Likelihood Ratio Combination of Multiple Biomarkers and Change Point Detection in Functional Time Series

Zhiyuan Du

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Pang Du, Chair
Xinwei Deng
Inyoung Kim
Laura Sands

September 4, 2024
Blacksburg, Virginia

Keywords: Multiple biomarkers, Smoothing splines density estimation, Functional time series, Change point detection.

Likelihood Ratio Combination of Multiple Biomarkers and Change Point Detection in Functional Time Series

Zhiyuan Du

(ABSTRACT)

Utilizing multiple biomarkers in medical research is crucial for the diagnostic accuracy of detecting diseases. An optimal method for combining these biomarkers is essential to maximize the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC). The optimality of the likelihood ratio has been proven but the challenges persist in estimating the likelihood ratio, primarily on the estimation of multivariate density functions. In this study, we propose a non-parametric approach for estimating multivariate density functions by utilizing Smoothing Spline density estimation to approximate the full likelihood function for both diseased and non-diseased groups, which compose the likelihood ratio. Simulation results demonstrate the efficiency of our method compared to other biomarker combination techniques under various settings for generated biomarker values. Additionally, we apply the proposed method to a real-world study aimed at detecting childhood autism spectrum disorder (ASD), showcasing its practical relevance and potential for future applications in medical research.

Change point detection for functional time series has attracted considerable attention from researchers. Existing methods either rely on FPCA, which may perform poorly with complex data, or use bootstrap approaches in forms that fall short in effectively detecting diverse change functions. In our study, we propose a novel self-normalized test for functional time series implemented via a non-overlapping block bootstrap to circumvent reliance on FPCA. The SN factor ensures both monotonic power and adaptability for detecting diverse change

functions on complex data. We also demonstrate our test's robustness in detecting changes in the autocovariance operator. Simulation studies confirm the superior performance of our test across various settings, and real-world applications further illustrate its practical utility.

Likelihood Ratio Combination of Multiple Biomarkers and Change Point Detection in Functional Time Series

Zhiyuan Du

(GENERAL AUDIENCE ABSTRACT)

In medical research, it is crucial to accurately detect diseases and predict patient outcomes using multiple health indicators, also known as biomarkers. Combining these biomarkers effectively can significantly improve our ability to diagnose and treat various health conditions. However, finding the best way to combine these biomarkers has been a long-standing challenge. In this study, we propose a new, easy-to-understand method for combining multiple biomarkers using advanced estimation techniques. Our method takes into account various factors and provides a more accurate way to evaluate the combined information from different biomarkers. Through simulations, we demonstrated that our method performs better than other existing methods under a variety of scenarios. Furthermore, we applied our new method to a real-world study focusing on detecting childhood autism spectrum disorder (ASD), highlighting its practical value and potential for future applications in medical research.

Detecting changes in patterns over time, especially shifts in averages, has become an important focus in data analysis. Existing methods often rely on techniques that may not perform well with more complex data or are limited in the types of changes they can detect. In this study, we introduce a new approach that improves the accuracy of detecting changes in complex data patterns. Our method is flexible and can identify changes in both the mean and variation of the data over time. Through simulations, we demonstrate that this approach

is more accurate than current methods. Furthermore, we applied our method to real-world climate research data, illustrating its practical value.

Acknowledgments

First and foremost, I would like to express my deepest appreciation to my advisor, Prof. Pang Du, for his unwavering support throughout my Ph.D. journey. His patience, motivation, and extensive knowledge have been invaluable assets in my research and thesis-writing process. Prof. Du's guidance has been instrumental at every step of the way, and I couldn't have imagined a more suitable mentor for my doctoral studies.

Besides my advisor, I would like to thank Prof. Danping Liu and Prof. Aiyi Liu from the National Institutes of Health (NIH) for offering their brilliant idea of the Pseudo-likelihood Ratio and providing the Autism data.

Additionally, I would like to extend my heartfelt gratitude to the other members of my thesis committee: Prof. Xinwei Deng, Prof. Inyoung Kim, and Prof. Laura Sands. Their encouragement, thought-provoking feedback, and challenging questions have been invaluable in shaping my research. I am truly grateful for their contributions and support.

Furthermore, I would like to extend my appreciation to all the faculty and staff members of the Department of Statistics and the Center for Gerontology at Virginia Tech for their assistance and support. I am also grateful to my fellow students, who have not only provided help when needed but also offered encouragement and friendship during my time at the university.

Last but not least, I want to thank my parents and my wife. Their love and support have been the foundation of my life. They have provided countless spiritual support throughout my journey. I am eternally grateful for their unconditional love and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Biomarkers play an important role in diagnostic medicine. In medical research, biomarkers are objective medical signs used to measure the presence or progress of the disease, or the effects of treatments. For example, several types of cancer cells show traces of Carcinoembryonic antigen (CEA) on their surfaces, and high levels of CEA detected in the blood is a possible symptom of colon rectal cancers CEA [1]. In real practice, there may be many sources of information in the detection to help us predict the appearance of the disease. Making good use of them could unsurprisingly improve prediction accuracy.

Nowadays, advances in technology have led to the availability of high-resolution functional time series data. Functions observed consecutively over time, such as daily or yearly observations, are commonly in fields like econometrics and climate research, examples including intraday financial curves and temperature trends. Detecting changes in these curves is crucial, as it can have a direct impact on our daily lives, and accurate detection can lead to better decision-making and improved outcomes.

1.1 Motivation

1.1.1 Medical diagnosis with multiple biomarkers

In the context of public health, the ability of a biomarker to designate a patient correctly with the diseased status is the True Positive Rate (TPR), also called sensitivity. The specificity, also known as the True Negative Rate (TNR), is the ability of a biomarker to designate a subject without disease as negative, and the False Positive Rate (FPR) is one minus the specificity. Commonly, a highly accurate biomarker requires a high TPR and a low FPR, so that most diseased patients are detected and treated while nondiseased patients are exempted from unnecessary treatments. For instance, for a biomarker on a continuous scale with higher values indicating the disease, values higher than a decision threshold c, are considered positive results.

Binary classifications of subjects through their biomarker measures shows the accuracy of the biomarker. For example, many clinical trials are implemented to check the performance of a specific biomarker in diagnosing a disease. Usually, the classification accuracy is measured in terms of the Receiver Operating Characteristic (ROC) curve, which plots the TPR versus the FPR for all possible values of the threshold. The Area under the ROC Curve (AUC) is a critical measure as a summary of the ROC curve, or a performance metric measuring the ability of a biomarker to distinguish between the two classes. In public health research, more biomarkers mean more potential resources looking for signs of a disease. Thus, a sound method of combining multiple biomarkers can really improve the diagnostic accuracy.

A common way to compare the biomarkers is to compare their TPRs at a fixed FPR f_0, such that the one with the highest TPR is the optimal biomarker at f_0. However, biomarker may be optimal at some FPRs but not at others. Suppose we have p biomarkers and the

results of the pth biomarker is denoted as $x_j, j = 1, ..., p$, the vector of biomarkers' results as $x = (x_1, ..., x_p)$. Let D be a binary variable denoting the disease status with $D = 1$ for positive results and $D = 0$ for negative results. We say a rule combining the x as a combing rule so as to classify the patients as diseased or nondiseased. By the Neyman Peasrson Lemma, the likelihood rule based on x would provide the highest TPR among all possible rules based on x for any f_0. That is to say, a patient is diagnosed as positive whenever

$$LR(x) > c(f_0),$$

where $LR(x) = f(x)/g(x)$ is the likelihood ratio with $f(x)$ and $g(x)$ being respectively the likelihoods when $D = 1$ and $D = 0$, and $c(f_0)$ is the value so that $f_0 = P(LR(x) > c(f_0)|D = 0)$. We may be familiar with these results in the context of statistical hypothesis testing [33] and the analogy was explicitly described by McIntosh and Pepe [31]. The Neyman-Pearson results state that the likelihood ratio rule is the Uniformly Most Powerful (UMP) test achieving the highest statistical power among all tests with the same Type I error, which is the FPR concerning the classification rule in our case. Accordingly, a rule based on the likelihood ratio $LR(X)$ exceeding a threshold can achieve the highest TPR among all possible combing rules based on X with FPR=f_0. Given the optimality of the likelihood ratio, the question then reduces to how to best estimate it in practice. With this goal, we introduce a nonparametric approach based on the smoothing spline density estimation of the likelihoods for the diseased and nondiseased groups. Its non-reliance on any specific distributional assumption addresses the difficulty mentioned by McIntosh and Pepe [35].

1.1.2 Change point testing in functional time series

In the context of change point testing in scalar time series, Shao [42] introduced an innovative method to bypass the challenges of estimating the long-run variance, which motivated us to extend this approach into the functional data setting. Traditional long-run variance estimation often depends on the selection of a bandwidth parameter, as discussed by Vogelsang [47] and Deng and Perron [11], who noted that no optimal selection exists to resolve the non-monotonic power issue. To overcome this, Shao [42] proposed a self-normalization technique to avoid estimating nuisance parameters and empirically demonstrated that it results in monotonic power. Given a time series $\{X_t\}_{t=1}^n$, they define the CUSUM process $T_n(k) = n^{-1/2}\sum_{t=1}^k (X_t - \bar{X}_n)$, where $\bar{X}_n = n^{-1}\sum_{t=1}^n X_t$, and the normalization process $V_n(\cdot)$

$$V_n(k) = n^{-2}\left[\sum_{t=1}^{k}\{S_{1,t} - (t/k)S_{1,k}\}^2 + \sum_{t=k+1}^{n}\{S_{t,n} - (n-t+1)/(n-k)S_{k+1,n}\}^2\right], \quad k = 1, \ldots, n-1,$$

where $S_{t_1,t_2} = \sum_{j=t_1}^{t_2} X_j$ if $t_1 \leq t_2$ and 0 otherwise. The self-normalization test statistic is then defined as $G_n = \sup_{1\leq k\leq n-1} T_n(k)' V_n^{-1}(k) T_n(k)$, whose null distribution can be easily derived. We shall extend this method to the functional data setting by applying it to a functional time series, preserving its beneficial properties.

1.2 Literature Review

1.2.1 Multiple Biomarkers Combination

The most straightforward way to combine multiple biomarkers is a linear combination. One of the first papers that considered the linear combination of the biomarkers was Su and Liu [46] who used the normal linear discriminant analysis to combine the biomarkers. But the method is based on the solid assumption of multivariate normality. They also studied the optimality of ROC curves by maximizing the AUC. Pepe et al. [36] then proved the robustness of maximizing AUC as the criterion. Now AUC maximization has become the common criterion for evaluating new approaches to combine multiple biomarkers. Pepe and Thompson [35] proposed to obtain an empirical solution to the optimal linear combination that maximizes the Mann–Whitney statistic, an empirical estimate of the AUC. This procedure is distribution-free and thus robust against distributional assumptions. However, when the number of biomarkers is relatively large, this empirical optimization procedure can be computationally intensive. Liu et al. [27] proved the optimality of combining the minimum and maximum values of the biomarkers but this method is unstable since it does not consider all information about the biomarkers, and standardization is needed. The standardization may be difficult especially when data are skewed, and inappropriate standardization may significantly deteriorate the efficiency. McIntosh and Pepe [31] showed the optimality of the likelihood ratio combination based on the Neyman-Pearson Lemma but did not provide any specific combination method.

1.2.2 Probability Density Estimation

Probability density estimation has long been discussed in theory and applied statistics. The classical parametric way is to estimate the unknown parameters when we have prior knowledge about the parametric family that f belongs to. However, nonparametric methods must be used when we have little information about f. According to Good and Gaskins [15], the estimation of f lies in the minimization of a penalized likelihood score $L(f) + \lambda J(f)$. $L(f)$ is usually taken as the log-likelihood, $J(f)$ is the roughness term and λ is the smoothing parameter controlling the trade-off. Good and Gaskin's idea is to set $J(f)$ as a quadratic function in $\sqrt{f}$, which makes up the positivity constraint but leaves the unity constraint to the numeric problem.

The penalized likelihood approach was pioneered by Good and Gaskins [15]. The success of penalty smoothing (smoothing spline) has been proved as one of the most successful multivariate methods available by Gu and Wahba [3]. Unlike the regression splines, two intrinsic constraints must be enforced for density estimation which are the positivity constraint and the unity constraint. Then Leonard [26] proposed the logistic density transform $f = e^g / \int e^g$ and proved that g is satisfying both constraints but a many-to-one feature of transform in usual function spaces introduces extra theoretical and computational inconvenience. Silverman [45] proposed to estimate $g = log\ f$ which is free of the positivity constraint and to argue the penalized likelihood score by a term $\int e^g$ to effectively enforce the unity constraint when $J(f)$ is a quadratic function in derivatives of $g = log\ f$. Based on Silverman's idea, O'Sullivan developed an algorithm to calculate Silverman's estimator using B-spline approximations in one dimension. Then Klonias [22] established the convergence rate for the Good-Gaskins $\sqrt{f}$-based penalties and Silverman established the convergence rate for the Leonard-Silverman log f-based penalties. Gu and Qiu [18] proposed a simple surgery on the usual function spaces to make the logistic density transform one-to-one, developed

an asymptotic theory in parallel to that of Silverman [45], and proposed and justified an adaptive semiparametric estimator, which paved the way for the current development. The development of Gu and Qiu [18] is in a dimensionless generic setup, and hence the current development is dimensionless.

The choice of smoothing parameter λ has a major impact on the performance of the resulting estimate. In the context of kernel density estimation, various cross-validation schemes have been developed to automatically select the smoothing parameter there. Wahba [48] developed a generalized cross-validation method for use in the context of orthogonal series density estimation. O'Sullivan [34] adapted a certain cross-validation score in kernel method literature to choose λ in the calculation of Silverman's [45] estimator. Gu [16] then developed a dimensionless automatic algorithm that updated g and λ jointly which is also a self-voting generalized cross-validation method.

1.2.3 Change Point Analysis

Change point analysis for scalar time series has a rich literature, including two key works by Csörgö and Horváth [9] and Perron et al. [37]. In mean shift testing, obtaining a consistent long-run variance (LRV) estimate that accounts for temporal dependence is crucial for parameter-free asymptotic analysis under the null hypothesis. Despite extensive research on bandwidth selection for LRV, challenges like the non-monotonic power issue highlighted by Vogelsang [47] and Deng and Perron [11] persist. To address this, Shao and Zhang [43] proposed a self-normalization (SN) based test that considers dependence and ensures monotonic power.

Extensive research on change point testing for functional data has primarily focused on CUSUM-based methods, the most widely used approach for functional time series. Berkes

et al. [2] introduced a CUSUM-based test for detecting mean function changes in independent functional data, but it did not account for temporal dependence. Noticing its limitation of not accounting for temporal dependence, Hörmann and Kokoszka [19] incorporated a consistent LRV estimator, later extended by Horváth et al. [21]. However, both approaches face challenges with bandwidth or tuning parameter selection, which can result in non-monotonic power. Zhang et al. [52] incorporated the self-normalization matrix into the functional context using the CUSUM process built up on empirical score vectors from the functional principle component analysis (FPCA).

1.2.4 Bootstrapped Change Point Tests

For tests not relying on FPCA, bootstrap methods offer a practical solution, especially when dealing with the inestimable infinite-dimensional covariance operator in their asymptotic distributions. Among these methods, the block bootstrap is particularly effective for time series data. Its utility in both independent and dependent settings is extensively studied by Lahiri and Lahiri [24]. Carlstein [6] explored the non-overlapping block bootstrap, showing its asymptotic convergence, while Sharipov et al. [44] extended it to change-point detection in Hilbert space-valued random variables. The dependent wild bootstrap, initially introduced by Shao [41], has also been adapted for functional time series by researchers like Bucchia and Wendler [4] and Wegner and Wendler [49].

1.3 Smoothing Spline Density Estimation

Since the probability densities to be estimated here are all multivariate ones, we now introduce the smoothing spline density estimation method on a multivariate domain.

