X_{i+h} - \hat{\mu}\|) \rangle \} \right.$$
$$\left. - \langle \hat{\mu} \left\| X_i \right\|, X_{i+h} \left\| X_{i+h} \right\| \rangle - \langle X_i \left\| X_i \right\|, \hat{\mu} \left\| X_{i+h} \right\| \rangle + \langle \hat{\mu} \left\| X_i \right\|, \hat{\mu} \left\| X_{i+h} \right\| \rangle \right|$$
$$= \frac{1}{\hat{m}_{i,h}} \left| S_{N,i}^{(1)} - S_{N,i}^{(2)} - S_{N,i}^{(3)} + S_{N,i}^{(4)} \right| \leq \frac{1}{\hat{m}_{i,h}} \left(\left| S_{N,i}^{(1)} \right| + \left| S_{N,i}^{(2)} \right| + \left| S_{N,i}^{(3)} \right| + \left| S_{N,i}^{(4)} \right| \right),$$

where

$$S_{N,i}^{(1)} = \langle X_i \left\| X_i \right\|, X_{i+h} \left\| X_{i+h} \right\| \rangle - \langle X_i (\|X_i - \hat{\mu}\|), X_{i+h} (\|X_{i+h} - \hat{\mu}\|) \rangle,$$
$$S_{N,i}^{(2)} = \langle \hat{\mu} \left\| X_i \right\|, X_{i+h} \left\| X_{i+h} \right\| \rangle,$$
$$S_{N,i}^{(3)} = \langle X_i \left\| X_i \right\|, \hat{\mu} \left\| X_{i+h} \right\| \rangle, \quad \textit{and } S_{N,i}^{(4)} = \langle \hat{\mu} \left\| X_i \right\|, \hat{\mu} \left\| X_{i+h} \right\| \rangle.$$

By adding and subtracting $\langle X_i \left\| X_i \right\|, \left\| X_{i+h} \right\| \left\| X_{i+h} - \hat{\mu} \right\| \rangle$ *and using the triangle inequality, we have*

$$\frac{1}{\hat{m}_{i,h}} \left| S_{N,i}^{(1)} \right| = \frac{1}{\hat{m}_{i,h}} \left| \{ \langle X_i \left\| X_i \right\|, X_{i+h} \left\| X_{i+h} \right\| \rangle - \langle X_i \left\| X_i \right\|, X_{i+h} \left\| X_{i+h} - \hat{\mu} \right\| \rangle \} \right.$$
$$\left. + \{ \langle X_i \left\| X_i \right\|, X_{i+h} \left\| X_{i+h} - \hat{\mu} \right\| \rangle - \langle X_i (\|X_i - \hat{\mu}\|), X_{i+h} (\|X_{i+h} - \hat{\mu}\|) \rangle \} \right|$$
$$=: \frac{1}{\hat{m}_{i,h}} \left| S_{N,i}^{(1,1)} + S_{N,i}^{(1,2)} \right| \leq \frac{1}{\hat{m}_{i,h}} \left| S_{N,i}^{(1,1)} \right| + \frac{1}{\hat{m}_{i,h}} \left| S_{i,i}^{(1,2)} \right|,$$

where

$$S_{N,i}^{(1,1)} = \langle X_i \left\| X_i \right\|, X_{i+h} \left\| X_{i+h} \right\| \rangle - \langle X_i \left\| X_i \right\|, X_{i+h} \left\| X_{i+h} - \hat{\mu} \right\| \rangle \quad \textit{and}$$
$$S_{N,i}^{(1,2)} = \langle X_i \left\| X_i \right\|, X_{i+h} \left\| X_{i+h} - \hat{\mu} \right\| \rangle - \langle X_i (\|X_i - \hat{\mu}\|), X_{i+h} (\|X_{i+h} - \hat{\mu}\|) \rangle.$$

Using the reverse triangle inequality and Cauchy–Schwarz inequality, we obtain that

$$\frac{1}{\hat{m}_{i,h}} \left| S_{N,1}^{(1,1)} \right| = \frac{1}{\hat{m}_{i,h}} \left| \langle X_i \left\| X_i \right\|, X_{i+h} (\|X_{i+h}\| - \|X_{i+h} - \hat{\mu}\|) \rangle \right|$$
$$= \frac{(\|X_{i+h}\| - \|X_{i+h} - \hat{\mu}\|) \left\| X_i \right\|}{\hat{m}_{i,h}} \langle X_i, X_{i+h} \rangle$$
$$\leq \frac{\|\hat{\mu}\| \left\| X_i \right\|}{\hat{m}_{i,h}} \langle X_i, X_{i+h} \rangle \leq \frac{\|\hat{\mu}\| \left\| X_i \right\|^2 \left\| X_{i+h} \right\|}{\hat{m}_{i,h}}$$
$$\leq \frac{\|\hat{\mu}\| \left\| X_i \right\|^2 \left\| X_{i+h} \right\|}{\left\| X_i \right\| \left\| X_{i+h} \right\| \left\| X_i - \hat{\mu} \right\| \left\| X_{i+h} - \hat{\mu} \right\|}$$

$$= \frac{\|X_i\| \, \|\hat{\mu}\|}{\|X_i - \hat{\mu}\| \, \|X_{i+h} - \hat{\mu}\|}.$$

For $(X_i, X_{i+h}) \in A_N$*, it follows that*

$$\frac{\|X_i\| \, \|\hat{\mu}\|}{\|X_i - \hat{\mu}\| \, \|X_{i+h} - \hat{\mu}\|} \leq 4 \frac{\|X_i\| \, \|\hat{\mu}\|}{\|X_i\| \, \|X_{i+h}\|} = \frac{4 \, \|\hat{\mu}\|}{\|X_{i+h}\|}.$$

Combining the above arguments, it follows that

$$\frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n\}} \frac{1}{\hat{m}_{i,h}} \left| S_{N,i}^{(1,1)} \right| \leq \frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n\}} \frac{4}{\|X_{i+h}\|} \|\hat{\mu}\|$$

$$= 4 \, \|\hat{\mu}\| \, \frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n\}} \frac{1}{\|X_{i+h}\|} \leq 4 \, \|\hat{\mu}\| \, \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\|X_{i+h}\|}.$$

Since $\|\hat{\mu}\| \xrightarrow{a.s.} 0$ *and* $\frac{1}{N} \sum_{i=1}^{N} \|X_{i+h}\|^{-1} \xrightarrow{a.s.} E\|X_{i+h}\|^{-1} < \infty$ *by the ergodic Theorem, we have that*

$$\frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n\}} \frac{1}{\hat{m}_{i,h}} \left| S_{N,i}^{(1,1)} \right| \xrightarrow{a.s.} 0. \tag{B.12}$$

A similar argument shows that

$$\frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n\}} \frac{1}{\hat{m}_{i,h}} \left| S_{N,i}^{(1,2)} \right| \xrightarrow{a.s.} 0. \tag{B.13}$$

Combining (B.12) *and* (B.13)*, we get that*

$$\frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n\}} \frac{1}{\hat{m}_{i,h}} \left| S_{N,i}^{(1)} \right| \xrightarrow{a.s.} 0.$$

As for $S_{N,i}^{(2)}$*, using the Cauchy-Schwarz inequality, we have that*

$$\begin{aligned} \frac{1}{\hat{m}_{i,h}} \left| S_{N,i}^{(2)} \right| &= \frac{1}{\hat{m}_{i,h}} \left| \langle \hat{\mu} \, \|X_i\| \, , X_{i+h} \, \|X_{i+h}\| \rangle \right| \\ &\leq \frac{\|\hat{\mu}\| \, \|X_i\| \, \|X_{i+h}\|^2}{\|X_i\| \, \|X_{i+h}\| \, \|X_i - \hat{\mu}\| \, \|X_{i+h} - \hat{\mu}\|} \\ &= \frac{\|\hat{\mu}\| \, \|X_{i+h}\|}{\|X_i - \hat{\mu}\| \, \|X_{i+h} - \hat{\mu}\|}. \end{aligned}$$

For $(X_i, X_{i+h}) \in A_N$*, it follows that*

$$\frac{\|\hat{\mu}\| \, \|X_{i+h}\|}{\|X_i - \hat{\mu}\| \, \|X_{i+h} - \hat{\mu}\|} \leq \frac{4 \, \|\hat{\mu}\| \, \|X_{i+h}\|}{\|X_i\| \, \|X_{i+h}\|} = 4 \frac{\|\hat{\mu}\|}{\|X_i\|}.$$

Thus,

$$\frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n\}} \frac{1}{\hat{m}_{i,h}} \left| S_{N,i}^{(2)} \right| \leq \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\hat{m}_{i,h}} \left| S_{N,i}^{(2)} \right| \leq 4 \, \|\hat{\mu}\| \, \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\|X_i\|} \xrightarrow{a.s.} 0.$$

Nearly identical arguments give that

$$\frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n\}} \frac{1}{\hat{m}_{i,h}} \left| S_{N,i}^{(3)} \right| \xrightarrow{a.s.} 0.$$

Finally considering $S_{N,i}^{(4)}$*, we have that using the Cauchy-Schwarz inequality that*

$$\begin{aligned}
&\frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n\}} \left| S_{N,i}^{(4)} \right| = \frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n\}} \frac{1}{\hat{m}_{i,h}} \left| \langle \hat{\mu} \, \|X_i\| \, , \hat{\mu} \, \|X_{i+h}\| \rangle \right| \\
&\leq \frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n\}} \frac{\|\hat{\mu}\|^2 \, \|X_i\| \, \|X_{i+h}\|}{\|X_i\| \, \|X_{i+h}\| \, \|X_i - \hat{\mu}\| \, \|X_{i+h} - \hat{\mu}\|} \\
&\leq 4 \, \|\hat{\mu}\|^2 \, \frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n\}} \frac{1}{\|X_i\| \, \|X_{i+h}\|} \\
&\leq 4 \, \|\hat{\mu}\|^2 \, \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\|X_i\| \, \|X_{i+h}\|} \xrightarrow{a.s.} 0,
\end{aligned}$$

since $\|\hat{\mu}\|^2 \xrightarrow{a.s.} 0$*, and* $\frac{1}{N} \sum_{i=1}^{N} \|X_i\|^{-1} \, \|X_{i+h}\|^{-1} \xrightarrow{a.s.} E[\|X_i\|^{-1} \, \|X_{i+h}\|^{-1}] < \infty$ *by Assumption 3.2.4. Thus, we conclude that*

$$S_{N,i}^{(1)} = \frac{1}{N} \sum_{\{i:(X_i, X_{i+h}) \in A_n\}} \left| \langle S(X_i - \hat{\mu}), S(X_{i+h} - \hat{\mu}) \rangle - \langle S(X_i), S(X_{i+h}) \rangle \right| \xrightarrow{a.s.} 0.$$

Turning to $S_{N,2}$*. The argument here is similar to that presented in the proof of Theorem 1 of Boente et al. (2019), but we provide the details so that the proof is self-contained. First note that*

$$\|\langle S(X_i - \hat{\mu}), S(X_{i+h} - \hat{\mu}) \rangle\| = \|\langle S(X_i), S(X_{i+h} - \hat{\mu}) \rangle\| = 1. \tag{B.14}$$

Let $d_{N,x} = \{(x,y) \in \mathcal{H} \times \mathcal{H} : 2\|x - \hat{\mu}\| \geq \|x\|\}$, $d_{N,y} = \{(x,y) \in \mathcal{H} \times \mathcal{H} : 2\|y - \hat{\mu}\| \geq \|y\|\}$.