We first construct a tensor product reproducing kernel Hilbert space by incorporating the ANOVA decomposition of a multivariate function. Every non-negative definite function R on a domain $\mathcal{X}$ corresponds to a reproducing kernel Hilbert space with R as its reproducing kernel. Given $\mathcal{H}_{\langle 1\rangle}$ on $\mathcal{X}_1$ with reproducing kernel $R_{\langle 1\rangle}$ and $\mathcal{H}_{\langle 2\rangle}$ on $\mathcal{X}_2$ with reproducing kernel $R_{\langle 2\rangle}$, $R = R_{\langle 1\rangle}R_{\langle 2\rangle}$ is non-negative definite on $\mathcal{X}_1 \times \mathcal{X}_2$.The reproducing kernel Hilbert space corresponding to such an R is called the tensor product space of $\mathcal{H}_{\langle 1\rangle}$ and $\mathcal{H}_{\langle 2\rangle}$, and is denoted by $\mathcal{H}_{\langle 1\rangle} \otimes \mathcal{H}_{\langle 2\rangle}$. In general, a multiple-term reproducing kernel Hilbert space can be written as $\mathcal{H} = \oplus_\beta \mathcal{H}_\beta$, with subspaces $\mathcal{H}_\beta$ having inner products $(f_\beta, g_\beta)_\beta$ and reproducing kernels R_β. Allowing for inter-module rescaling of the metrics, an inner product in $\mathcal{H}$ can be specified via

$$J(f, g) = \sum_\beta \theta_\beta^{-1}(f_\beta, g_\beta)_\beta \tag{1.1}$$

where $\theta_\beta \in (0, \infty)$ are tunable parameters. The reproducing kernel associated with (1.1) is $R_J = \sum_\beta \theta_\beta R_\beta$.

Example 1.3.1 (Tensor Product RKHS on $[0,1]^2$) Consider two continuous variables on $\mathcal{X} = [0, 1]^2$. The RHKS $\mathcal{H}_{\nu,\mu} = \mathcal{H}_{\nu\langle 1\rangle} \otimes \mathcal{H}_{\mu\langle 2\rangle}, \nu, \mu = 00, 01, 1$, with inner products $(f, g)_{\nu,\mu}$ and reproducing kernels $R_{\nu,\mu} = R_{\nu\langle 1\rangle}R_{\mu\langle 2\rangle}$. One may set

$$J(f, g) = \theta_{1,00}^{-1}(f, g)_{1,00} + \theta_{1,01}^{-1}(f, g)_{1,01} + \theta_{00,1}^{-1}(f, g)_{00,1} + \theta_{01,1}^{-1}(f, g)_{01,1} + \theta_{1,1}^{-1}(f, g)_{1,1}$$

and the associated reproducing kernel is

$$R_J = \theta_{1,00}R_{1,00} + \theta_{1,01}R_{1,01} + \theta_{00,1}R_{00,1} + \theta_{01,1}R_{01,1} + \theta_{1,1}R_{1,1}$$

For the tensor product cubic spline, the corresponding reproducing kernel is $R_{00}(x_{\langle 1\rangle}, x_{\langle 2\rangle}) = 1$, $R_{01}(x_{\langle 1\rangle}, x_{\langle 2\rangle}) = k_1(x_{\langle 1\rangle})k_1(x_{\langle 2\rangle})$, and $R_1(x_{\langle 1\rangle}, x_{\langle 2\rangle}) = k_2(x_{\langle 1\rangle})k_2(x_{\langle 2\rangle}) - k_4(x_{\langle 1\rangle} - x_{\langle 2\rangle})$, the

details and examples of the reproducing kernel could be found in Sec 2.3 of Gu [17]

1.4 Functional Autoregressive Model

In functional time series, various types of data can be considered. Independent functional observations are usually treated as realizations of the standard Brownian Motion (BM) or Brownian Bridge (BG), while the Functional Autoregressive process of a specified order is always used to simulate the data with temporal dependence. The functional autoregressive sequences $\{X_t\}_{t=1}^n$ of order 1 are defined by the equation

$$X_t(s) = \int_0^1 \psi(s,u)X_{t-1}(u)du + \epsilon_t(s), \quad 0 \leq s \leq 1, \quad t = 1, ..., n, \tag{1.2}$$

where $\psi(s,u)$ represents the kernel function, and ϵ_t are functional innovation sequences that can be independent Brownian Motions or Brownian Bridges in the interval $[0,1]$. To ensure the FAR(1) process remains stationary, the operator norm associated with ψ has to maintain a value less than one, that is, $\|\psi\|_{HS}^2 = \int_0^1 \int_0^1 \psi^2(s,u)dsdu < 1$, where $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm. Two kernel functions are introduced by Gabrys and Kokoszka [14]. They are, the Gaussian kernel $\psi(s,u) = C\ \exp(\frac{s^2+u^2}{2})$ and the Wiener kernel $\psi(s,u) = C\ \min(s,u)$, where the norm $\|\psi\|_{HS}$ scales differently with the change of C.

1.5 New Challenges

The combination of biomarkers has been a long-standing topic of interest in medical research. Historically, the focus has been on linear combinations based on the AUC criterion, until the introduction of the likelihood ratio rule, which involves the ratio of the densities of the dis-

eased and non-diseased groups. Existing approaches often assume specific forms of densities, which can reduce the efficiency of classification. This dissertation aims to estimate multivariate probability densities in a nonparametric manner without parametric assumptions on the distributions of the biomarkers. In Chapter 2, we propose using smoothing spline density estimation to estimate densities for the diseased and non-diseased groups separately. These estimates are then used to compute the likelihood ratio. Specifically, each density is estimated by smoothing splines to minimize the corresponding penalized likelihood. A pseudo-likelihood form is also introduced to handle cases with a high number of biomarkers. Simulations and real-world data applications demonstrate the excellent performance of our proposed method compared to other approaches.

Detecting change points in functional time series has gained significant attention recently, owing to the increased availability of such data in modern studies. The self-normalization test has found great successes in scalar time series. In Chapter 3, we propose a novel test building upon the cumulative sum (CUSUM) approach. Our method offers the advantages of monotonic power in detecting change points by incorporating the self-normalization. Additionally, we extend it to identify change points in the lag-1 autocovariance operator. Simulation studies illustrate the superiority of our method in detecting several types of changes across various types of complex data and the real-world applications demonstrate its practical utility.

Chapter 2

Combination of Multiple Biomarkers via Likelihood Ratio with Smoothing Spline Estimated Densities

2.1 Introduction

Biomarkers are crucial for medical professionals and researchers to diagnose diseases. In medical diagnosis, patients are typically classified as diseased if their biomarker values exceed a threshold, which can be seen as a statistical hypothesis test where the False Positive Rate (FPR) represents the Type I error and the True Positive Rate (TPR) represents the statistical power. With rapid advancements in medical techniques, there are now numerous biomarkers whose combined information is crucial for disease diagnosis. Hence, finding the most effective way to combine these biomarkers is a major area of focus for researchers in diagnostic medicine.

The utilization of linear combinations has been widely acknowledged as a prevalent method for the combination of multiple biomarkers. The pioneering study in this area was conducted by Su and Liu [46], who considered a multivariate normal distribution and employed normal linear discriminant analysis for the combination of the biomarkers. Their proposed approach

seeks to determine the optimal linear combination by maximizing the area under the receiver operating characteristic (ROC) curve under the framework of multivariate normal distributions. Pepe and Thompson [35] introduced a distribution-free approach for the combination of multiple biomarkers, in which the empirical linear combination of the biomarkers can attain the optimal empirical area under the curve (AUC) without assuming multinormality. However, the computational burden increases dramatically as the grid of coefficients expands with the addition of more biomarkers. Chunling [27] showed that the optimal combination can be achieved by combining only the minimum and maximum of the biomarkers, but this approach is prone to instability, as it does not take into account the information of all the biomarkers. Standardization is required but can be difficult. When data are skewed, incorrect standardization can significantly impact efficiency. By building upon the linear combination principles discussed above, Yin and Tian [51] utilized a linear combination of biomarkers with an emphasis on optimizing the Youden index when determining an ideal cut-off point. Similarly, Yan et al. [50] directed their focus towards the linear combination of biomarkers in order to elevate the partial AUC (pAUC), especially in contexts where only a limited spectrum of FPR (or TPR) is of clinical significance. However, in their adoption of the aforementioned concepts, both studies encountered the inherent limitations of the aforementioned methodologies.

McIntosh and Pepe [31] studied the optimality of the likelihood ratio combination, which is the ratio of the multivariate density of diseased individuals to the multivariate density of non-diseased individuals, based on the decision theory of the Neyman-Pearson Lemma. They showed that [33] both the likelihood ratio and the risk score, as a transformation of the likelihood ratio, are optimal in the sense of the Neyman-Pearson theory. That is, the likelihood ratio combination achieves the highest TPR among all the combinations with the same Type I error. In other words, this ensures point-wise optimality across a spectrum

of indices derived from the ROC curve, including but not limited to metrics such as AUC, Youden index, and pAUC. To illustrate the robustness and versatility of our methodology, we focus on AUC as a representative index in the ensuing discussion. Drawing upon the principles of the likelihood ratio, Qin and Zhang [38] proposed a semiparametric approach for estimating the exponential tilting model of the density ratio. Chen et al. [7] improved the model by considering a semiparametric monotonic density ratio model. However, the parametric component of the density ratio can be difficult to specify when incorporating a large number of biomarkers, as polynomials and interaction terms must be included. Liu et al. [28] proposed approaches for estimating the density ratio by decomposing the multivariate likelihood ratio statistic in various ways, including estimating the marginal densities of the biomarkers, the conditional densities given the relationships among the biomarkers, and a mixture of both. They argue that the multivariate density can be expressed as the product of the marginal densities under the assumption of independence among the biomarkers, and those conditional densities can be estimated if the relationships among the biomarkers are known. However, in practice, it is challenging to specify the relationships among the biomarkers. All of these approaches, to some extent, require a complete or partial parametric modeling or an independence assumption of the biomarkers, suggesting that additional work is required.

The optimality of the likelihood combination has been given theoretically, the difficulty lies in finding an appropriate method for estimating the multivariate density functions of the diseased and non-diseased groups. To address this challenge, we present a novel, nonparametric approach for combining multiple biomarkers. Our approach employs smoothing splines to estimate the probability densities for the diseased and non-diseased groups, the estimated densities are then assembled to compute the likelihood ratio. Unlike previous methods, our approach does not impose any parametric assumptions on the distributions of the biomarkers.

For the estimation of each density function, we first apply a logit transformation to generate a constraint-free function. Then the transformed function is estimated by smoothing splines through the optimization of a penalized likelihood, consisting of three parts: the negative log-likelihood representing the goodness-of-fit, a roughness penalty to enforce a certain level of smoothness on the estimate, and a smoothing parameter to balance the trade-off. This optimization process can be performed using the standard Newton-Raphson procedure.

Our approach is validated by its exceptional performance with large AUCs in both simulation studies and real-world examples. To address the challenge of combining a high number of biomarkers, we present a novel method known as the Pseudo-Likelihood ratio approach. Our method first performs variable selection through the SpAM algorithm [40] to identify biomarkers that are strongly correlated with the patient's status, which are the least squares minimizers of the additive model. Subsequently, we employ the Ramp AUC (RAUC) method proposed by Fong et al. [13] to estimate the coefficients of the Pseudo-Likelihood Ratio, effectively mitigating the curse of dimensionality. RAUC leads to a consistent and asymptotically normal estimator of the linear marker combination and outperforms other linear combination methods such as the smoothed AUC method [30].

The rest of the paper is organized as follows. In section 2, we provide the details of our approach, including the smoothing spline density estimation and Pseudo-Likelihood ratio for dealing with a high number of biomarkers. Section 3 presents the results of our simulation studies, comparing the performance of our method with other commonly used biomarker combination methods. In section 4, we apply our method to the childhood autism/ASD data, and provide a conclusion and discussion in section 5.

2.2 Methodology

2.2.1 Notation and Model

Suppose there are p biomarkers and n independent observations $x_i = (x_{i1}, x_{i2}, ..., x_{ip}), i = 1, ...n$. Without loss of generality, assume that the first n_0 observations $x_i, i = 1, ..., n_0$, are from the diseased group and follow a common distribution with the density function $f(x)$, and that the remaining $n - n_0$ observations $x_i, i = n_0 + 1, ..., n$, are from the non-diseased group and follow a common distribution with the probability density function $g(x)$. According to McIntosh and Pepe [31], the likelihood rule would provide the optimal diagnostic accuracy if the biomarkers are combined in the form of the full likelihood ratio:

$$LR(x_i) = \frac{f(x_i)}{g(x_i)} \tag{2.1}$$

or equivalently, its logarithm transformation. In this paper, we propose a nonparametric approach to estimating the log-likelihood ratio through smoothing spline modeling of the two density functions [17]. That is, we don't assume the density functions $f(x)$ or $g(x)$ belong to any parametric families except that they are both smooth functions.

Here we only describe the procedure for estimating $f(x)$ with details, with the notion that $g(x)$ can be estimated in a similar way. For a function $f(x)$ to be a probability density function it must satisfy that $f(x) > 0$ for any x on its supporting domain $\mathcal{X}$ and $\int_{\mathcal{X}} f(x)dx = 1$. Therefore, our first step in estimation is to transform the density function into a constraint-free function. Consider the logistic density transformation $f(x) = \frac{e^{\eta(x)}}{\int_{\mathcal{X}} e^{\eta(x)} dx}$. Note that replacing $\eta(x)$ by $\eta(x) + c$ for any constant c would also work for the transformation. We need an additional constraint on η, say, $\int_{\mathcal{X}} \eta(x)dx = 0$, to make it uniquely associated with

$f(x)$. Then we will estimate $\eta(x)$ by the minimizer of the penalized likelihood (PL)

$$-\frac{1}{n_0}\sum_{i=1}^{n_0}\eta(x_i)+\log\int_{\mathcal{X}}e^{\eta(x)}dx+\frac{\lambda}{2}J(\eta) \tag{2.2}$$

on a reproducing kernel Hilbert space $\mathcal{H}$ to be introduced in the next section. In the PL, the first two terms form the negative log-likelihood $L(\eta)$, which measures the goodness-of-fit of η, $J(\eta)$ is the roughness penalty enforcing a certain level of on η, and $\lambda(>0)$ is the smoothing parameter balances the trade-off.

2.2.2 Smoothing Spline Density Estimation

In this section, we describe the details of the estimation of η through the minimization of the PL (2.2). The minimization is performed on a reproducing kernel Hilbert space (RKHS) of functions. An RKHS is a Hilbert space $\mathcal{H}$ where the evaluation functional $[x]:\mathcal{H}\to\mathbb{R},\eta\mapsto\eta(x)$ is continuous for every $x\in\mathcal{X}$. Each RKHS is uniquely associated with a reproducing kernel $R:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ such that $R(x,u)=\langle R_x,R_u\rangle$, where R_x is the representer of the evaluation functional $[x]$ guaranteed by the Riesz Representation Theorem. It possesses the so-called reproducing property $\langle R_x,\eta\rangle=[x](\eta)=\eta(x)$, where $\langle\cdot,\cdot\rangle$ is the inner product on $\mathcal{H}$. The penalty functional J in (2.2) is a squared semi-norm on $\mathcal{H}$. Its null space $\mathcal{H}_0=\{f:J(f)=0\}$ induces a direct sum decomposition $\mathcal{H}=\mathcal{H}_0\oplus\mathcal{H}_J$, with $\mathcal{H}_J$ being the complement of $\mathcal{H}_0$ in $\mathcal{H}$. Correspondingly, the reproducing kernel decomposes as $R=R_0+R_J$, where R_0 and R_J are respectively the reproducing kernels on the subspaces $\mathcal{N}_J$ and $\mathcal{H}_J$. See, e.g., Gu [17, Chapter 2] for more details on RKHSs.