Recall $A_N = \{(x,y) \in \mathcal{H} \times \mathcal{H} : 2\|x - \hat{\mu}\| > \|x\| \text{ and } 2\|y - \hat{\mu}\| > \|y\|\}$. *Consequently* $A_N^c = \{(x,y) \in \mathcal{H} \times \mathcal{H} : 2\|x - \hat{\mu}\| \leq \|x\| \text{ or } 2\|y - \hat{\mu}\| \leq \|y\|\}$. *Clearly* $A_n^c = d_{N,x}^c \cup d_{n,y}^c$. *Consequently,*

$$\mathbb{1}_{A_N^c}(x,y) \leq \mathbb{1}_{d_{N,x}^c}(x,y) + \mathbb{1}_{d_{N,y}^c}(x,y). \tag{B.15}$$

Using the triangle inequality that

$$\begin{aligned}
S_{N,2} &= \frac{1}{N} \sum_{\{i:(X_i,X_{i+h}) \in A_n^c\}} \|\langle S(X_i - \hat{\mu}), S(X_{i+h} - \hat{\mu})\rangle - \langle S(X_i), S(X_{i+h})\rangle\| \\
&\leq \frac{2}{N} \sum_{\{i:(X_i,X_{i+h}) \in A_n^c\}} 1 = \frac{2}{N} \sum_{i=1}^{N} \mathbb{1}_{A_N^c}(X_i, X_{i+h}) \\
&\leq \frac{2}{N} \sum_{i=1}^{N} \left(\mathbb{1}_{d_{N,x}^c}(X_i, X_{i+h}) + \mathbb{1}_{d_{N,y}^c}(X_i, X_{i+h}) \right) \\
&= \frac{2}{N} \sum_{i=1}^{N} \mathbb{1}_{d_{N,x}^c}(X_i, X_{i+h}) + \frac{2}{N} \sum_{i=1}^{N} \mathbb{1}_{d_{N,y}^c}(X_i, X_{i+h}) \\
&=: J_{N,1} + J_{N,2}.
\end{aligned} \tag{B.16}$$

We aim to show that $J_{N,1} \overset{a.s.}{\to} 0$ *and* $J_{N,2} \overset{a.s.}{\to} 0$.

Note that the assumption $\mathbb{E}\|X\|^{-1} < \infty$ *implies that* $P(\|X\| = 0) = 0$. *Thus, by continuity of the probability measure, for every* $\epsilon > 0$, *there exists* $\delta > 0$, *such that* $P(\|X\| \in B_{\delta,x}) \leq \epsilon$, *where* $B_{\delta,x} = \{(X,Y) \in \mathcal{H} \times \mathcal{H} : \|X\| \leq \delta\}$. *Then*

$$\begin{aligned}
&\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{d_{N,x}^c}(X_i, X_{i+h}) \\
&\leq \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{B_{\delta,x}}(X_i, X_{i+h}) + \frac{1}{N} \sum_{i=1}^{N} \left(\mathbb{1}_{d_{N,x}^c}(X_i, X_{i+h}) - \mathbb{1}_{B_{\delta,x}}(X_i, X_{i+h}) \right)_+ \\
&=: J_{N,1}^{(1)} + J_{N,1}^{(2)},
\end{aligned}$$

where $a_+ = \max(a, 0)$. *Using the ergodic theorem,* $J_{N,1}^{(1)} \overset{a.s.}{\to} P(\|X\| \in B_{\delta,x}) \leq \epsilon$. *Since* ϵ *is arbitrary, we take* $\epsilon \downarrow 0^+$ *and obtained* $J_{N,1}^{(1)} \overset{a.s.}{\to} 0$. *As for the* $J_{N,2}$, *note that* $\{(X_i, X_{i+h}) \in$

$\mathcal{H} \times \mathcal{H} : \|\hat{\mu}\| \leq \delta/2\} \subset \{(X_i, X_{i+h}) \in \mathcal{H} \times \mathcal{H} : \mathbb{1}_{d^c_{N,x}}(X_i, X_{i+h}) - \mathbb{1}_{B_{\delta,x}}(X_i, X_{i+h}) = 0\}$. *Hence, using Assumption 3.2.3, there exists a null set* $\mathcal{F}$ *such that for* $\omega \notin \mathcal{F}$*, there exists an* $M \in \mathbb{Z}$ *such that for every integer* $N \geq M$*,* $\|\hat{\mu}\| \leq \delta/2$*. This implies*

$$J^{(2)}_{N,1} = \frac{2}{N} \sum_{i=1}^{N} \mathbb{1}_{d^c_{N,y}}(X_i, X_{i+h}) = 0.$$

Combining the arguments above, we obtain $J_{N,1} \overset{a.s.}{\to} 0$*. An identical argument shows also that* $J_{N,2} \overset{a.s.}{\to} 0$*. Therefore,* $S_{N,2} \overset{a.s.}{\to} 0$*, which completes the proof.*

B.2 Additional Simulation Results

B.2.1 Accuracy of Estimator $\|\hat{C}_P\|_2$

We examined the accuracy of variance estimator $\|\hat{C}_P\|_2$ by generating data for which the value $\|C_P\|_2$ may be computed explicitly. Consider the two-dimensional process

$$X_i(t) = \sqrt{2\lambda_1} Z_{1,i} \sin(2\pi t) + \sqrt{2\lambda_2} Z_{2,i} \cos(2\pi t),$$

where $Z_{j,i}$ are a family of independent and identically distributed $\mathcal{N}(0,1)$ random variables. Straightforward calculation shows that $\|C_P\|_2$ may be calculated based on the moments ratios of quadratic forms of normal random variables. Such moments are computed explicitly in Magnus (1986). It may be shown when $\lambda_1 = \lambda_2 = 1$, then covariance kernel of the projected data is

$$C_P(t,s) = \cos(2\pi t)\cos(2\pi s) + \sin(2\pi t)\sin(2\pi s) = \cos(2\pi(s-t)),$$

and hence in this case $\|C_P\|_2^2 = 1/2$. Similarly, when $\lambda_1 = 1, \lambda_2 = 2$, $\|C_P\|_2^2 = 9 - 6\sqrt{2}$.

In order to examine the effect of variance estimation on the coverage of the confidence intervals in (3.9), we considered four different ways for constructing the confidence interval: 1) Normal confidence interval with true variance; 2) t-confidence interval with true variance with degrees of freedom $N-1$; 3) normal confidence interval with estimated variance and 4) t-confidence interval with estimated variance with degrees of freedom $N-1$. The

empirical coverage rates for each of these intervals are reported in Table B.1. We observed that the difference between the four construction methods was quite small, with the t-confidence interval being relatively narrow which is a phenomenon that is widely known. With the coverage rates that are close to the nominal levels, it evidently supports the derived asymptotic distribution given in (3.5).

Table B.1: The comparison between the variance estimate in a finite-dimensional case: 1) $Z_{\alpha/2}\|C_P\|_2^2$ is normal confidence interval with true variance; 2) $t_{\alpha/2,N-1}\|C_P\|_2^2$ is t-confidence interval with true variance with degrees of freedom $N-1$; 3) $Z_{\alpha/2}\|\hat{C}_P\|_2^2$ is normal confidence interval with estimated variance; and 4) $t_{\alpha/2,N-1}\|\hat{C}_P\|_2^2$ is t-confidence interval with estimated variance with degrees of freedom $N-1$.

			$\lambda_1=1$ and $\lambda_2=2$						$\lambda_1=\lambda_2=1$					
N	α	h	$Z_{\alpha/2}\|C_P\|_2^2$	$t_{\alpha/2,N-1}\|C_P\|_2^2$	$Z_{\alpha/2}\|\hat{C}_P\|_2^2$	$t_{\alpha/2,N-1}\|\hat{C}_P\|_2^2$	N	α	h	$Z_{\alpha/2}\|C_P\|_2^2$	$t_{\alpha/2,N-1}\|C_P\|_2^2$	$Z_{\alpha/2}\|\hat{C}_P\|_2^2$	$t_{\alpha/2,N-1}\|\hat{C}_P\|_2^2$
100	0.01	1	0.008	0.006	0.007	0.006	100	0.01	1	0.007	0.006	0.007	0.006
		5	0.008	0.006	0.007	0.006			5	0.009	0.007	0.009	0.006
		10	0.009	0.008	0.008	0.008			10	0.008	0.007	0.007	0.006
	0.05	1	0.048	0.046	0.045	0.045		0.05	1	0.047	0.046	0.047	0.043
		5	0.049	0.046	0.047	0.044			5	0.048	0.045	0.047	0.045
		10	0.044	0.041	0.041	0.038			10	0.045	0.04	0.043	0.04
	0.10	1	0.09	0.087	0.09	0.088		0.10	1	0.094	0.093	0.094	0.09
		5	0.102	0.099	0.103	0.098			5	0.11	0.105	0.109	0.105
		10	0.088	0.084	0.086	0.084			10	0.094	0.091	0.092	0.088
250	0.01	1	0.011	0.011	0.011	0.011	250	0.01	1	0.008	0.008	0.008	0.008
		5	0.008	0.007	0.007	0.006			5	0.006	0.006	0.006	0.006
		10	0.009	0.008	0.008	0.006			10	0.007	0.007	0.007	0.007
	0.05	1	0.047	0.045	0.045	0.044		0.05	1	0.048	0.047	0.048	0.047
		5	0.050	0.048	0.050	0.050			5	0.045	0.045	0.045	0.044
		10	0.037	0.035	0.036	0.036			10	0.034	0.033	0.034	0.033
	0.10	1	0.101	0.100	0.101	0.101		0.10	1	0.102	0.100	0.100	0.009
		5	0.091	0.089	0.088	0.087			5	0.093	0.093	0.093	0.093
		10	0.084	0.083	0.083	0.082			10	0.076	0.075	0.074	0.074
500	0.01	1	0.009	0.009	0.009	0.009	500	0.01	1	0.008	0.008	0.008	0.008
		5	0.009	0.008	0.01	0.009			5	0.006	0.006	0.006	0.006
		10	0.010	0.01	0.011	0.011			10	0.007	0.007	0.007	0.007
	0.05	1	0.045	0.044	0.043	0.042		0.05	1	0.051	0.050	0.05	0.049
		5	0.047	0.046	0.047	0.047			5	0.045	0.044	0.045	0.044
		10	0.046	0.046	0.047	0.047			10	0.044	0.044	0.044	0.044
	0.10	1	0.111	0.111	0.111	0.111		0.10	1	0.104	0.104	0.104	0.104
		5	0.100	0.098	0.101	0.1			5	0.094	0.093	0.094	0.093
		10	0.093	0.093	0.093	0.093			10	0.091	0.089	0.09	0.089

B.3 Details of the Estimation of the FSAR model

The FSAR model was introduced in Horváth et al. (2020). The estimation method that we use is based on a least-squares principle and FPCA. The FSAR model is defined as

$$X_n(t) = \Phi_{\ell_1}\left(X_{n-\ell_1}\right)(t) + \cdots + \Phi_{\ell_k}\left(X_{n-\ell_k}\right)(t) + \varepsilon_n(t), \tag{B.17}$$

where each $\ell_i \in \{1, \ldots, N\}$ and $\ell_i \neq \ell_j$ for $i \neq j$. Assuming that the kernels ϕ_{ℓ_i} and the observations X_i's can be well approximated by the first p FPCs, we then let

$$\boldsymbol{X}_i = \left(\langle X_i, \hat{v}_1\rangle, \ldots, \langle X_i, \hat{v}_p\rangle\right)^\top \in \mathbb{R}^p.$$

and

$$\mathbf{X}_{L,p,i} = \begin{pmatrix} \boldsymbol{X}_{i-\ell_1} \\ \boldsymbol{X}_{i-\ell_2} \\ \vdots \\ \boldsymbol{X}_{i-\ell_k} \end{pmatrix} = \begin{pmatrix} \langle X_{i-\ell_1}, \hat{v}_1\rangle \\ \langle X_{i-\ell_1}, \hat{v}_2\rangle \\ \vdots \\ \langle X_{i-\ell_k}, \hat{v}_p\rangle \end{pmatrix} \in \mathbb{R}^{kp}$$