For p biomarkers in the product domain $\mathcal{X}=\mathcal{X}_1\times...\times\mathcal{X}_p$, we consider the RKHS of tensor product smoothing splines on $\mathcal{X}$, which is formed as the tensor product of marginal RKHS on $\mathcal{X}_j, j=1,\ldots,p$. We now give an example of such RKHS configuration for $p=2$ and

$\mathcal{X}_1 = \mathcal{X}_2 = [0,1]$, The common domain $[0,1]$ is used here only for notation simplification, with the notion that a configuration for general domains $[a_j, b_j]$ can be defined similarly after a simple normalization of variables.

Example 2.2.1 (Tensor Product Cubic Smoothing Splines)**.** *Suppose* $x = (x_1, x_2) \in [0,1]^2$ *is a bivariate biomarker variable. We start with the definitions of the marginal RKHS* $\mathcal{H}_{\langle j\rangle}$, $j = 1, 2$. *A popular choice of* $J_{\langle j\rangle}(f)$ *is* $\int_0^1 (f'')^2 dx_j$, *yielding the space of cubic splines* $\mathcal{H}_{\langle j\rangle} = \{f : J_{\langle j\rangle}(f) < \infty\}$. *If the inner product in* $\mathcal{H}_{0\langle j\rangle}$, *the null space of* $J_{\langle j\rangle}$, *is* $(\int_0^1 f\, dx_j)(\int_0^1 g\, dx_j) + (\int_0^1 f'\, dx_j)(\int_0^1 g'\, dx_j)$, *then* $\mathcal{H}_{J\langle j\rangle} = \mathcal{H}_{\langle j\rangle} \ominus \mathcal{H}_{0\langle j\rangle} = \{f : \int_0^1 f dx_j = \int_0^1 f' dx_j = 0, J_{\langle j\rangle}(f) < \infty\}$ *and the reproducing kernel* $R_{J_{\langle j\rangle}}(s,t) = k_2(s)k_2(t) - k_4(|s-t|)$, *where* $k_\nu(t) = B_\nu(t)/\nu!$ *are scaled Bernoulli polynomials with* $k_0(t) = 1$ *and* $k_1(t) = t - 0.5$ *for* $t \in [0,1]$. *The null space* $\mathcal{H}_{0\langle j\rangle}$ *has a dimension of 2, spanned by the basis functions* $\{1, k_1(x_j)\}$. *The marginal space* $\mathcal{H}_{\langle j\rangle}$ *is decomposed as*

$$\begin{aligned}\mathcal{H}_{\langle j\rangle} &= \big\{f : \textstyle\int_0^1 (f'')^2 dx_j < \infty\big\} = \mathcal{H}_{00\langle j\rangle} \oplus \mathcal{H}_{01\langle j\rangle} \oplus \mathcal{H}_{J\langle j\rangle} \\ &= span\{1\} \oplus span\{k_1(x_j)\} \oplus \big\{f : \textstyle\int_0^1 f dx_j = \int_0^1 \dot{f} dx_j = 0,\ \int_0^1 (f'')^2 dx_j < \infty\big\}.\end{aligned}$$

Taking tensor product of $\mathcal{H}_{\langle 1\rangle}$ *and* $\mathcal{H}_{\langle 2\rangle}$, *one obtains nine tensor sum terms* $\mathcal{H}_{\nu,\mu} = \mathcal{H}_{\nu\langle 1\rangle} \otimes \mathcal{H}_{\mu\langle 2\rangle}$ *on* $\mathcal{T} \times \mathcal{Z}$, $\nu = 00, 01, J$ *and* $\mu = 00, 01, J$. *The four subspaces with* $\nu = 00, 01$ *and* $\mu = 00, 01$ *can be lumped together as* $\mathcal{H}_0$. *The other five subspaces can be put together as* $\mathcal{H}_J$. *For interpretation, the nine subspaces define an ANOVA decomposition*

$$\eta(x_1, x_2) = \eta_\emptyset + \eta_1(x_1) + \eta_2(x_2) + \eta_{12}(x_1, x_2)$$

for functions on $\mathcal{X}_1 \times \mathcal{X}_2$, *with* $\eta_\emptyset \in \mathcal{H}_{00\langle 1\rangle} \otimes \mathcal{H}_{00\langle 2\rangle}$ *being the constant term,* $\eta_1 \in \{\mathcal{H}_{01\langle 1\rangle} \oplus \mathcal{H}_{1\langle 1\rangle}\} \otimes \mathcal{H}_{00\langle 2\rangle}$ *the main effect of* x_1, $\eta_2 \in \mathcal{H}_{00\langle 1\rangle} \otimes \{\mathcal{H}_{01\langle 2\rangle} \oplus \mathcal{H}_{1\langle 2\rangle}\}$ *the main effect of* x_2, *and* $\eta_{12} \in \{\mathcal{H}_{01\langle 1\rangle} \oplus \mathcal{H}_{1\langle 1\rangle}\} \otimes \{H_{01\langle 2\rangle} \oplus H_{1\langle 2\rangle}\}$ *the interaction. In this paper, we only consider*

the additive model for fitting the function η, that is, the model with $\eta_{12} = 0$. $\square$

The function space $\mathcal{H}$ is usually infinite-dimensional, and so the exact minimizer of (2.2) is not computable. Therefore, Gu [17] proposes to minimize (2.2) in a data-adaptive finite dimensional space $\mathcal{H}^* = \mathcal{H}_0 \oplus \text{span}\{R_J(Z_i, \cdot), i = 1, ..., q_{n_0}\}$, where Z_i is a size-q_{n_0} random subset of $\{x_1, \ldots, x_{n_0}\}$. It can be shown that the minimizers in $\mathcal{H}^*$ and $\mathcal{H}$ share the same asymptotic convergence rate when tensor product cubic splines with $q_{n_0} \asymp n_0^{2/9+\epsilon}$ are used, where $\epsilon > 0$ is an arbitrarily small number. For the simplicity of notation, we will drop the subscript from q_{n_0} and simply use q from now on.

Let η_λ be the minimizer of (2.2) in the adaptive space $\mathcal{H}^*$. A function in $\mathcal{H}^*$ has the expression

$$\eta(x) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(x) + \sum_{i'=1}^{q} c_{i'} R_J(Z_{i'}, x) = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c} \tag{2.3}$$

where $\boldsymbol{\phi}$ and $\boldsymbol{\xi}$ are vectors of functions and $\mathbf{c}$ and $\mathbf{d}$ are vectors of coefficients. Plugging (2.3) into (2.2) reduces the problem to the minimization of

$$A_\lambda(\mathbf{c}, \mathbf{d}) = -\frac{1}{n_0} \mathbf{1}^T (S\mathbf{d} + R\mathbf{c}) + \log \int_{\mathcal{X}} \exp(\phi^T \mathbf{d} + \xi^T \mathbf{c}) dx + \frac{\lambda}{2} \mathbf{c}^T Q \mathbf{c} \tag{2.4}$$

with respect to $\mathbf{c}$ and $\mathbf{d}$, where S is $n_0 \times m$ with the (i, ν) the entry $\phi_\nu(x_i)$, R is $n_0 \times q$ with the (i, i')th entry $\xi_{i'}(X_i) = R_J(Z_{i'}, X_i)$, and Q is $q \times q$ with the (i', k)th entry $R_J(Z_{i'}, Z_k)$.

Since (2.4) is a convex function of $\mathbf{c}$ and $\mathbf{d}$, its minimization with a fixed smoothing parameter λ can be carried out by a standard Newton-Raphson procedure. In an outer loop, the smoothing parameter λ can be selected to optimize a cross-validation score derived from the Kullback-Leibler distance between the true density and the estimated one based on η_λ; see Sec 7.3 in [17] for more details.

2.2.3 Extension to High Number of Biomarkers

The smoothing spline density estimation technique is effective for densities with 5 or fewer biomarkers, but becomes computationally expensive when a higher number of biomarkers are involved. To address this issue, we introduce a new approach to reduce the dimension of the smoothing spline densities. The key idea is to utilize the pseudo-likelihood ratio, which combines all possible combinations of likelihood ratios linearly, thus retaining the essential information of the biomarkers in a computationally efficient manner.

Suppose we have p candidate biomarkers, and most of them are irrelevant to predicting patients' status. When p is ultra high, we can use a nonparametric screening procedure such as the Sparse Additive Model (SpAM) [40], Nonparametric Independence Screening (NIS) [12], or Sure Independent Ranking and Screening (SIRS) [53] to reduce the dimensional to a relatively small number K with $K \ll p$. When K is still larger than 5, we apply the triple-wise pseudo-likelihood ratio idea as detailed below. Our choice is based on our pilot simulations that show the triple-wise approach is faster and yields a higher AUC compared to the pair-wise method.

For notation simplicity, we will still use the $x_i = (x_{i1}, ..., x_{iK})$ to denote the ith subject's values for the K selected biomarkers, with the notion that the x_{ij} here does not necessarily match with x_{ij} in the original data. Let $T = \binom{K}{3}$ and $(k_t, k_t', k_t''), t = 1, \ldots, T$, be all the index combinations with $k_t \neq k_t' \neq k_t''$. Then the pseudo-likelihood ratio of subject i is defined as a linear combination of triple-wise log likelihood ratios,

$$\sum_{t=1}^{T} \beta_t \log \frac{f(x_{ik_t}, x_{ik_t'}, x_{ik_t''})}{g(x_{ik_t}, x_{ik_t'}, x_{ik_t''})} \equiv \mathbf{h}_i^T \boldsymbol{\beta}, i = 1, ..., n. \tag{2.5}$$

Here $\mathbf{h}_i$ is the vector of the log likelihood ratios and can be estimated using the corresponding smoothing spline density estimates for f and g. The vector $\boldsymbol{\beta}$ consists of all the coefficients

and can be computed by the Ramp AUC in Fong et al. [13]. The Ramp AUC is a transformation of the empirical AUC, but provides the estimation of linear coefficients based on a convex objective function. The aim is to maximize the empirical AUC defined as

$$\frac{1}{n_0(n-n_0)}\sum_{i=1}^{n_0}\sum_{j=n_0+1}^{n}\big[1-I\big\{(\mathbf{h}_i-\mathbf{h}_j)^T\boldsymbol{\beta}>0\big\}\big] \equiv \frac{1}{n_0(n-n_0)}\sum_{r=1}^{n_0(n-n_0)}\big[1-I\big\{\phi_r>0\big\}\big], \quad (2.6)$$

where the right hand side is a re-indexing of the double indices i and j on the left hand side to a single index r such that $\phi_r = (\mathbf{h}_{i_r} - \mathbf{h}_{j_r})^T\boldsymbol{\beta}$. To find the best combination coefficients $\boldsymbol{\beta}$, Fong et al. [13] used the ramp function as a continuous approximation to the step function $1 - I(x)$, which is defined as

$$u(x) = \begin{cases} 1 & \text{if} \quad x \leqslant -\frac{1}{2} \\ -x+\frac{1}{2} & \text{if} \quad -\frac{1}{2} \leqslant x < \frac{1}{2}. \\ 0 & \text{if} \quad x > \frac{1}{2} \end{cases}$$

To build a convex and smooth objective function, they write the ramp function u as the difference between two convex functions $u(x) = u_1(x) - u_2(x)$, where $u_1(x) = (\frac{1}{2} - x)_+$ and $u_2(x) = (-\frac{1}{2} - x)_+$. Then the penalized RAUC loss is proposed:

$$\min_{\beta} \sum_{r=1}^{n_0(n-n_0)} u\big(\phi_r\big) + \frac{1}{2}\lambda_n\|\boldsymbol{\beta}\|_2^2 \quad (2.7)$$

Here, constraining $\boldsymbol{\beta}$ by the l_2-norm $||\boldsymbol{\beta}||_2$ rather than the l_1-norm $||\boldsymbol{\beta}||_1$ ensures a more uniform approximation of the AUC loss across the parameter space. Other details of selecting $||\boldsymbol{\beta}||_2$ are in [13]. Writing the ramp function as a difference of two convex functions allows for the use of Differential Convex Analysis (DCA)[29], such that a series of convex optimization problems can be used to approximate the non-convex problem. The algorithm detail is referred to [13] and already implemented in the R package `aucm`.

2.3 Simulation Study

In this section, we present a numerical evaluation of various combination approaches for predicting disease status with multiple biomarkers. Among the existing methods, the distribution-free approach [35] is a commonly used method that linearly combines biomarkers by maximizing the AUC without assuming any specific distribution (referred to as **LIN**). Another approach is the risk score method using a monotone increasing transformation of the likelihood ratio function [31]. A logistic regression model can then be fitted based on the risk scores over the linear combinations of biomarkers (referred to as **LOGIT**). Additionally, we evaluate the linear combination of the maximum and minimum values of the biomarkers proposed by Liu et al. [27] (referred to as **Min-Max**). Yan et al. [50] proposed the kernel-based approach estimating the densities for either biomarker and then linearly combining them, referred to as **KERN**. As a traditional machine learning algorithm designed for binary classification, the support vector machine (**SVM**) is renowned for its computational efficiency and is thus considered. We compare these approaches to the proposed Smoothing Spline Density Estimation method (denoted as **SSD**) in terms of the prediction accuracy represented by the empirical AUC, which is defined as the Mann-Whitney U statistic

$$\widehat{AUC} = \frac{1}{n_0(n-n_0)} \sum_{i=1}^{n_0} \sum_{j=n_0+1}^{n} \left\{ I\left[\hat{Y}_i > \hat{Y}_j{}^*\right] + \frac{1}{2} I\left[\hat{Y}_i = \hat{Y}_j{}^*\right] \right\}, \tag{2.8}$$

where $\hat{Y}_i, i = 1, \ldots, n_0$ and $\hat{Y}_j^*, j = n_0 + 1, \ldots, n$ are respectively the estimated combination scores for the diseased and non-diseased patients from the corresponding method. As the benchmark of the comparisons, we also calculate the AUC of the true likelihood ratio statistic (referred to as **TRUE**).

For the low-dimensional case, we consider 3 biomarkers and simulate three scenarios of joint distributions for the biomarkers with N=2000 subjects. Disease prevalence is an important

factor and imbalanced sample sizes between the case and controls may influence the combination approaches so we assume a disease prevalence of 0.2, 0.5, or 0.8. We randomly split the subjects into two sets of 1000 subjects each, with one set used as the training sample and the other as the testing sample. We repeat the simulation 500 times to obtain a reliable estimate of the mean and standard deviation of the AUCs. The means and standard deviations of the AUCs provide a performance measure of each approach and allow us to compare their precision and assess their robustness and generalizability.

2.3.1 Multivariate normal with equal covariance

Let $D = 0$ denote the controls and $D = 1$ the cases. Firstly, we consider both the cases and the controls follow multivariate normal distributions sharing a covariance matrix: $x|D = d \sim MVN(\mu_d, \Sigma_d), d = 0, 1$, where $\mu_0 = (0, 0, 0)^T$, $\mu_1 = (1.5, 1, 0)^T$, $\Sigma_0 = \Sigma_1 = \begin{pmatrix} 1.5 & 0.5 & 0.5 \\ 0.5 & 2 & 0.5 \\ 0.5 & 0.5 & 2.2 \end{pmatrix}$.

Methods / Prevalence	True	SSD	LIN	LOGIT	Min-Max	SVM	KERN
0.2	0.828 (0.013)	0.823 (0.014)	0.823 (0.014)	0.825 (0.014)	0.733 (0.018)	0.824 (0.038)	0.823 (0.017)
0.5	0.83 (0.013)	0.829 (0.012)	0.828 (0.013)	0.828 (0.013)	0.736 (0.015)	0.829 (0.014)	0.828 (0.014)
0.8	0.827 (0.015)	0.825 (0.016)	0.826 (0.015)	0.826 (0.015)	0.735 (0.019)	0.826 (0.017)	0.824 (0.018)

Table 2.1: The means and standard deviations (in parentheses) of the simulated AUCs by six different methods of combining the three biomarkers generated from the multivariate normal distribution with equal variance-covariance under three prevalence

Table 2.1 shows that, except for the Min-Max method, as expected, all the methods (including the proposed SSD method) can yield mean AUCs very close to the mean AUCs from

the TRUE method. The LIN approach, which is considered optimal under the multivariate normality assumption, performs well as expected. This echoes with the findings of the simulation studies in Chunling et al. [27]. The lower mean AUCs from the Min-Max method indicate that linear combinations of only the minimum and maximum biomarkers are not sufficient to capture all the biomarker information to accurately predict disease status in this setting.

2.3.2 Multivariate normal with unequal covariance

For this scenario, we still consider that the cases and the controls follow multivariate normal distributions, but with different variance-covariance structures: $\Sigma_0 = \begin{pmatrix} 1.5 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2.2 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 1.8 & 0.68 & 1.55 \\ 0.68 & 2.2 & 0.4 \\ 1.55 & 0.4 & 2 \end{pmatrix}$, all the other parameter settings are the same as Sec 2.3.1.

Methods / Prevalence	True	SSD	LIN	LOGIT	Min-Max	SVM	KERN
0.2	0.888 (0.011)	0.884 (0.012)	0.829 (0.016)	0.828 (0.015)	0.711 (0.026)	0.809 (0.022)	0.833 (0.017)
0.5	0.89 (0.013)	0.885 (0.012)	0.834 (0.012)	0.835 (0.013)	0.714 (0.017)	0.832 (0.013)	0.834 (0.011)
0.8	0.884 (0.01)	0.879 (0.012)	0.822 (0.016)	0.83 (0.016)	0.71 (0.018)	0.793 (0.022)	0.832 (0.018)

Table 2.2: The means and standard deviations (in parentheses) of the simulated AUCs by six different methods of combining the three biomarkers generated from the multivariate normal distribution with unequal variance-covariance under three prevalence

Table 2.2 shows that the proposed SSD method performs the best among all the approaches,

with all the mean AUCs around 0.88, very close to the mean AUCs obtained by the true likelihood ratios. The linear combination (LIN), kernel-based approach (KERN) and logistic regression (LOGIT) approaches perform similarly, with lower mean AUCs than the SSD method. The SVM method follows closely behind these methods. On the other hand, the Min-Max approach still suffered in this setting. It is also noteworthy that, as expected, all the biomarker combination methods are not affected much by the different levels of disease prevalence influences since the AUC criterion does not depend on the disease prevalence. On the other hand, as a classification method optimizing the so-called hinge loss, the SVM seems to suffer slightly from an unbalanced prevalence value. A possible reason is that the relatively smaller sample size in one of two subject groups may have caused some instability in the identification of the support vectors used for constructing the decision boundary.

2.3.3 Multivariate Gamma

In the last setting, we consider the biomarkers from the multivariate gamma distributions, which is consistent with the situation when some extreme biomarker values exist. For the cases, $x_1|d = 1 \sim Gamma(15, 0.04)$, $x_2|d = 1 \sim Gamma(6.5, 0.006)$, $x_3|d = 1 \sim Gamma(19, 7.5)$ with correlation structure $R = \begin{pmatrix} 1 & 0.4 & 0.75 \\ 0.4 & 1 & 0.2 \\ 0.75 & 0.2 & 1 \end{pmatrix}$. For the controls, $x_1|d = 0 \sim Gamma(16, 0.05)$, $x_2|d = 0 \sim Gamma(7.2, 0.01)$, $x_3|d = 0 \sim Gamma(23, 11)$, and they are independent of each other. In practice, density estimation requires the specification of the domain where data live in. Different from normal distributions, Gamma distributions can often result in a small number of exceptionally extreme values. This creates difficulty in specifying a sufficiently large but data-informative domain. We use the following standard approach to counter this problem. Before fitting any density, we repeat

the data generation simulations a large number of times and calculate the ranges of x_1, x_2, and x_3 separately. We then use the means of upper and lower bounds of these ranges as our data domains. When generating samples, only values within the domains are accepted.

Methods / Prevalence	True	SSD	LIN	LOGIT	Min-Max	SVM	KERN
0.2	0.867 (0.014)	0.854 (0.015)	0.775 (0.019)	0.81 (0.019)	0.763 (0.018)	0.807 (0.018)	0.782 (0.019)
0.5	0.867 (0.013)	0.863 (0.012)	0.78 (0.012)	0.813 (0.013)	0.769 (0.017)	0.814 (0.014)	0.784 (0.015)
0.8	0.866 (0.013)	0.856 (0.014)	0.779 (0.015)	0.811 (0.014)	0.765 (0.016)	0.722 (0.019)	0.78 (0.014)

Table 2.3: The means and standard deviations (in parentheses) of the simulated AUCs by six different methods of combining the three biomarkers generated from the multivariate gamma distributions under three prevalence

Under this setting, the biomarker values are more spreadout than in the previous two settings. However, this does not affect the performance of the SSD method at all, as indicated by the results in Table 2.3. Once again, the SSD approach yields a larger mean AUC than the other methods. Its mean AUCs of more than 0.85 under all prevalence are also very close to the mean AUC from the true likelihood ratios. The logistic regression (LOGIT) approach follows the SSD method with mean AUCs of around 0.81, indicating its flexibility in being applied under different distribution settings. However, the linear combination approaches, such as the LIN approach, the KERN approach, and the Min-Max approach, seem to be less efficient here, as indicated by their relatively low mean AUCs. Although the linear combination approaches claim to be distribution-free, they actually appear to suffer when the data distributions deviate from the normal distributions. The SVM approach can achieve performance comparable to logistic regression when there are equal numbers of diseased and non-diseased samples. Analogous to the previous setting, its mean AUC values under the

other two prevalence levels markedly decline, indicating pronounced instability.

2.3.4 Mixture of Multivariate Normal

In the last scenario, a more intricate setting is considered to handle instances where the biomarkers display bimodalities. This can be represented effectively through the application of a mixture of two distributions:

$$x|D = d \sim MVN(\mu_1, \Sigma_1)I(\delta = 0) + MVN(\mu_2, \Sigma_2)I(\delta = 1),$$

where $I(\cdot)$ is the indication function and $\delta \sim Bernoulli(0.5)$. When $D = 1$, we set $\mu_1 = (1,1,1)^T$, $\mu_2 = (2,2,2)^T$, $\Sigma_1 = \begin{pmatrix} 0.25 & 0.35 & 0.27 \\ 0.35 & 1 & 0.54 \\ 0.27 & 0.54 & 0.6 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & 0.35 & 0.99 \\ 0.35 & 0.25 & 0.49 \\ 0.99 & 0.49 & 2 \end{pmatrix}$. While when $D = 0$, we set $\mu_1 = (0,0,0)^T$, $\mu_2 = (1,1,1)^T$, $\Sigma_1 = \begin{pmatrix} 0.5 & 0.8 & 0.57 \\ 0.8 & 2 & 1.13 \\ 0.57 & 1.13 & 1 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & 0.35 & 0.99 \\ 0.35 & 0.25 & 0.49 \\ 0.99 & 0.49 & 2 \end{pmatrix}$.

Table 2.4 presents the performance outcomes of various methods across three distinct prevalence levels when the biomarkers follow bimodal distributions. Among all the methods evaluated, the SSD approach consistently registers the highest mean AUCs. All the other methods manifest performances comparable to each other, with mean AUCs under 0.8. The SVM method still has reduced AUCs under unbalanced prevalence levels in this simulation setting.

Methods / Prevalence	True	SSD	LIN	LOGIT	Min-Max	SVM	KERN
0.2	0.925 (0.009)	0.864 (0.013)	0.761 (0.016)	0.747 (0.017)	0.765 (0.015)	0.675 (0.121)	0.777 (0.017)
0.5	0.926 (0.007)	0.856 (0.018)	0.758 (0.016)	0.755 (0.015)	0.765 (0.017)	0.754 (0.016)	0.774 (0.015)
0.8	0.926 (0.008)	0.853 (0.021)	0.758 (0.022)	0.759 (0.02)	0.763 (0.018)	0.71 (0.103)	0.767 (0.02)

Table 2.4: The means and standard deviations (in parentheses) of the simulated AUCs by six different methods of combining the three biomarkers generated from the mixture of two multivariate normal distributions under three prevalence

Methods / Settings	SSD	LIN	LOGIT	Min-Max	SVM	KERN
1	28.42	52.07	0.02	6.21	0.08	3.11
2	28.07	53.56	0.02	5.89	0.12	3.07
3	20.95	57.83	0.03	6.32	0.09	4.22
4	28.27	120.02	0.03	5.5	0.1	4.87

Table 2.5: The average computational time in seconds of the simulations by six different methods of combining the three biomarkers under the above four data settings

To compare the computational efficiencies of these methods, we have also recorded their average running times in all the above simulation settings in Table 2.5. The simulations were run on a Windows laptop with an intel Core i7-10510U CPU at 1.80GHz-2.30GHz and 8GB of RAM. The LOGIT and SVM methods are the fastest with average running times within a second. Both Min-Max and KERN could run in several seconds. The SSD method takes half a minute and the LIN method takes a minute or so. In summary, all the methods can run in a practically reasonable length of time. The small computational time "disadvantage" of the SSD method can clearly offsetted by its superior and robust numerical performance of the SSD method in all the simulation settings.

2.3.5 High number of Biomarkers

In this section, we will compare the Pseudo-Likelihood extension of the proposed SSD method with the TRUE approach in the scenario of a high number of biomarkers. In this simulation, we consider a setting with fifteen biomarkers in total, specific numbers of which are generated from our specified distributions, while the rest are generated from a normal distribution with mean 0 and variance 1 (i.e., noise biomarkers). We assume that there are 100 cases and 100 controls in our sample, and we will repeat this simulation 500 times. We specify the distributions of the four biomarkers as follows: $x|D = d \sim MVN(\mu_d, \Sigma_d),\ d = 0, 1$, where $\mu_0 = (0, 0, 0, 0)^T$, $\mu_1 = (1.5, 1, 2, 2.5)^T$, $\Sigma_0 = \Sigma_1 = \begin{pmatrix} 1.5 & 0.5 & 0.5 & 0.5 \\ 0.5 & 2 & 0.5 & 0.5 \\ 0.5 & 0.5 & 2.2 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}$.

We specify similar settings for the six biomarkers' case where μ_0 is a zero vector and $\mu_1 = (1.5, 1, 2, 2.5, 1.5, 2)$ and $diag(\Sigma_0) = diag(\Sigma_1) = (1.5, 2, 2.2, 1, 1, 1.5)$ and all 0.5 for the covariance. A higher dimensional is considered that eight biomarkers are generated from a similar setting where $\mu_1 = (1.5, 1, 2, 2.5, 1.5, 2, 1, 2)$ and $diag(\Sigma_0) = diag(\Sigma_1) = (1.5, 2, 2.2, 1, 1, 1.5, 1.2, 1.5)$ with all 0.5 for the covariance. We apply the SpAM to identify the biomarkers with high functional norms that are most useful for predicting the status of the subjects. The results of one simulation for specified four biomarkers using the SpAM approach are summarized in Figure 2.1. We can observe that some of the biomarkers have much higher functional norms than others, indicating that they are more informative for predicting disease status.

From the results of our simulations, it is clear that the four biomarkers selected by SpAM consistently have higher functional norms than the other biomarkers. Similarly, higher functional norms are observed for the specified six and eight biomarkers. We then used the

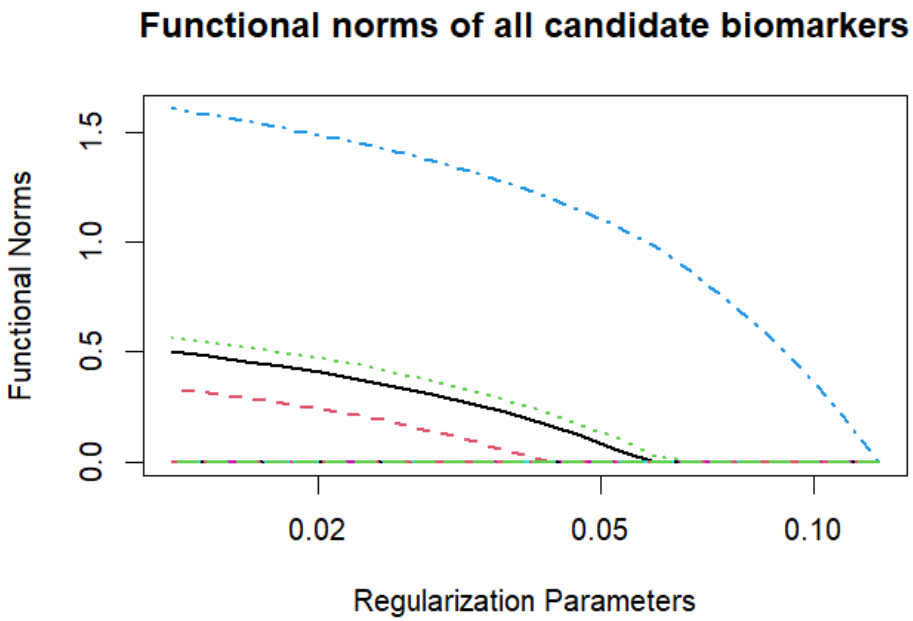

Figure 2.1: The functional norms of the estimated components is plotted against the tunning parameter λ. The solid, dotted, and dashed lines and dotted-dashed lines correspond to the four biomarkers generated from our specified distribution.

selected biomarkers to make up the Pseudo-likelihood ratio and implemented the RAUC approach to estimate the coefficients of the ratios. Also, other approaches mentioned above are applied to the selected biomarkers for comparisons.

Methods / Dimension	True	Pseudo	LIN	LOGIT	Min-Max	SVM	KERN
4	0.977 (0.007)	0.975 (0.012)	0.945 (0.015)	0.966 (0.011)	0.911 (0.02)	0.967 (0.01)	0.96 (0.011)
6	0.978 (0.01)	0.973 (0.012)	0.956 (0.013)	0.97 (0.013)	0.916 (0.017)	0.969 (0.011)	0.967 (0.011)
8	0.974 (0.011)	0.971 (0.014)	0.957 (0.013)	0.972 (0.013)	0.913 (0.016)	0.971 (0.011)	0.965 (0.015)

Table 2.6: The means and standard deviations (in parentheses) of the simulated AUCs by five different methods of combining the selected biomarkers after screening

The results are in Table 2.6. Our approach yields a mean empirical AUC of more than 0.97, very close to the mean empirical AUCs from the benchmark TRUE approach, which leverages the true likelihood ratio statistic. Excellent performance is also achieved by the logistic regression (LOGIT) approach, the KERN approach, and the SVM approach, especially with a higher specified dimension. They clearly outperform the LIN approach and the Min-Max approach.

2.4 Data Example: diagnosis of autism by combining growth-related hormones

In this section, we apply the SSD method to a study investigating the relationship between growth-related hormones and autism spectrum disorders (ASD). The data were from an autism study conducted by the Eunice Kennedy Shriver National Institute of Child Health and Human Development between 2002 and 2005. The study included 161 children aged 4-8 years, of which 81 were diagnosed with ASD and 80 were controls. Blood samples were extracted from the participants, and the levels of six growth-related hormones were measured, including insulin-like growth factor binding protein (IGFBP3), dehydroepiandrosterone (DHEA), growth hormone binding protein (GHBP), and two insulin-like growth factors (IGFI, IGFII). Following the guideline in Mills et al. [32], we excluded the samples for girls due to their low sample size and those without complete blood samples, resulting in a sample size of 71 cases and 59 controls. We also excluded the biomarker DHEAS due to its high proportion of missing values.

Since the number of biomarkers is exactly the maximum dimension available in current smoothing spline density estimation software, we used both the SSD method and its faster

Pseudo-Likelihood ratio extension in our analysis, together with the other five existing methods. Figure 2.2 shows that all five methods performed well, with AUCs larger than 0.7. Again, the SSD method outperformed the other methods, with an AUC of 0.897, indicating its superiority in identifying the disease status correctly. Additionally, the Pseudo-Likelihood ratio extension also performed similarly to the SSD method, with an AUC of 0.879. The existing methods all had AUCs between 0.7 and 0.8. The AUCs of the LIN, KERN, LOGIT, and SVM approaches were respectively 0.797, 0.795, 0.789, and 0.784, indicating their lower accuracy in identifying the disease status. The Min-Max approach yields an AUC of 0.71. This result is notably lower compared to other methods.