Then

$$\begin{aligned}
\boldsymbol{X}_{L,p} &= \begin{pmatrix} \mathbf{X}_{L,p,\ell_k+1}^\top & \mathbf{X}_{L,p,\ell_k+2}^\top & \cdots & \mathbf{X}_{L,p,N}^\top \end{pmatrix} \\
&= \begin{pmatrix} \langle X_{\ell_k+1-\ell_1}, \hat{v}_1\rangle & \langle X_{\ell_k+1-\ell_1}, \hat{v}_2\rangle & \cdots & \langle X_1, \hat{v}_1\rangle & \cdots & \langle X_1, \hat{v}_p\rangle \\ \langle X_{\ell_k+2-\ell_1}, \hat{v}_1\rangle & \langle X_{\ell_k+2-\ell_1}, \hat{v}_2\rangle & \cdots & \langle X_2, \hat{v}_1\rangle & \cdots & \langle X_2, \hat{v}_p\rangle \\ \vdots & & & & & \\ \langle X_{n-\ell_1}, \hat{v}_1\rangle & \langle X_{n-\ell_1}, \hat{v}_2\rangle & \cdots & \langle X_{n-\ell_k}, \hat{v}_1\rangle & \cdots & \langle X_{n-\ell_k}, \hat{v}_p\rangle \end{pmatrix} \\
&\in \mathbb{R}^{(n-\ell_k)\times kp}, \\
\boldsymbol{X}_{R,p} &= \begin{pmatrix} \boldsymbol{X}_{\ell_k+1}^\top \\ \vdots \\ \boldsymbol{X}_n^\top \end{pmatrix} \\
&= \begin{pmatrix} \langle X_{\ell_k+1}, \hat{v}_1\rangle & \langle X_{\ell_k+1}, \hat{v}_2\rangle & \ldots & \langle X_{\ell_k+1}, \hat{v}_p\rangle \\ \vdots & & & \\ \langle X_n, \hat{v}_1\rangle & \langle X_n, \hat{v}_2\rangle & \ldots & \langle X_n, \hat{v}_p\rangle \end{pmatrix} \in \mathbb{R}^{(n-\ell_k)\times p}.
\end{aligned}$$

The least-squares estimator is:

$$\widehat{\mathbf{\Phi}} = \left(\mathbf{X}_{L,p}^{\top}\mathbf{X}_{L,p}\right)^{-1}\mathbf{X}_{L,p}^{\top}\mathbf{X}_{R,p} =: \begin{pmatrix} \hat{\mathbf{\Phi}}_{\ell_1} \\ \vdots \\ \hat{\mathbf{\Phi}}_{\ell_k} \end{pmatrix} \in \mathbb{R}^{kp\times p}$$

Then the estimates of the kernel function $\psi_{\ell_i}(t,s)$ is

$$\hat{\phi}_{\ell_i}(t,s) \approx \sum_{j=1}^{p}\sum_{r=1}^{p} \hat{\mathbf{\Phi}}_{\ell_i}[j,r]\hat{v}_j(t)\hat{v}_r(s), \quad \hat{\mathbf{\Phi}}_i \in \mathbb{R}^{p\times p},$$

where the square bracket $[\cdot,\cdot]$ shows the indices of the matrix.
It may be re-written in the matrix form as

$$\hat{\phi}_{\ell_i}(t,s) \approx \begin{pmatrix}\hat{v}_1(t), & \hat{v}_2(t), & \cdots & \hat{v}_p(t)\end{pmatrix} \hat{\mathbf{\Phi}}_{\ell_i} \begin{pmatrix} \hat{v}_1(s) \\ \hat{v}_2(s) \\ \vdots \\ \hat{v}_p(s) \end{pmatrix}.$$

The forecasts are then made by using the estimated kernel as

$$\hat{X}_i(t) = \int \hat{\phi}_{\ell_1}(t,s)X_{i-\ell_1}(s)ds + \cdots + \int \hat{\phi}_{\ell_k}(t,s)X_{i-\ell_k}(s)ds.$$

Appendix C

Technical Details of Chapter 4

C.1 Closed-form Expression for $f_{X|N}(x \mid \overline{N})$

This section provides a step-by-step derivation of the closed-form expression for $f_{\boldsymbol{X}|\boldsymbol{N}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \mid \overline{\boldsymbol{N}})$, adapted from Krupskii and Genton (2017). Before demonstrating this result, let us recall some notations that will be used throughout this sections. Let K be the number of locations and T be the number of time points. We denote $\mathcal{T} = \{t_1, \cdots, t_T\}$ the (common) set of time points where the FD are observed. Let $\boldsymbol{N} := (N_1, N_2, \ldots, N_T)^\top$ denote the collection random observations of the Poisson process $N(t)$ at each time in the set $\mathcal{T}$, and we denote the realization of $\boldsymbol{N}$ by $\overline{\boldsymbol{N}} = (\overline{N}_1, \overline{N}_2, \ldots, \overline{N}_T)^\top$. At each time t_j for $1 \leq j \leq T$, we denote $\boldsymbol{x}_j = (x_{1,t_j} \cdots x_{K,t_j})^\top$ and $\boldsymbol{\beta}_j = (\beta_{1,t_j} \cdots \beta_{K,t_j})^\top$. Furthermore, we set $\boldsymbol{x} = (\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \cdots, \boldsymbol{x}_T^\top)^\top$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \cdots, \boldsymbol{\beta}_T^\top)^\top$.

Without loss of generality, we assume that $\mu(t) = 0$, otherwise one may obtains the same result after subtracting the mean process as $X'(t) = X(t) - \hat{\mu}(t)$, where $\hat{\mu}(t) = K^{-1} \sum_{k=1}^{K} X(\boldsymbol{s}_k; t)$. Under this assumption, we have $\boldsymbol{x} = \boldsymbol{w}$, where the latter is defined in (4.12).

The following theorem demonstrates the analytical representation of the conditional density function $f_{\boldsymbol{X}|\boldsymbol{N}}$ in (4.13). The proof of this theorem is motivated and adapted by Krupskii and Genton (2017).

Theorem C.1.1 *Suppose that the random variables ξ_i are i.i.d.* $\exp(1)$*-distributed, then the density function $f_{\boldsymbol{X}|\boldsymbol{N}}$ admits the following form:*

$$f_{\boldsymbol{X}|\boldsymbol{N}}(\boldsymbol{y}_1,\dots,\boldsymbol{y}_T \mid \overline{\boldsymbol{N}}) = (2\pi)^{(m-KT)/2}|\boldsymbol{\Sigma}|^{-1/2}|\boldsymbol{\Sigma}_*|^{1/2}\exp(C_0)\Phi_{\boldsymbol{\Sigma}_*}(\xi_1^*,\cdots,\xi_m^*),$$

where

$$\begin{aligned}&\boldsymbol{\Sigma}_*^{-1} = \boldsymbol{B}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{B}, \qquad \boldsymbol{\xi}^* = \boldsymbol{\Sigma}_*\left(-\mathbb{I}_m + \boldsymbol{B}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)\mathbb{1}_m,\\ &C_0 = -\frac{1}{2}\left((\boldsymbol{X}\mathbb{1}_m)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\mathbb{1}_m - (\boldsymbol{B}\boldsymbol{\xi}^*)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{\xi}^*\right).\end{aligned} \tag{C.1}$$