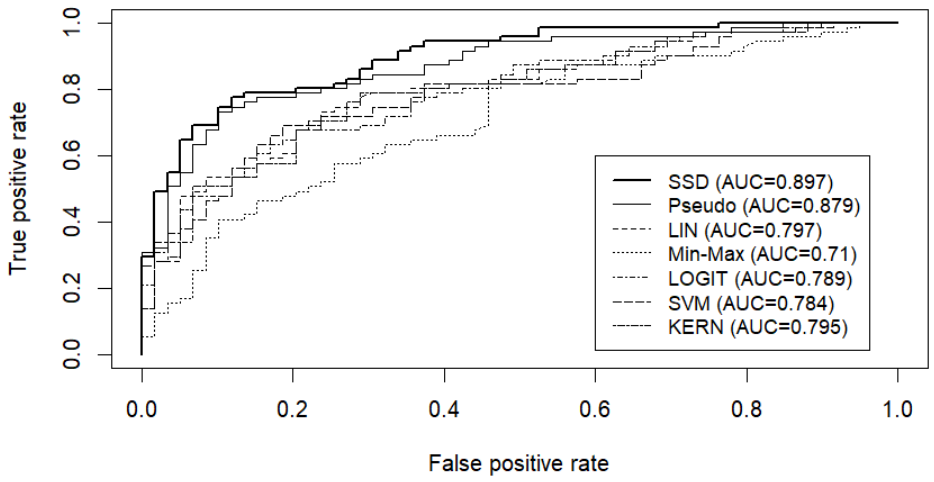

Figure 2.2: Seven methods of combining IGF1, IGFII, IGFBP3, GHBP, DHEA for diagnosis of autism

2.5 Discussion

The diagnostic accuracy of a biomarker is often evaluated using the area under the ROC curve (AUC), which reflects the ability of the biomarker to discriminate between diseased

and non-diseased individuals. The challenge of combining multiple biomarkers to achieve a high AUC is a common goal in medical research. While many methods for estimating the likelihood ratio have been proposed, these approaches are often limited by assumptions about the underlying distributions and the number of biomarkers, which may not hold in practice. In contrast, estimating the multivariate density without assuming its form is a more accurate and innovative approach, as demonstrated in this article.

Although the SSD approach has demonstrated its robust excellent performance in the simulation studies and the application, there are still challenges with this method. For example, the domains of the biomarkers need to be specified for the density estimation procedure. This would require the exclusion of some extreme values in the data since little information is available in those regions to fit a reasonable density estimate. Therefore, the estimated density is actually a truncated version of the boundless true density, such as the normal or gamma density. Additionally, the computational time increases quickly with the number of biomarkers increasing, and high collinearity among the biomarkers may lead to the non-convergence of the model. Therefore, some exploratory analysis addressing these issues is recommended prior to fitting the SSD approach.

The pseudo-likelihood ratio extension provides a solution for fitting high-dimensional biomarkers by first fitting all combinations of biomarkers in low dimensions and then estimating the linear coefficients after predictor screening. However, the efficiency of this approach is lower than fitting all biomarkers directly, as demonstrated in the data example above.

In this paper we have used smoothing splines to estimate the probability densities. An alternative is the logspline method [23]. However, as well-known in the smoothing method literature [39], a roughness penalty approach like smoothing splines can often yield more visually appealing and numerically accurate estimates than a simple basis expansion approach like the logspline. This is because the smoothing parameter in a roughness penalty allows

a continuous control on the smoothness of the function estimate. On the other hand, a simple basis expansion approach only has a discrete control on the smoothness of the estimate through the selection of the number of basis functions. This advantage is even more significant for multivariate density estimation.

Future research may focus on estimating the likelihood ratio directly, based on its optimality, instead of estimating the densities of the diseased and non-diseased groups separately. Furthermore, improving the estimation of the linear coefficients in the Pseudo-likelihood ratio method by finding a closed form of the objective function may also be a promising avenue for future investigation. A much more challenging future direction is statistical inference for the proposed method. This would require the derivation of asymptotic distributions for the smoothing splines density estimates and then translate the results to provide confidence intervals for the corresponding ROC curve. However, such inference theory is a non-trivial task. The most recent work in this direction is the inference theory for smoothing splines regression models [8]. The extension of this work to density estimation by itself already warrants another paper. Therefore, further investigation in this direction is out of the scope of this paper.

Chapter 3

Self-normalization Tests for Change Points in Functional Time Series

3.1 Introduction

With the rise of advanced technology, we're seeing an influx of functional data that generalizes the traditional types of scalar or vector data and calls for more sophisticated statistical methods. When functional data are collected over time, various traditional time series models, such as the autoregressive model, have been extended to accommodate this functional feature in the research community.

One interesting topic in time series is the detection of possible change points. There is an extensive literature on change point analysis for scalar time series, with two representative examples in Csörgö and Horváth [9] and Perron et al. [37]. In the realm of mean shift testing, ensuring a consistent estimate of the long-run variance (LRV) that accounts for temporal data dependence becomes paramount, since such an estimate is necessary for parameter-free asymptotic analysis under the null hypothesis. Despite numerous literature on selecting the bandwidth parameter inherent to LRV, challenges persist, such as the non-monotonic power problem pointed out by Vogelsang [47] and Deng and Perron [11]. To tackle this issue, Shao and Zhang [43] introduced a self-normalization (SN) based test, where the normalization matrix not only factors in dependence but also ensures monotonic power for the test statistic.

In this work, we aim to extend the SN approach to the functional time series setting.

Extensive research has been conducted on change point testing for functional time series and the CUSUM-based (cumulative sum) methodology remains the most commonly used approach. Berkes et al. [2], among the first functional time series studies, considered a CUSUM-based test for the mean function change of independent functional data. Noticing its limitation of not accounting for temporal dependence, Hörmann and Kokoszka [19] proposed a test by incorporating a consistent LRV estimator, which was further extended by Horváth et al. [21]. However, both versions, like their counterparts for scalar time series, have the issue of bandwidth or other tuning parameter selection. Poor or inaccurate estimation can lead to non-monotonic powers. Zhang et al. [52] incorporated the self-normalization matrix into the functional context using the CUSUM process built up on empirical score vectors from the functional principle component analysis (FPCA). Note that all the aforementioned tests share a common reliance on FPCA and its efficacy may be questioned in the context pointed out by Cai and Yuan [5]. For example, if the covariance processes for the functional data groups have very different eigenfunctions, the performance of FPCA representations can deteriorate significantly. In the context of functional time series, this can happen when the covariance process also changes dramatically together with the mean function.

For tests not relying on the FPCA, bootstrap methods present a fitting solution, especially when dealing with the inestimable infinite-dimensional covariance operator in the asymptotic distributions. Among them, the block bootstrap method, renowned for handling time series data, is particularly attractive. Its efficacy in both independent and dependent data scenarios is extensively discussed in Lahiri and Lahiri [24]. Carlstein [6] delved into the non-overlapping block bootstrap, demonstrating its asymptotic convergence, while Sharipov et al. [44] applied it to change point detection in Hilbert space-valued random variables. Bucchia and Wendler [4] and Wegner and Wendler [49] provided generalizations of the dependent wild bootstrap

introduced by Shao [41] to functional time series.

Here we develop a functional extension of the SN test [43] for univariate time series to test for a mean shift in a functional time series. Given a candidate time point for the change point, we first construct the CUSUM process and its normalization factor from the functional time series. Then our functional SN test statistic is defined as the supreme of their ratios over all the time points. Based on the functional central limit theorem in Sharipov et al. [44], we first derive the asymptotic distribution of the test statistic under the null hypothesis of no change point and then establish its test consistency under both fixed and local alternatives. Due to the structural complexity of the limiting distribution, a Monte Carlo simulation approach to obtaining it similar to Shao and Zhang [43] is impossible. Instead, we resort to the non-overlapping sequential block bootstrap procedure introduced in Sharipov et al. [44]. First, the whole functional time series is divided into a number of non-overlapping blocks. Then an equal number of blocks are drawn from the original blocks with replacement to obtain a bootstrap sample. A bootstrap version of the CUSUM process, its normalization factor, and the test statistic are then obtained from a large number of such bootstrap samples. We show that the bootstrap test statistic share the same null distribution with the original one, even under a sequence of local alternatives converging to the null at the rate of $n^{-1/2}$. We further extend both the raw and the bootstrap data versions of the mean test procedure to test on the stationarity of the lag-1 autocovariance operator and thus examine the validity of a functional AR(1) model for the functional time series of interest. Our simulations show that our functional SN (FSN) test is better at detecting mean changes in areas with small variabilities than the existing ones, due to the additional normalization factor in its test statistics. We then illustrate the application of the proposed testing procedure to two meteorological examples.

The rest of the paper is organized as follows. In section 3.2, we provide the details of our

Functional Self-Normalized (FSN) test statistic by deriving its null limit distribution, consistency, and the proof of asymptotic properties for its bootstrapped counterpart. Detection on both mean and Lag autocovariance are introduced in this part. In Section 3.3, we showcase numerical studies by comparing the performance of our FSN test against previously established tests for both mean and Lag-1 autocovariance. Section 3.4 provides a practical application of our test using real-world data. Finally, Section 3.5 offers a concluding discussion. All the technical proofs for the theorems are collected in the Appendix.

3.2 Methodology

3.2.1 Mean Hypotheses and Notation

Suppose our functional time series observations are $X_t \in H, t = 1, \ldots, n$, where H is a separable Hilbert space with the corresponding inner product and norm respectively denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$. For an H-valued random variable X, we define its mean μ, also denoted by EX, to be an element in H such that $E\langle X, h \rangle = \langle \mu, h \rangle$ for all $h \in H$. The covariance of X is defined as the operator $S : H \to H$ such that

$$\langle Sh_1, h_2 \rangle = E\left[\langle X - EX, h_1 \rangle \langle X - EX, h_2 \rangle\right] \quad \text{for } h_1, h_2 \in H.$$

Our primary goal is to test for a change point in the mean function of the series with the hypothesis

$$H_0 : EX_1 = \cdots = EX_n$$
$$\text{vs. } H_1 : EX_1 = \cdots = EX_{\tau_0} \neq EX_{\tau_0+1} = \cdots = EX_n$$

where τ_0 represents the potential change point position.

We first introduce a notion of weak independence from [10] for a sequence of random variables in H. Let $(\xi_i)_{i\in\mathbb{Z}}$ be a stationary sequence of random variables taking values in some separable measurable space. A stationary sequence $(X_n)_{n\in\mathbb{Z}}$ in H is called L_p-near epoch dependent (NED(p)) on $(\xi_i)_{i\in\mathbb{Z}}$ if there exists a sequence $(a_k)_{k\in\mathbb{N}}$ with $\lim_{k\to\infty} a_k = 0$ such that $E\left[\left\|X_0 - E\left[X_0 \mid \mathcal{F}_{-k}^{k}\right]\right\|^p\right] \leq a_k$, where $\mathcal{F}_{-l}^{m} = \sigma(\xi_{-l}, \ldots, \xi_m)$ denotes the σ-field generated by $\xi_{-l}, \ldots, \xi_m$. Define $\beta(k) = \left|E \sup_{A\in\mathcal{F}_k^\infty}\left[P(A \mid \mathcal{F}_{-\infty}^{0}) - P(A)\right]\right|$. The sequence $(\xi_i)_{i\in\mathbb{Z}}$ is called absolutely regular if $\lim_{k\to\infty}\beta(k) \to 0$.

Let $D_H[0,1]$ be the space of all cadlag functions mapping from $[0,1]$ to H. Define the Skorohod metric on $D_H[0,1]$ as $d(f,g) = \inf_{\lambda\in\Lambda}\left\{\sup_{t\in[0,1]}\|f(t) - g\circ\lambda(t)\| + \|id - \lambda\|_\infty\right\}$, where Λ is the class of strictly increasing, continuous mappings of $[0,1]$ onto itself, $id : [0,1] \to [0,1]$ is the identity function, $\circ$ denotes composition of functions, and $\|\cdot\|_\infty$ is the supremum norm. Then $D_H[0,1]$ is a separable Banach space equipped with the Skorohod metric d.

3.2.2 Functional Self-Normalization Test Statistics

Given the functional time series $\{X_n\}_{n\in\mathbb{Z}}$ and a range $\tau = 1, ..., n$, the partial sum process is

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{\tau}(X_t - \mu). \tag{3.1}$$

Let $\tau = \lfloor nr\rfloor, r \in [0,1]$. We first cite a convergence result of the partial sum process from Sharipov et al. [44], for which we need the definition of a Brownian motion in H.

An H-valued random variable N is Gaussian if $\langle N, h\rangle$ is normally distributed for all $h \in H\setminus\{0\}$. A random element W of $D_H[0,1]$ is called a Brownian motion in H if: (i) $W(0) = 0$ almost surely, (ii) $W \in C_H[0,1]$ almost surely, where $C_H[0,1]$ is the set of all continuous

functions from $[0,1]$ to H, (iii) the increments on disjoint intervals are independent, and (iv) for all $0 \leq u < u+v \leq 1$, the increment $W(u+v) - W(u)$ is Gaussian with mean zero and covariance operator S, where $S : H \to H$ does not depend on u or v.

Theorem 3.1 (Theorem 1.1 of Sharipov et al. [44]). *Let $(X_n)_{n\in\mathbb{Z}}$ be L_1-near epoch dependent on a stationary and absolutely regular sequence $(\xi_n)_{n\in\mathbb{Z}}$ with $EX_1 = \mu \in H$ and assume that for some $\delta > 0$,*

$$E\left\|X_1\right\|^{4+\delta} < \infty, \quad \sum_{m=1}^{\infty} m^2 \left(a_m\right)^{\delta/(\delta+3)} < \infty, \quad \sum_{m=1}^{\infty} m^2 (\beta(m))^{\delta/(\delta+4)} < \infty.$$

Then

$$\left(\frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor nr \rfloor} (X_t - \mu)\right)_{r\in[0,1]} \Rightarrow (W(r))_{r\in[0,1]}, \tag{3.2}$$

where $(W(r))_{r\in[0,1]}$ is a Brownian motion in H and $W(1)$ has the covariance operator $S : H \to H$, defined by

$$\langle Sx, y\rangle = \sum_{t=-\infty}^{\infty} E\left[\langle X_0 - \mu, x\rangle \langle X_t - \mu, y\rangle\right], \quad \text{for } x, y \in H.$$

Now, our CUSUM process is defined as

$$D_{n,\tau} = \frac{1}{\sqrt{n}} \sum_{t=1}^{\tau} \left(X_t - \bar{X}_n\right) \tag{3.3}$$

where $\bar{X}_n = n^{-1}\sum_{t=1}^{n} X_t$ is the mean function of the whole functional time series. We can also write it as $D_{n,\lfloor nr \rfloor} = \frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor nr \rfloor}(X_t - \bar{X}_n), r \in [0,1]$. To extend the self-normalization (SN) test statistic in Shao and Zhang [43], we let $A_{t_1,t_2} = \sum_{t=t_1}^{t_2} X_t$ if $t_1 \leq t_2$ and 0 otherwise. The normalization factor is defined as

$$V_{n,\tau} = \frac{1}{n^2}\Big[\sum_{t=1}^{\tau}\{A_{1,t} - (t/\tau)A_{1,\tau}\}^2 + \sum_{t=\tau+1}^{n}\{A_{t,n} - (n-t+1)/(n-\tau)A_{\tau+1,n}\}^2\Big]. \tag{3.4}$$

Subsequently, our Functional Self-Normalization (FSN) test statistic is

$$T_n = \sup_{1<\tau<n-1}\left\|\frac{D_{n,\tau}}{\sqrt{V_{n,\tau}}}\right\|, \tag{3.5}$$

which generalizes the SN test statistic in Shao and Zhang [43] for univariate time series to functional time series. Based on the convergence result for the partial sum process in (3.2), we can derive the asymptotic distribution of our test statistic under the null hypothesis. See proof in the appendix.

Theorem 3.2. *Suppose Assumptions in Theorem 3.1 hold. Then under H_0, we have*

$$T_n \Rightarrow \sup_{r\in[0,1]}\left\|\frac{W(r) - rW(1)}{\sqrt{V(r)}}\right\|, \tag{3.6}$$

where $V(r) = \int_0^r\{W(u) - u/rW(r)\}^2du + \int_r^1\{W(1) - W(u) - (1-u)/(1-r)[W(1) - W(r)]\}^2du$.

Now let's delve into the power of T_n. Under the alternative hypothesis H_a, we designate τ_0 as the unknown change point. Let $\Delta_n := EX_{\tau_0+1} - EX_{\tau_0}$ be the difference between the expected values before and after the change point. We consider the scenarios of a fixed alternative where $\Delta_n \equiv \Delta_0$ with Δ_0 being a nonzero element in H and local alternatives where $\Delta_n = n^{-1/2}\Delta_0$ as $n \to 0$.

Theorem 3.3. *Suppose assumptions in Theorem 3.1 hold. Let Δ_n be the mean difference before and after the change point under H_a. Under the fixed alternative with $\Delta_n = \Delta_0 \neq 0$, T_n diverges to ∞ in probability. Under local alternatives, if $\Delta_n = n^{-1/2}\Delta_0$ and Δ_0 is a*

nonzero element in H, then $\lim_{\|\Delta_0\|\to\infty} \lim_{n\to\infty} \|T_n\| = \infty$ *in probability.*

3.2.3 Non-overlapping Sequential Block Bootstrap

While we have derived the weak convergence result for our test statistic T_n in Theorem 3.2, using Monte Carlo simulations to simulate the corresponding asymptotic null distribution similar to Shao and Zhang [43] may not be realistic here. The unknown covariance operator S is an infinite dimensional parameter. An accurate approximation of such a covariance operator may not be possible through simple Monte Carlo simulations. Therefore, we consider generalizing the non-overlapping block bootstrap technique introduced by Sharipov et al. [44] here and aim at constructing a process whose limiting distribution is the same as $\frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor nr \rfloor} (X_i - \mu)$ in (3.2).

Given a block length p, we can identify $k = \lfloor n/p \rfloor$ blocks $I_j = \left(X_{(j-1)p+1}, \ldots, X_{jp}\right), j = 1, 2, \ldots, k$. To construct our bootstrap sample, we draw from these blocks k times independently and with replacement. The resulting selections, or the bootstrap blocks, satisfy $P\left(\left(X^*_{(j-1)p+1}, \ldots, X^*_{jp}\right) = I_i\right) = \frac{1}{k}, i, j = 1, \ldots, k$. From this, we can introduce a bootstrapped version of the partial sum process:

$$W^*_{n,p}(r) = \frac{1}{\sqrt{kp}} \sum_{t=1}^{\lfloor kpr \rfloor} \left(X^*_t - EX^*_t\right). \tag{3.7}$$

Sharipov et al. [44] proved the next result which establishes the asymptotic distribution of the process $W^*_{n,p}(r)$.

Theorem 3.4 (Theorem 1.2 of Sharipov et al. [44])**.** *Let $(X_n)_{n\in\mathbb{Z}}$ be L_1-near epoch dependent on a stationary and absolutely regular sequence $(\xi_n)_{n\in\mathbb{Z}}$ with $EX_1 = \mu$ and assume that for*

some $\delta > 0$,

$$E\left\|X_1\right\|^{4+\delta} < \infty, \quad \sum_{m=1}^{\infty} m^2 \left(a_m\right)^{\delta/(\delta+3)} < \infty, \quad \sum_{m=1}^{\infty} m^2 (\beta(m))^{\delta/(\delta+4)} < \infty.$$

Further, let the block length be nondecreasing, $p(n) = O\left(n^{1-\epsilon}\right)$ *for some* ϵ *and* $p(n) = p\left(2^l\right)$ *for* $n = 2^{l-1} + 1, \cdots, 2^l$, *for all* $l \in \mathbb{N}$. *Then*

$$\left(W_{n,p}^*(r)\right)_{r\in[0,1]} \Rightarrow (W(r))_{r\in[0,1]} \ a.s. \ ,$$

where $(W(r))_{r\in[0,1]}$ *is the Brownian motion defined in Theorem 3.1.*

Leveraging the findings from Theorem 3.4, We can derive the asymptotic distributions of $D_{n,p}^*(r)$ and $V_{n,p}^*(r)$, which are defined as

$$D_{n,p}^*(r) = \frac{1}{\sqrt{kp}} \sum_{t=1}^{\lfloor kpr \rfloor} \left(X_t^* - \bar{X}_n^*\right), \tag{3.8}$$

$$\begin{aligned} V_{n,p}^*(r) =& \frac{1}{(kp)^2} \Big[\sum_{t=1}^{\lfloor kpr \rfloor} \{A_{1,t}^* - (t/\lfloor kpr \rfloor) A_{1,\lfloor kpr \rfloor}^*\}^2 \\ &+ \sum_{t=\lfloor kpr \rfloor+1}^{kp} \{A_{t,kp}^* - (kp - t + 1)/(kp - \lfloor kpr \rfloor) A_{\lfloor kpr \rfloor+1,kp}^*\}^2 \Big]. \end{aligned} \tag{3.9}$$

Let $\tau = \lfloor kpr \rfloor$ and these definitions can be generalized to define $D_{n,p,\tau}^*$ and $V_{n,p,\tau}^*$. Then our bootstrapped test statistic T_n^* is

$$T_n^* = \sup_{1<\tau<kp-1} \left\| \frac{D_{n,p,\tau}^*}{\sqrt{V_{n,p,\tau}^*}} \right\|. \tag{3.10}$$

Corollary 3.5. *Under the conditions of Theorem 3.4, we have:*

$$T_n^* \Rightarrow \sup_{r\in[0,1]} \left\| \frac{W(r) - rW(1)}{\sqrt{V(r)}} \right\|. \tag{3.11}$$

Next, we derive the asymptotic distribution of the bootstrap test statistic under a sequence of local alternatives drifting towards the null at the rate of $n^{-1/2}$. This is different from evaluating tests against a fixed alternative. The goal is to ensure that the limiting distribution under local alternatives remains the same as the one under the null so that the critical values derived from the bootstrap test remain trustworthy.

Corollary 3.6. *If the conditions of Theorem 3.4 are satisfied, then under the local alternatives with $\Delta_n = n^{-1/2}\Delta_0$ for some nonzero $\Delta_0 \in H$, we have:*

$$T_n^* \Rightarrow \sup_{r\in[0,1]} \left\| \frac{W(r) - rW(1)}{\sqrt{V(r)}} \right\|, \tag{3.12}$$

Based on Theorem 3.3, Corollary 3.5, and Corollary 3.6, our bootstrap test has an asymptotically nontrivial power. Outlined below are the steps for the bootstrap test:

1. Calculate T_n.

2. Generate $T_{n,j}^*$ values for iterations $j = 1, ..., J$.

3. Based upon the independent random variables $T_{n,1}^*, ..., T_{n,J}^*$, determine the empirical $(1-\alpha)$-quantile, denoted as $q_{n,J}(\alpha)$.

4. If T_n exceeds $q_{n,J}(\alpha)$, the null hypothesis is rejected.

3.2.4 Testing for a Change Point in the Lag-1 autocovariance operator

In this part, we will explore an extension of our test, specifically focusing on assessing the stability of the autocovariance operator at lag one. This is crucial as it captures the inherent dependence structure within dependent functional data. Horváth et al. [20] introduced a test to evaluate the constancy of the FAR(1) (functional autoregressive model of order one) autocovariance operator within the context of a change point alternative. Zhang et al. [52] extended the process of the approximation of the Lag-1 autocovariance operator by checking the action of the Lag-1 autocovariance operator through the most important principle components of the observations. Our proposed test distinguishes itself from the aforementioned ones in several ways. Primarily, we employ the empirical approximation directly, bypassing the need for functional principal component analysis. Furthermore, our methodology is versatile enough to be extended, allowing for the detection of the change point in the Lag-d autocovariance operator for any $d \geq 1$.

Similar to Horváth et al. [20], the Lag-1 autocovariance operator at time t is formulated as:

$$R_{x,t} = E[\langle X_t, x\rangle X_{t+1}] = X_{t+1}(u) \int X_t(v)x(v)dv, u, v \in [0,1]$$

Our primary focus is to test the hypotheses

$$H_0 : R_{x,1} = R_{x,2} = \cdots = R_{x,n-1} \text{ versus } H_1 : R_{x,1} = \cdots = R_{x,\tau_0} \neq R_{x,\tau_0+1} = \cdots = R_{x,n-1}.$$

For this purpose, we introduce $R(u, v)$ as $E[X_t(u)X_{t+1}(v)]$ such that $R_{x,t} = \int R(u, v)x(v)dv$. The constancy of $R_{x,t}$ is approximately equivalent to the constancy of the new operator $R(u, v)$. Empirically, $R(u, v)$ can be represented as $\hat{R}(u, v) = X_t(u)X_{t+1}(v)$. Suppose there

are p dimensions for X_t. Then we define our new p^2-dimentional functional observation as $Y_t = [\hat{R}(1,1), \hat{R}(1,p), \hat{R}(2,1), ..., \hat{R}(2,p), ..., \hat{R}(p,1), ..., \hat{R}(p,p)]^T$. Subsequently, we can define $D_{n-1,\tau} = (n-1)^{-1/2} \sum_{t=1}^{\tau} \left(Y_t - \bar{Y}_{n-1}\right), \tau = 1, ..., n-2$, where $\bar{Y}_{n-1} = \frac{1}{n-1} \sum_{t=1}^{n-1} Y_t$ and $A_{t_1,t_2} = \sum_{t=t_1}^{t_2} Y_t$ if $t_1 \leq t_2$ and 0 otherwise. The self-normalization in this case is

$$V_{n-1,\tau} = \frac{1}{(n-1)^2} \Big[\sum_{t=1}^{\tau} \{A_{1,t} - (t/\tau) A_{1,\tau}\}^2 + \sum_{t=\tau+1}^{n-1} \{A_{t,n-1} - (n-t)/(n-\tau-1) A_{\tau+1,n-1}\}^2 \Big]. \tag{3.13}$$

And our test statistic is

$$G_{n-1} = \sup_{1<\tau<n-2} \left\| \frac{D_{n-1,\tau}}{\sqrt{V_{n-1,\tau}}} \right\|, \tag{3.14}$$

The null distribution for our test on the Lag-1 autocovariance operator can be derived similarly to that of the mean test.

Corollary 3.7. *Suppose Assumptions in Theorem 3.1 hold. Then under H_0, we have*

$$G_{n-1} \Rightarrow \sup_{r \in [0,1]} \left\| \frac{W(r) - rW(1)}{\sqrt{V(r)}} \right\|, \tag{3.15}$$

where $V(r) = \int_0^r \{W(u) - u/rW(r)\}^2 du + \int_r^1 \{W(1) - W(u) - (1-u)/(1-r)[W(1) - W(r)]\}^2 du$.

The bootstrap test statistic G^*_{n-1} based on the sequential bootstrap samples Y^*_t can be defined similarly, as well as the theoretical results corresponding to Theorem 3.4, Corollary 3.5, and Corollary 3.6. The test can be executed in the following steps:

1. Calculate and obtain $Y_t, t = 1, ..., n-1$.
2. Calculate G_{n-1}.
3. Generate $G^*_{n-1,j}$ values for iterations $j = 1, ..., J$.

4. Based upon the independent random variables $G^*_{n-1,1}, ..., G^*_{n-1,J}$, determine the empirical $(1-\alpha)$-quantile, denoted as $q_{n,J}(\alpha)$.

5. If G_{n-1} exceeds $q_{n,J}(\alpha)$, the null hypothesis is rejected.

3.3 Numerical Studies

We now report several simulation studies on our Functional Self-Normalization (FSN) test. In our simulations, we used 1000 replications, deriving critical values from a set of 500 bootstrap iterations. For the mean shift detection, we compared with the following methods, which, except for the first one, are all based on the FPCA:

1) **CUSUM**: A non-overlapping bootstrapped test proposed by Sharipov et al. [44] where X_t is considered to be the random functions in the Hilbert space relying solely on the CUSUM process.

2) **SN**: A self-normed CUSUM process based on the principle component score vectors for the sequences proposed by Zhang et al. [52].

3) **HOR**: A CUSUM-centric statistic by Horváth et al. [21] that projects sequences onto eigenfunctions, capturing component information explaining 90% of the variance.

4) **BGHK**: The CUSUM strategy employing score vectors with the diagonal matrix of eigenvalues as its long-run variance determinant [2].

5) **HK**: A CUSUM statistic based on the empirical score vectors with the dependence structure incorporated into the estimated LRV matrix [19].

For detecting shifts in the Lag-1 autocovariance, we compared with the approaches in Zhang et al. [52] and Horváth et al. [20], which are respectively denoted by **SN-lag** and **HHK**.

3.3.1 Detect the mean change in curves

In this section, we examine the finite sample properties of the FSN test for detecting a change in the mean function. A critical component of our method is the block length which determines the efficiency of the bootstrapping. To navigate through this, we experiment with multiple block lengths with the recommendations set forth by Carlstein [6] regarding the optimal length. As the primary performance metrics, the empirical sizes and powers are recorded to represent the efficacies of the tests. Furthermore, we account for the scenarios of independent or dependent data. For the independent scenario, our data generation process involves X_t drawn from independent and identically distributed standard Brownian motion (BM). For the dependent data scenarios, we derive functional sequences from the FAR(1) process defined as

$$X_t(s) = \int_0^1 \psi(s,u)X_{t-1}(u)du + \epsilon_t(s), \quad 0 \leq s \leq 1, \quad t = 1, ..., n, \tag{3.16}$$

Here, $\psi(s,u)$ represents the kernel function, and ϵ_t are independent Brownian Bridges in the interval $[0,1]$. To ensure the FAR(1) process remains stationary, the operator norm associated with ψ has to maintain a value less than one, that is, $\|\psi\|_{HS}^2 = \int_0^1 \int_0^1 \psi^2(s,u)dsdu < 1$, where $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm. Following Gabrys and Kokoszka [14], we use the Gaussian kernel $\psi(s,u) = C\ \exp(\frac{s^2+u^2}{2})$. In out simulations, we choose C to ensure $\|\psi\|_{HS}$ is approximately 0.5.

We consider functional time series of two lengths: $n = 100$ and $n = 150$. The percentage of rejections at the 5% and 10% significance levels are recorded. To simulate a change in the data, we introduced a mean function shift $f(s)$ after the change point τ_0, which is set to be $n/2$. Different choices of $f(s)$ corresponding to different alternative hypotheses H_a are considered:

(i) H_a^1: $f(s) = s$;

(ii) H_a^2: $f(s) = 0.3$

(iii) H_a^3: $f(s) = \begin{cases} 1, & s = [0, 0.05] \\ 0, & s = (0, 05, 1]; \end{cases}$

(iv) H_a^4: $f(s) = 1 - 3s$;

(v) H_a^5: $f(s) = 1 - 4s$.