Proof C.1.1 (Proof of Theorem C.1.1.) *Let us begin the proof with the following alternative representation of vectors $\boldsymbol{x}$ and $\boldsymbol{\beta}$. To this end, consider the index set*

$$\mathcal{I}_i = \left\{1 \le j \le T \middle| \overline{N}_j = i\right\},$$

for $1 \le i \le m$, and denote $n_i \coloneqq card(\mathcal{I}_i)$. The vectors $\widetilde{\boldsymbol{\beta}}_i := (\boldsymbol{\beta}_{j_1}^\top\cdots\boldsymbol{\beta}_{j_{n_i}}^\top)^\top$ and $\widetilde{\boldsymbol{x}}_i := (\boldsymbol{x}_{j_1}^\top\cdots\boldsymbol{x}_{j_{n_i}}^\top)^\top$ are the collections of vectors $\boldsymbol{\beta}_j$'s and $\boldsymbol{x}_j$'s respectively, where all of the indices $j_1,\cdots,j_{n_i}$ belong to the index set $\mathcal{I}_i$. This leads to the following representation of vectors $\boldsymbol{\beta}$ and $\boldsymbol{x}$ as

$$\boldsymbol{\beta} = (\widetilde{\boldsymbol{\beta}}_1^\top,\cdots,\widetilde{\boldsymbol{\beta}}_m^\top)^\top \quad \text{and} \quad \boldsymbol{x} = (\widetilde{\boldsymbol{x}}_1^\top,\cdots,\widetilde{\boldsymbol{x}}_m^\top)^\top. \tag{C.2}$$

Applying the above identities yields the following results:

$$\begin{aligned}f_{\boldsymbol{X}|\boldsymbol{N}}(\boldsymbol{y}_1,\dots,\boldsymbol{y}_T \mid \overline{\boldsymbol{N}}) &= \int_0^\infty\cdots\int_0^\infty \Phi_{\boldsymbol{\Sigma}_Z}(\boldsymbol{y}_1,\dots,\boldsymbol{y}_T)\exp\left(-\sum_{k=1}^m\xi_k\right)d\xi_1\dots d\xi_m\\ &= \int_{\mathbb{R}_+^m}\Phi_{\boldsymbol{\Sigma}_Z}\left(\boldsymbol{x}_1-\xi_{\overline{N}_1}\boldsymbol{\beta}_1,\dots,\boldsymbol{x}_T-\xi_{\overline{N}_T}\boldsymbol{\beta}_T\right)\exp\left(-\sum_{k=1}^m\xi_k\right)d\xi_1\dots d\xi_m\\ &= \int_{\mathbb{R}_+^m}\Phi_{\boldsymbol{\Sigma}_Z}\left(\widetilde{\boldsymbol{x}}_1-\xi_1\widetilde{\boldsymbol{\beta}}_1,\dots,\widetilde{\boldsymbol{x}}_m-\xi_m\widetilde{\boldsymbol{\beta}}_m\right)\exp\left(-\sum_{k=1}^m\xi_k\right)d\xi_1\dots d\xi_m,\end{aligned}$$

where the last equality holds due to the representation (C.2) and the fact that $\xi_{\overline{N}_j} = i$ for $j \in \mathcal{I}_i$. Comparing the above identity with (4.13), it remains to show that

$$\begin{aligned}&-\frac{1}{2}\left((\widetilde{\boldsymbol{x}}_1-\xi_1\widetilde{\boldsymbol{\beta}}_1)^\top\cdots(\widetilde{\boldsymbol{x}}_m-\xi_m\widetilde{\boldsymbol{\beta}}_m)^\top\right)\boldsymbol{\Sigma}^{-1}\left((\widetilde{\boldsymbol{x}}_1-\xi_1\widetilde{\boldsymbol{\beta}}_1)^\top\cdots(\widetilde{\boldsymbol{x}}_m-\xi_m\widetilde{\boldsymbol{\beta}}_m)^\top\right)^\top\\ &= C_0 - \frac{1}{2}(\boldsymbol{\xi}-\boldsymbol{\xi}^*)^\top\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\xi}-\boldsymbol{\xi}^*) + \sum_{k=1}^m\xi_k,\end{aligned} \tag{C.3}$$

where $C_0, \boldsymbol{\xi}^*$ and Σ^* are of the corresponding forms in (C.1). Provided that the above equality holds, the result follows from

$$
\begin{aligned}
f_{\boldsymbol{X}|\boldsymbol{N}}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T \mid \overline{\boldsymbol{N}}) &= \int_{\mathbb{R}_+^m} \Phi_{\boldsymbol{\Sigma}_Z}\left(\widetilde{\boldsymbol{x}}_1 - \xi_1\widetilde{\boldsymbol{\beta}}_1, \ldots, \widetilde{\boldsymbol{x}}_m - \xi_m\widetilde{\boldsymbol{\beta}}_m\right) \exp\left(-\sum_{k=1}^m \xi_k\right) d\xi_1 \ldots d\xi_m \\
&= (2\pi)^{-KT/2}|\boldsymbol{\Sigma}|^{-1/2} \int_{\mathbb{R}_+^m} \exp\left\{C_0 - \frac{1}{2}(\boldsymbol{\xi} - \boldsymbol{\xi}^*)^\top \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\xi} - \boldsymbol{\xi}^*)\right\} d\xi_1 \cdots d\xi_m \\
&= \int_{\mathbb{R}_+^m} \frac{(2\pi)^{(m-KT)/2}|\boldsymbol{\Sigma}|^{-1/2}|\boldsymbol{\Sigma}_*|^{1/2}\exp(C_0)}{(2\pi)^{m/2}|\boldsymbol{\Sigma}_*|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\xi} - \boldsymbol{\xi}^*)^\top \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\xi} - \boldsymbol{\xi}^*)\right\} d\xi_1 \cdots d\xi_m \\
&= (2\pi)^{(m-KT)/2}|\boldsymbol{\Sigma}|^{-1/2}|\boldsymbol{\Sigma}_*|^{1/2}\exp(C_0)\Phi_{\boldsymbol{\Sigma}_*}(\xi_1^*, \ldots, \xi_m^*),
\end{aligned}
$$