N	Test	p	H_0	H_a^1	H_a^2	H_a^3	H_a^4	H_a^5
100	FSN	8	5.6	86.5	91.3	100	100	100
		10	4.4	86.2	89	100	100	100
		15	4.4	64.7	60.1	98.3	98.3	97.8
	CUSUM	8	3.3	92.2	45.9	28.2	74.3	56.4
		10	5.1	92.2	46.9	26.8	62.1	47.7
		15	7.7	85.7	41	27.7	64.3	54.3
	SN		4.8	81.2	37.6	4.2	5.2	7.3
	BGHK		5.1	98	46.4	3.5	7.4	8.3
	HK		4.2	95.7	41.4	3.6	5.1	6
	HOR		4.4	69.7	34.6	32.2	68.2	53.4
150	FSN	12	5.8	93.3	97.9	100	100	100
		15	5.1	90.3	94.7	95.3	100	100
		18	4.9	86.3	91.8	100	100	100
	CUSUM	12	3.6	99	63.3	37.1	95.1	81.7
		15	2.6	98.6	55.6	37.3	91.8	75.3
		18	3.3	91.7	51.6	37.3	92.3	74.1
	SN		5.8	97.2	49.6	4.2	6.5	6.6
	BGHK		6.4	99.8	64.1	4.8	7.2	7.8
	HK		4.4	99.4	58.4	4.2	7.1	6.8
	HOR		4.4	92.8	53.5	53.8	93.6	86.5

Table 3.1: Empirical size and power results in percentage for tests for the Brownian motion data when $\alpha = 0.05$

Tables 3.1-3.4 show the empirical sizes and powers of the tests for the independent (BM) and dependent (FAR(1)) data scenarios. Based on these results, we have the following

N	Test	p	H_0	H_a^1	H_a^2	H_a^3	H_a^4	H_a^5
100	FSN	8	9.4	89.3	98.6	100	100	100
		10	9.8	89.1	97.4	100	100	100
		15	11.3	83.4	88.2	100	100	100
	CUSUM	8	11.1	98.4	59.7	42.6	88.6	75.4
		10	12	97.8	65.2	42.2	80.6	64.3
		15	15.3	97.1	60.6	45.3	82.6	76.7
	SN		9.4	96.4	51.5	9.6	12.3	11.8
	BGHK		12.4	99.2	51.5	8.1	12.2	14.4
	HK		12	98.4	56.1	9.5	11.2	14.3
	HOR		11.8	92.2	60.4	60.1	92.6	85.4
150	FSN	12	11.1	96.4	99.3	100	100	100
		15	12.5	96.8	99.3	100	100	100
		18	12.3	95.4	98.5	100	100	100
	CUSUM	12	11.2	100	79.2	59.9	99.3	92.3
		15	7.6	100	74.1	64.3	98.2	91.6
		18	12.7	100	73.6	61	99.3	90.2
	SN		8.6	99.2	64.4	9.7	11.1	12.8
	BGHK		9.8	100	75.7	9.6	15.1	14
	HK		9.2	100	72.4	9.6	13.6	13.8
	HOR		10.6	99.4	79	83.5	99.2	97.9

Table 3.2: Empirical size and power results in percentage for tests for the Brownian motion data when $\alpha = 0.1$

conclusions:

- In terms of size controls, the FSN test consistently exhibits superior performance, irrespective of the underlying dependence structures in the data. The SN and HOR tests yield accurate sizes but are not as good as the FSN test. Notably, the CUSUM, BGHK, and HK tests exceed the alpha levels for the FAR(1) data, indicating inflated size performance. Their size control performances improve on the Brownian motion data, although the sizes yielded are slightly lower than the alpha levels. This variation underscores the fluctuating performance of these tests across diverse types of datasets.

- In terms of empirical powers, the FSN, CUSUM, and HOR tests outperform the other

N	Test	p	H_0	H_a^1	H_a^2	H_a^3	H_a^4	H_a^5
100	FSN	8	4.7	98.5	54.1	91.5	99.2	97.4
		10	5	96.2	58.1	90.6	91.6	93.4
		15	5.1	85.5	37.4	72.5	82.5	78.3
	CUSUM	8	8.8	99.5	62	52.4	92.9	85.8
		10	5.7	95.2	54.2	45.3	97.7	85.6
		15	10.3	94.5	51.6	45.5	85.1	76.4
	SN		7.1	92.6	57.6	9.8	20.2	14.2
	BGHK		31.2	99.8	92.9	38.8	53.4	42.8
	HK		10.8	98.6	67.4	10.4	24.8	16
	HOR		4.2	98	38.3	36	73.1	61.3
150	FSN	12	4.4	98.7	74	96.1	98.9	98.6
		15	5	98.7	70.7	94	99.3	98.8
		18	4.1	94.7	62	91.3	96.2	96.3
	CUSUM	12	9.2	100	80.6	67.3	98.8	95.4
		15	7.1	100	69.3	64.4	99.2	95.9
		18	8.4	100	68.6	67.5	98.5	99.3
	SN		6.2	99.3	64.8	7.9	22.2	12.3
	BGHK		30.2	100	95.4	39.7	57.4	43.6
	HK		8	99.8	77.4	12	26.4	16.1
	HOR		2.4	98.4	54.1	69.3	97.2	90.2

Table 3.3: Empirical size and power results in percentage for tests for the FAR(1) model with the Gaussian kernel when $\alpha = 0.05$

methods in terms of magnitude and stability under different alternatives when dealing with both the Brownian motions data and FAR(1) data. In particular, the FSN test has nearly perfect empirical powers under almost all alternative scenarios. Although the CUSUM and HOR tests exhibit strong performance under the first two alternatives, they fail to catch up with the FSN test under alternatives involving sparse or decreasing change functions. Similarly, the SN, BGHK, and HK tests exhibit strong empirical powers under the first two alternatives. But their powers under alternatives with sparse and decreasing change functions are also substantially lower than the FSN. In the case of Brownian motions data, their powers even drop down to only single-digit percentages.

N	Test	p	H_0	H_a^1	H_a^2	H_a^3	H_a^4	H_a^5
100	FSN	8	10.3	100	73.7	98.1	100	100
		10	9.8	99.3	78.5	97.3	100	100
		15	10.1	94.5	60.2	90.5	96.7	94.1
	CUSUM	8	16.2	100	74.1	75.3	97.5	94.7
		10	16.3	97.3	77.2	63.2	93.1	97.5
		15	18	99.1	70.8	70.5	96.5	88.9
	SN		11.6	96	66.6	15.6	28.6	21.6
	BGHK		44.2	100	94.8	51.4	64.1	55.7
	HK		18.5	99.2	80.7	21.2	36.4	24.9
	HOR		10.2	99.1	64.3	72.1	96.8	88.4
150	FSN	12	10.5	100	85.3	98.4	100	100
		15	9.1	100	86.1	98.7	100	100
		18	10.9	99.5	83.3	96.5	98.7	99.3
	CUSUM	12	15.4	100	91.2	88	100	99.2
		15	15.3	100	89.2	84.16	100	99.3
		18	16.4	100	86.7	82.6	100	100
	SN		10.1	99.5	75.7	16.2	30.8	18.4
	BGHK		41.8	100	97.2	53.3	67	57.6
	HK		14.3	99.8	88.6	21.5	38.3	24.8
	HOR		9.2	99.8	78.7	92.4	99.8	99

Table 3.4: Empirical size and power results in percentage for tests for the FAR(1) model with the Gaussian kernel when $\alpha = 0.1$

Also, it's not hard to tell the influence of the change function on test outcomes. In real-world analogs, change functions are unpredictable and can be of diverse types, our FSN test exhibits superior adaptability. Conversely, the efficacy of other tests diminishes markedly under some alternatives. This advantage of the FSN test can be ascribed to its self-normalization component which calibrates the significance of the change relative to the variance distribution across time, making the change more noticeable in areas with little variability. This is nontrivial since the monotone trend on variance distribution could be a common feature in standard functional datasets. Figure 3.1 clarifies this effect by depicting the variance distribution over time for both our generated FAR(1) model and the Brownian motion data.

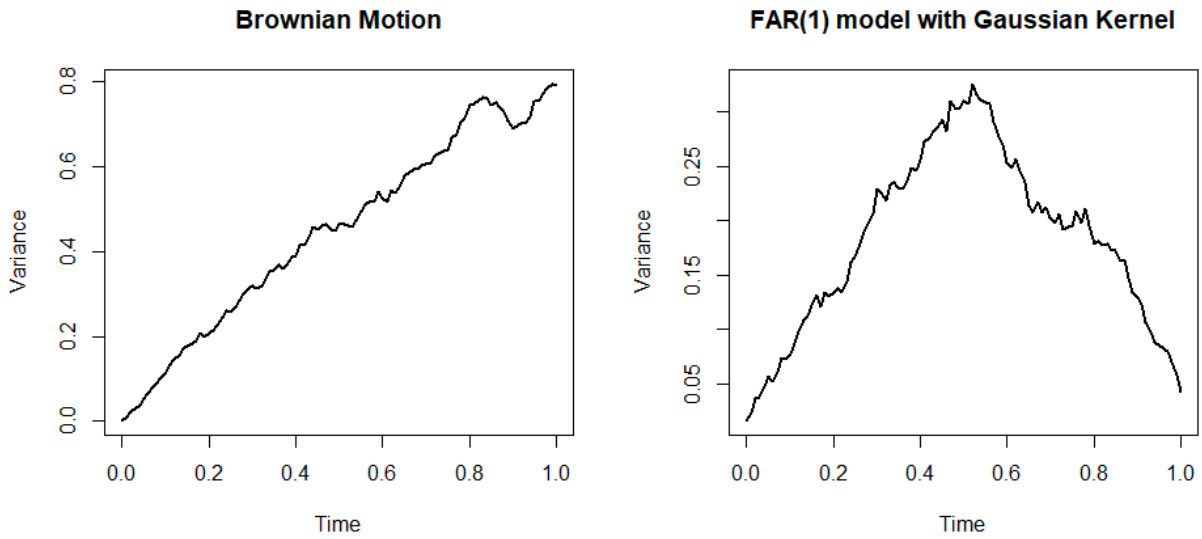

Figure 3.1: The variance functions of the standard Brownian motions and the FAR(1) model with Gaussian kernel across the time.

Next, we consider the non-constant covariance structure before and after the change point. We generate the functional sequences X_t from Gaussian processes. When there is no change point, the generating Gaussian process has mean 0 and covariance function $C_1(s,u) = \sum_{k\geq 1} 2k^{-2}\cos(k\pi s)\cos(k\pi u)$. When there is a change point at $t = n/2$, the functional data are generated from

$$X_t(s) \sim \begin{cases} N(\mathbf{0}, C_1(s,u)), & t = 1, ..., \frac{n}{2}, \\ N(\mu(\mathbf{s}), C_2(s,u)), & t = \frac{n}{2}+1, ..., n, \end{cases} \tag{3.17}$$

where $\mu(s) = s$, $C_2(s,u) = \sum_{k\geq 1} 2\theta_k \cos(k\pi s)\cos(k\pi u)$ and $\theta_k = (|k-k_0|+1)^{-2}$. In this setting, the leading eigenfunctions of $C_2(s,u)$ are located around the k_0th eigenfunction and in a certain sense control the misalignment between the covariance functions of $X_t(s)$ in the first and second halves of the sequence. Tables 3.5 and 3.6 reveal that the FSN test stands out as the only performer exhibiting ideal empirical size and power. While the CUSUM and

k_0	FSN	CUSUM	SN	BGHK	HK	HOR
20	4.6	6.2	10	9.3	7.9	2.7
40	4.2	7.2	12	5.4	6	2.6
60	4.3	6.7	15.4	9.6	7.5	66.6
20	80.6	88.4	48.2	55.1	52.6	68
40	83	89.3	50.5	51.1	49.3	72.6
60	80.3	86	47.6	48.1	44.6	99

Table 3.5: Empirical Size and Power in percentages for the functional data with non-constant covariance structure when $\alpha = 0.05$

k_0	FSN	CUSUM	SN	BGHK	HK	HOR
20	8.3	15.4	14	11.3	12.6	8.6
40	8.8	14.1	18.7	17.3	11.9	15.4
60	8.1	14.2	24	16.8	14.6	84.4
20	94.3	98	60.6	66	67.2	90.2
40	96.2	95.3	64.4	68.7	66	96
60	91.6	96.2	59.3	61	60.4	100

Table 3.6: Empirical Size and Power in percentages for the functional data with non-constant covariance structure when $\alpha = 0.1$

HOR tests have good powers, their size controls are not as good. This poor performance in size control is particularly apparent for the HOR test when k_0 reaches 60, and the CUSUM test consistently reports sizes exceeding the alpha level. In comparison, the SN, BGHK, and HK tests fall short on both fronts, with rejection rates higher than the alpha level under the null and powers lower than the FSN and CUSUM tests. This simulation study highlights the limitations of FPCA-based methods in handling data with misaligned covariance functions.

3.3.2 Detect the change in Lag-1 autocovariance operator

In this subsection, we explore the finite sample performance of our proposed FSN test on the detection of shifts in the Lag-1 autocovariance operator. We generate functional sequences based on a mean-zero FAR(1) model, employing the aforementioned Gaussian kernel with various Hilbert-Schmidt norm values.

When the null hypothesis is true, all the functional observations are from the same mean-zero FAR(1) model. When the alternative hypothesis is true, we adopt the following data-generating process:

$$\begin{cases} X_t(s) = \int_0^1 \psi_1(s,u) X_{t-1}(u) du + \epsilon_t, & t = 1, \ldots, n/2 \\ X_t(s) = \int_0^1 \psi_2(s,u) X_{t-1}(u) du + \epsilon_t, & t = n/2+1, \ldots, n. \end{cases} \tag{3.18}$$

Here, $\Psi_1(s,u)$ and $\Psi_2(s,u)$ are both Gaussian kernels. We consider two settings for their norms:

(i) $\|\psi_1\| = 0.3$, $\|\psi_2\| = 0.8$;

(ii) $\|\psi_1\| = 0.1$, $\|\psi_2\| = \begin{cases} 0.9 - 3su, & 0 \le su \le 0.2 \\ 0.3, & 0.2 < su \le 1 \end{cases}$.

We denote the hypotheses corresponding to these two settings respectively by H_0^1, H_a^1 and H_0^2, H_a^2. Errors are generated from the independent Brownian bridges on $[0,1]$. Two sample sizes, $n = 100$ and $n = 150$, are considered. The results are in Tables 3.7 and 3.8.

In terms of size control, most tests perform adequately with the notable exception of the CUSUM test. Its empirical sizes are higher than the alpha levels under H_0^1, and lower than the alpha levels under H_0^2. On the other hand, the powers of the CUSUM test under H_a^1

N	Test	p	H_0^1	H_0^2	H_a^1	H_a^2
100	FSN	8	5.9	5.4	41.3	33.4
		10	5.1	4.6	29.6	26.6
		15	4.1	6.5	22.1	86.3
	CUSUM	8	8.9	8.5	46.2	21.9
		10	6.3	6.6	27.1	14.1
		15	8.1	9.3	40.4	72.8
	SN-lag		5.3	3.8	30.6	12.5
	HHK		7.2	5.6	42.2	16.9
150	FSN	12	5.7	5.1	46.7	41.5
		15	4.6	4.9	41.8	46.8
		18	5.6	4.6	52.5	34.8
	CUSUM	12	6.4	3.3	61.8	24.7
		15	7.1	2.8	36.7	16.2
		18	8.4	3	39.1	22.1
	SN-lag		4.6	3.6	44.9	15.8
	HHK		4.7	5.2	66.4	23.6

Table 3.7: Empirical size and power results in percentage for tests on the FAR(1) model with Gaussian kernel on detecting the lag-1 autocovariance when $\alpha = 0.05$

align well with the HHK test, both providing powers more than 0.5. There aren't significant differences between the powers of the tests under H_a^1, with each achieving powers between 0.4 and 0.7. Note that the FSN test performs better than the others, delivering considerably higher powers under H_a^2.

Another important observation is the influence of block length on bootstrapped tests. Specifically, for the FSN and CUSUM tests under H_a^2, power values exceeding 0.7 are attained when the block length is set to 15 and $n = 100$. Interestingly, this pronounced effect diminishes when $n = 150$, resulting in comparable power values across different block lengths.