where the final equality holds since

$$
\begin{aligned}
&\int_{\mathbb{R}_+^m} \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}_*|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\xi} - \boldsymbol{\xi}^*)^\top \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\xi} - \boldsymbol{\xi}^*)\right\} d\xi_1 \cdots d\xi_m \\
&= \int_0^\infty \cdots \int_0^\infty \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}_*|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\xi} - \boldsymbol{\xi}^*)^\top \boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\xi} - \boldsymbol{\xi}^*)\right\} d\xi_1 \cdots d\xi_m \\
&= \int_{s_m=-\xi_m^*}^\infty \cdots \int_{s_1=-\xi_1^*}^\infty \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}_*|^{1/2}} \exp\left\{-\frac{1}{2}\boldsymbol{s}^\top \boldsymbol{\Sigma}_*^{-1}\boldsymbol{s}\right\} ds_1 \cdots ds_m \\
&= \int_{s_m=-\xi_m^*}^\infty \cdots \int_{s_1=-\xi_1^*}^\infty \phi_{\boldsymbol{\Sigma}_*}(s_1, \ldots, s_m) ds_1 \ldots ds_m \\
&= 1 - \Phi_{\boldsymbol{\Sigma}_*}(-\xi_1^*, \ldots, -\xi_m^*) \\
&= \Phi_{\boldsymbol{\Sigma}_*}(\xi_1^*, \ldots, \xi_m^*).
\end{aligned}
$$

Now, it remains to show the equality in (C.3) holds. To this end, note that

$$
\boldsymbol{B}\boldsymbol{\xi} = \begin{pmatrix} \boldsymbol{\beta}_1^* & \boldsymbol{\beta}_2^* & \cdots & \boldsymbol{\beta}_m^* \end{pmatrix} \boldsymbol{\xi} = \begin{pmatrix} \widetilde{\boldsymbol{\beta}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \widetilde{\boldsymbol{\beta}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \widetilde{\boldsymbol{\beta}}_m \end{pmatrix} \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_m \end{pmatrix} = \begin{pmatrix} \xi_1\widetilde{\boldsymbol{\beta}}_1 \\ \vdots \\ \xi_m\widetilde{\boldsymbol{\beta}}_m \end{pmatrix}. \tag{C.4}
$$

Similarly, we have

$$\boldsymbol{X}\mathbb{1}_m = \begin{pmatrix} \boldsymbol{x}_1^* & \boldsymbol{x}_2^* & \cdots & \boldsymbol{x}_m^* \end{pmatrix} \mathbb{1}_m = \begin{pmatrix} \widetilde{\boldsymbol{x}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \widetilde{\boldsymbol{x}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \widetilde{\boldsymbol{x}}_m \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \widetilde{\boldsymbol{x}}_1 \\ \vdots \\ \widetilde{\boldsymbol{x}}_m \end{pmatrix}. \tag{C.5}$$

Thus, it follows from (C.4) that

$$-\frac{1}{2}\left(\boldsymbol{B}\boldsymbol{\xi}\right)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{\xi} = -\frac{1}{2}\begin{pmatrix} \xi_1\widetilde{\boldsymbol{\beta}}_1^\top & \cdots & \xi_m\widetilde{\boldsymbol{\beta}}_m^\top \end{pmatrix} \boldsymbol{\Sigma}^{-1} \begin{pmatrix} \xi_1\widetilde{\boldsymbol{\beta}}_1^\top & \cdots & \xi_m\widetilde{\boldsymbol{\beta}}_m^\top \end{pmatrix}^\top.$$

An analogous argument leads to

$$\begin{aligned} \boldsymbol{\xi}^\top\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\xi}^* &= \boldsymbol{\xi}^\top\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\Sigma}_*\left(-\mathbb{I}_m + \boldsymbol{B}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)\mathbb{1}_m \\ &= \boldsymbol{\xi}^\top\mathbb{I}_m\left(-\mathbb{I}_m + \boldsymbol{B}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)\mathbb{1}_m \\ &= -\boldsymbol{\xi}^\top\mathbb{I}_m\mathbb{1}_m + \boldsymbol{\xi}^\top\boldsymbol{B}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\mathbb{1}_m \\ &= -\sum_{k=1}^m \xi_k + \left(\boldsymbol{B}\boldsymbol{\xi}\right)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\mathbb{1}_m \\ &= -\sum_{k=1}^m \xi_k + \begin{pmatrix} \xi_1\widetilde{\boldsymbol{\beta}}_1^\top & \cdots & \xi_m\widetilde{\boldsymbol{\beta}}_m^\top \end{pmatrix} \boldsymbol{\Sigma}^{-1} \begin{pmatrix} \widetilde{\boldsymbol{x}}_1^\top & \cdots & \widetilde{\boldsymbol{x}}_m^\top \end{pmatrix}^\top. \end{aligned}$$

Finally, we have

$$\mathbb{1}_m^\top\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\mathbb{1}_m = \begin{pmatrix} \widetilde{\boldsymbol{x}}_1^\top & \widetilde{\boldsymbol{x}}_2^\top & \cdots & \widetilde{\boldsymbol{x}}_m^\top \end{pmatrix} \boldsymbol{\Sigma}^{-1} \begin{pmatrix} \widetilde{\boldsymbol{x}}_1^\top & \widetilde{\boldsymbol{x}}_2^\top & \cdots & \widetilde{\boldsymbol{x}}_m^\top \end{pmatrix}^\top.$$

Applying the above equalities gives

$$\begin{aligned} & C_0 + \sum_{k=1}^m \xi_k - \frac{1}{2}(\boldsymbol{\xi} - \boldsymbol{\xi}^*)^\top\boldsymbol{\Sigma}_*^{-1}(\boldsymbol{\xi} - \boldsymbol{\xi}^*) \\ &= C_0 + \sum_{k=1}^m \xi_k - \frac{1}{2}\left(\boldsymbol{B}\boldsymbol{\xi}\right)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{\xi} + \boldsymbol{\xi}^\top\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\xi}^* - \frac{1}{2}\left(\boldsymbol{B}\boldsymbol{\xi}^*\right)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{\xi}^* \\ &= -\frac{1}{2}\mathbb{1}_m^\top\boldsymbol{X}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\mathbb{1}_m + \sum_{k=1}^m \xi_k - \frac{1}{2}\left(\boldsymbol{B}\boldsymbol{\xi}\right)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{\xi} + \boldsymbol{\xi}^\top\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\xi}^* \\ &= -\frac{1}{2}\left((\widetilde{\boldsymbol{x}}_1 - \xi_1\widetilde{\boldsymbol{\beta}}_1)^\top \cdots (\widetilde{\boldsymbol{x}}_m - \xi_m\widetilde{\boldsymbol{\beta}}_m)^\top\right)\boldsymbol{\Sigma}^{-1}\left((\widetilde{\boldsymbol{x}}_1 - \xi_1\widetilde{\boldsymbol{\beta}}_1)^\top \cdots (\widetilde{\boldsymbol{x}}_m - \xi_m\widetilde{\boldsymbol{\beta}}_m)^\top\right)^\top. \end{aligned}$$

This shows that (C.3) holds, and hence, we complete the proof of this theorem.

C.2 Derivation of the Distribution and Density functions

This section shows the derivations of the marginal CDF and PDF for x. Similar calculations can be found in Krupskii et al. (2018) with the mean assumed to be zero.

We begin by showing the step-to-step calculation of the CDF.

$$\begin{aligned}
F_{kj}(x) &= \mathbb{P}(Y(\boldsymbol{s}_k; t_j) + \beta \mathcal{E}_{N(t)} \leq x) \\
&= \int_{-\infty}^{x} \mathbb{P}\left(\mathcal{E}_{N(t)} \leq \frac{x-y}{\beta}\right) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy \\
&= \int_{-\infty}^{x} \left(1 - e^{\frac{y-x}{\beta}}\right) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy \\
&= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy - \int_{-\infty}^{x} e^{-\frac{x}{\beta}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\frac{y^2-2\mu y+\mu^2}{2\sigma^2} - \frac{y}{\beta})} dy, \qquad (*)
\end{aligned}$$

where we obtain the last line by using integration by parts. We expand the exponent of the last term in the intergral,

$$\begin{aligned}
-\left(\frac{y^2-2\mu y+\mu^2}{2\sigma^2} - \frac{y}{\beta}\right) &= -\left(\frac{y^2-2\mu y+\mu^2}{2\sigma^2} - \frac{2y\sigma^2/\beta}{2\sigma^2}\right) \\
&= -\left(\frac{y^2 - 2(\mu+\frac{\sigma^2}{\beta})y + \mu^2 + (\mu+\frac{\sigma^2}{\beta})^2 - (\mu+\frac{\sigma^2}{\beta})^2}{2\sigma^2}\right) \\
&= -\frac{\left(y - (\mu+\frac{\sigma^2}{\beta})\right)^2}{2\sigma^2} + \frac{\mu^2 + \frac{2\sigma^2\mu}{\beta} + \frac{\sigma^4}{\beta^2} - \mu^2}{2\sigma^2} \\
&= -\frac{\left(y - (\mu+\frac{\sigma^2}{\beta})\right)^2}{2\sigma^2} + \frac{\mu}{\beta} + \frac{\sigma^2}{2\beta^2}.
\end{aligned}$$

Therefore we have

$$(*) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy - \int_{-\infty}^{x} e^{-\frac{x}{\beta}+\frac{\mu}{\beta}+\frac{\sigma^2}{2\beta^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(y-(\mu+\frac{\sigma^2}{\beta})\right)^2}{2\sigma^2}} dy$$

$$
\begin{aligned}
&= \mathbb{P}(\sigma Z + \mu \leq x) - \exp\left(\frac{-x+\mu}{\beta} + \frac{\sigma^2}{2\beta^2}\right) \mathbb{P}\left(\sigma Z + (\mu + \frac{\sigma^2}{\beta}) \leq x\right) \\
&= \mathbb{P}\left(Z \leq \frac{x-\mu}{\sigma}\right) - \exp\left(\frac{-x+\mu}{\beta} + \frac{\sigma^2}{2\beta^2}\right) \mathbb{P}\left(Z \leq \frac{x-(\mu+\frac{\sigma^2}{\beta})}{\sigma}\right) \\
&= \Phi\left(\frac{x-\mu}{\sigma}\right) - \exp\left(\frac{-x+\mu}{\beta} + \frac{\sigma^2}{2\beta^2}\right) \Phi\left(\frac{x-\mu-\sigma^2/\beta}{\sigma}\right) \\
&= \Phi\left(\frac{x-\mu}{\sigma}\right) - \exp\left(\frac{\sigma^2}{2\beta^2} - \frac{x-\mu}{\beta}\right) \Phi\left(\frac{x-\mu}{\sigma} - \frac{\sigma}{\beta}\right).
\end{aligned}
$$

We can obtain the marginal PDF by taking the derivative as $f_{kj}(x) = \frac{d}{dx} F_{kj}(x)$. By some elementary algebra, it yields

$$
\begin{aligned}
f_{kj}(x) = &\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) + \frac{1}{\beta}\exp\left(\frac{\sigma^2}{2\beta^2} - \frac{x-\mu}{\beta}\right) \Phi\left(\frac{x-\mu}{\sigma} - \frac{\sigma}{\beta}\right) \\
&- \exp\left(\frac{\sigma^2}{2\beta^2} - \frac{x-\mu}{\beta}\right) \phi\left(\frac{x-\mu}{\sigma} - \frac{\sigma}{\beta}\right) \frac{1}{\sigma}.
\end{aligned}
$$

Glossary

ACF Autocorrelation function 10

ARCH Autoregressive conditional heteroskedasticity 17

CIDR Cumulative intraday returns 86

CPV Cumulative percentage of total variance 7

FAR Functional autoregressive process 18

FDA Functional data analysis 1

FD Functional data 1

FPCA Functional principal component analysis 4

FPC Functional principal component 5

FSAR Functional seasonal autoregressive model 82

FTS Functional time series 1

GARCH Generalized autoregressive conditional heteroskedasticity 17

GLM Generalized linear models 40

HU Hyndman-Ullah 82

NBA National Basketball Association 35

PCA Principal component analysis 4

PC Principal component 5

SARIMA Seasonal autoregressive integrated moving average 15

SFD Spatial functional data 2

WN White noise 11

fACF Functional autocorrelation function 19

fSACF Functional spherical autocorrelation function 68

i.i.d. Independent and identically distributed 6