N	Test	p	H_0^1	H_0^2	H_a^1	H_a^2
100	FSN	8	10.8	9.3	49.6	56.5
		10	10.3	9	47.8	44.2
		15	8.7	11.1	41.7	88.8
	CUSUM	8	13.4	15.1	72.3	35.8
		10	11.2	15.3	50.2	26.9
		15	12	18.4	58.2	73.4
	SN-lag		9.5	8	44.7	21.4
	HHK		10.2	11.4	55.5	28.7
150	FSN	12	10.4	10.6	58.9	62.6
		15	10.6	10.1	55.7	63.5
		18	10.9	10.2	54	64.4
	CUSUM	12	15.4	6.4	84	36.6
		15	15.3	7.6	67.6	38.8
		18	16.4	6.8	60	44.8
	SN-lag		8.9	8.4	56.5	24.2
	HHK		10.1	9.3	77.7	36.4

Table 3.8: Empirical size and power results in percentage for tests on the FAR(1) model with Gaussian kernel on detecting the lag-1 autocovariance when $\alpha = 0.1$

3.4 Real-life data examples

3.4.1 Annual temperature profiles of Fairbanks, Alaska

We first apply our testing method for a mean change point to the annual temperature profile data of the Fairbanks area in Alaska (64.83°N, 147.77°W) from 1907 to 2022. Each year's profile consists of the 12 monthly highest temperatures and we have 116 curves corresponding to the years of 1907 to 2022. We are interested in determining whether there was any mean change point in these annual temperature profiles.

Figure 3.2 depicts the application of our FSN test to the data. In the left panel, we plotted all the values of $\left\| D_{n,\tau} \Big/ \sqrt{V_{n,\tau}} \right\|$ for $\tau = 1907, \ldots, 2022$. The test statistic is their maximum, which appeared at year 1972. Its value surpasses the bootstrapped 5% significance level with

an empirical p-value close to zero, suggesting a mean change in the temperature profiles at year 1972. The right panel of Figure 3.2 highlights this change by comparing the mean temperature profiles for the years up to and after 1972.

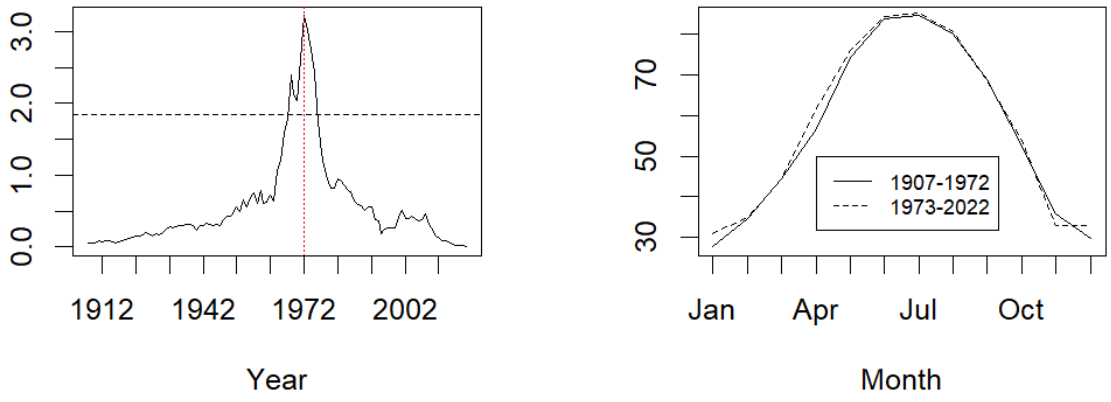

Figure 3.2: Annual temperatures of Fairbanks, Alaska in 1907-2022. Left panel: Plot of $\left\| D_{n,\tau} \,/\, \sqrt{V_{n,\tau}} \right\|$ for $\tau = 1907, \ldots, 2022$ (solid line), the 5% significance level (dashed line) computed from 500 bootstrap iterations, and the year of the detected mean profile change point (vertical red line). Right panel: Mean annual temperature profiles of the time periods 1907-1972 (solid line) and 1973-2022 (dashed line).

3.4.2 Global surface temperatures from NASA GISS

We now turn our attention to the global surface temperatures recorded in the NASA Goddard Institute for Space Studies (GISS) Surface Temperature Analysis (GISTEMP v4). This dataset contains the monthly Land-Ocean Temperature Indices (L-OTI) from 1880 to 2022. The L-OTI is a measure of how global average temperatures have changed over long periods of time and reflects temperature anomalies [25]. A temperature anomaly is how much warmer or cooler a particular time was compared to a 30-year average defined by the U.S. National Weather Service as "normal" temperature. For this data set, the 30-year average is computed

for the period 1951-1980. To remove the seasonality effect within the annual profiles, we followed the approach in Horváth et al. [20] to consider the time series of the 142 differences between the profiles of consecutive years, indexed by years 1881 to 2022. The first twenty profile differences are illustrated in the left panel of Figure 3.3. A characteristic pattern of an FAR(1) process with clusters of positive and negative observations is clearly seen. When applied to this functional time series, our FSN test on lag-1 auto-covariance fails to reject the null hypothesis, indicating that the FAR(1) model is indeed appropriate for the time series. As illustrated in the right panel of Figure 3.3, the quantity $\left\| D_{n,\tau} \big/ \sqrt{V_{n,\tau}} \right\|$ reaches its maximum in 1953, which lies substantially below the critical thresholds determined through the bootstrap procedure with an empirical p-value of 0.38. This confirms the stationarity of the Lag-1 autocovariance or the validity of the FAR(1) model for this time series.

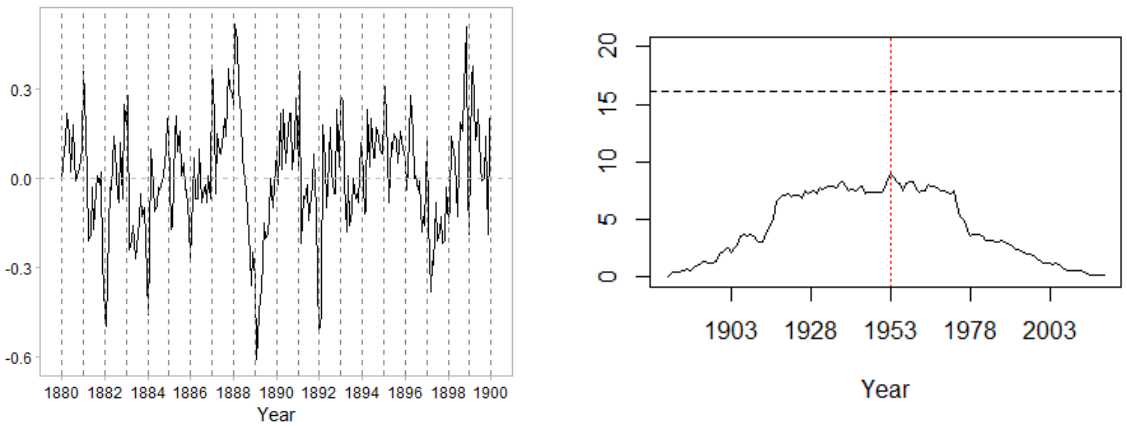

Figure 3.3: Land-Ocean Temperature Index differences from NASA GISS for 1880-2022. Left panel: the first 20 annual profile differences. Right panel: Plot of $\left\| D_{n,\tau} \big/ \sqrt{V_{n,\tau}} \right\|$ for $\tau = 1881, \ldots, 2022$ (solid line), the 5% significance level (dashed line) computed from 500 bootstrap iterations, and the year of the maximum $\left\| D_{n,\tau} \big/ \sqrt{V_{n,\tau}} \right\|$ (vertical red line).

3.5 Discussion

Mean and Lag-specific autocovariance change detection have been extensively studied in the literature of functional time series analysis. In this paper, we present a novel test for assessing the stationarity of functional time sequences, which generalizes the self-normalization test for univariate time series to the functional setting. We derive the limiting distribution of the test statistic under the null hypothesis and establish the test consistency under the alternative hypothesis. Calculating critical values poses a formidable challenge due to the estimation complexity of the covariance structure. To address this issue, we propose the use of a non-overlapping bootstrap method. Importantly, we demonstrate that the asymptotic null distribution and consistency of the bootstrapped version of our test align with those of the original method.

In our simulation studies, we compared our test against other established techniques that rely on bootstrap methods and FPCA. Our primary goal was to detect shifts in mean change and changes in the Lag-1 autocovariance. Our studies show that our method can reliably identify these changes in scenarios when the function changes are sparse or have a decreasing trend over the domain. Our FSN test performs much better than the others in terms of its accuracy and sensitivity to the change functions, highlighting its adaptability and broad applicability. To further validate our method, we applied it to real-world data, and the results were consistent, emphasizing its practical value in applied settings.

However, our test has some issues that require further exploration. Using the bootstrap method can significantly increase computation time, which is a critical consideration when choosing a testing approach. Additionally, the optimal selection of the block length is a complex task and can influence the effectiveness of our method. These directions all merit further research.

Bibliography

[1] Carcinoembryonic antigen. https://www.ncbi.nlm.nih.gov/books/NBK578172/. Accessed: 202-01-26.

[2] István Berkes, Robertas Gabrys, Lajos Horváth, and Piotr Kokoszka. Detecting changes in the mean of functional observations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):927–946, 2009.

[3] Leo Breiman. Discussion: Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):82–91, 1991. ISSN 00905364. URL http://www.jstor.org/stable/2241839.

[4] Beatrice Bucchia and Martin Wendler. Change-point detection and bootstrap for hilbert space valued random fields. *Journal of Multivariate Analysis*, 155:344–368, 2017.

[5] Tony Cai and Ming Yuan. Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107(499):1201–1216, 2012.

[6] Edward Carlstein. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The annals of statistics*, pages 1171–1179, 1986.

[7] Baojiang Chen, Pengfei Li, Jing Qin, and Tao Yu. Using a monotonic density ratio model to find the asymptotically optimal combination of multiple diagnostic tests. *Journal of the American Statistical Association*, 111(514):861–874, 2016. doi: 10.1080/01621459.2015.1066681. URL https://doi.org/10.1080/01621459.2015.1066681.

[8] Guang Cheng and Zuofeng Shang. Local and global asymptotic inference in smoothing spline models. *Ann Stat.*, 41(5):2608–2638, 2013. ISSN 0090-5364.

[9] Miklós Csörgö and Lajos Horváth. Limit theorems in change-point analysis. 1997.

[10] Herold Dehling, Olimjon Sh. Sharipov, and Martin Wendler. Bootstrap for dependent hilbert space-valued random variables with application to von Mises statistics. *Journal of Multivariate Analysis*, 133:200–215, 2015.

[11] Ai Deng and Pierre Perron. A non-local perspective on the power properties of the cusum and cusum of squares tests for structural change. *Journal of Econometrics*, 142 (1):212–240, 2008.

[12] Jianqing Fan, Yang Feng, and Rui Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557, 2011.

[13] Youyi Fong, Shuxin Yin, and Ying Huang. Combining biomarkers linearly and nonlinearly for classification using the area under the roc curve. *Statistics in medicine*, 35, 04 2016. doi: 10.1002/sim.6956.

[14] Robertas Gabrys and Piotr Kokoszka. Portmanteau test of independence for functional observations. *Journal of the American Statistical Association*, 102(480):1338–1348, 2007.

[15] I. J. Good and R. A. Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58(2):255–277, 1971. ISSN 00063444. URL http://www.jstor.org/stable/2334515.

[16] Chong Gu. Smoothing spline density estimation: A dimensionless automatic algorithm. *Journal of the American Statistical Association*, 88(422):495–504, 1993. doi: 10.1080/01621459.1993.10476300. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476300.

[17] Chong Gu. *Smoothing spline ANOVA models*, volume 297. Springer, 2013.

[18] Chong Gu and Chunfu Qiu. Smoothing Spline Density Estimation: Theory. *The Annals of Statistics*, 21(1):217 – 234, 1993. doi: 10.1214/aos/1176349023. URL https://doi.org/10.1214/aos/1176349023.

[19] Siegfried Hörmann and Piotr Kokoszka. Weakly dependent functional data. *The Annals of Statistics*, 38(3):1845–1884, 2010.

[20] Lajos Horváth, Marie Hušková, and Piotr Kokoszka. Testing the stability of the functional autoregressive process. *Journal of Multivariate Analysis*, 101(2):352–367, 2010.

[21] Lajos Horváth, Piotr Kokoszka, and Gregory Rice. Testing stationarity of functional time series. *Journal of Econometrics*, 179(1):66–82, 2014.

[22] V. K. Klonias. Consistency of Two Nonparametric Maximum Penalized Likelihood Estimators of the Probability Density Function. *The Annals of Statistics*, 10(3):811 – 824, 1982. doi: 10.1214/aos/1176345873. URL https://doi.org/10.1214/aos/1176345873.

[23] Charles Kooperberg and Charles J. Stone. A study of logspline density estimation. *Comput Stat Data Anal.*, 12(3):327–347, 1991. ISSN 0167-9473.

[24] SK Lahiri and SN Lahiri. *Resampling methods for dependent data*. Springer Science & Business Media, 2003.

[25] Nathan J. L. Lenssen, Gavin A. Schmidt, James E. Hansen, Matthew J. Menne, Avraham Persin, Reto Ruedy, and Daniel Zyss. Improvements in the gistemp uncertainty model. *Journal of Geophysical Research: Atmospheres*, 124(12):6307–6326, 2019. doi: https://doi.org/10.1029/2018JD029522. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018JD029522.

[26] Tom Leonard. Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(2):113–146, 1978. ISSN 00359246. URL http://www.jstor.org/stable/2984749.

[27] Chunling Liu, Aiyi Liu, and Susan Halabi. A min-max combination of biomarkers to improve diagnostic accuracy. *Statistics in medicine*, 30:2005–14, 07 2011. doi: 10.1002/sim.4238.

[28] Danping Liu, Yongli Han, and Aiyi Liu. Marginal, conditional, and pseudo likelihood ratio approaches for biomarker combination to predict a binary disease outcome. *Statistics in Medicine*, 41(14):2574–2585, 2022.

[29] Yufeng Liu, Xiaotong Shen, and Hani Doss. Multicategory ψ-learning and support vector machine: computational tools. *Journal of Computational and Graphical Statistics*, 14(1):219–236, 2005.

[30] Shuangge Ma and Jian Huang. Combining multiple markers for classification using roc. *Biometrics*, 63(3):751–757, 2007.

[31] Martin W. McIntosh and Margaret Sullivan Pepe. Combining several screening tests: Optimality of the risk score. *Biometrics*, 58(3):657–664, 2002. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/3068590.

[32] James Mills, Mary Hediger, Cynthia Molloy, George Chrousos, Patricia Manning-Courtney, Kai Yu, Mark Brasington, and Lucinda England. Elevated levels of growth-related hormones in autism and autism spectrum disorder. *Clinical endocrinology*, 67: 230–7, 09 2007. doi: 10.1111/j.1365-2265.2007.02868.x.

[33] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A,*

Containing Papers of a Mathematical or Physical Character, 231:289–337, 1933. ISSN 02643952. URL http://www.jstor.org/stable/91247.

[34] Finbarr O'Sullivan. Fast computation of fully automated log-density and log-hazard estimators. *Siam Journal on Scientific and Statistical Computing*, 9:363–379, 1988.

[35] M Pepe and M Thompson. Combining diagnostic test results to increase accuracy. *Biostatistics (Oxford, England)*, 1:123–40, 07 2000. doi: 10.1093/biostatistics/1.2.123.

[36] Margaret Sullivan Pepe, Tianxi Cai, and Gary Longton. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, 62 (1):221–229, 2006. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/3695724.

[37] Pierre Perron et al. Dealing with structural breaks. *Palgrave handbook of econometrics*, 1(2):278–352, 2006.

[38] Jing Qin and Biao Zhang. Best combination of multiple diagnostic tests for screening purposes. *Statistics in medicine*, 29:2905–19, 12 2010. doi: 10.1002/sim.4068.

[39] J. O. Ramsay and B. W. Silverman. *Functional data analysis (2nd Ed.)*. New York, NY: Springer Science+Business Media, Inc.; 2005.

[40] Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71 (5):1009–1030, 2009. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/40541567.

[41] Xiaofeng Shao. The dependent wild bootstrap. *Journal of the American Statistical Association*, 105(489):218–235, 2010.

[42] Xiaofeng Shao. A self-normalized approach to confidence interval construction in time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72 (3):343–366, 2010.

[43] Xiaofeng Shao and Xianyang Zhang. Testing for change points in time series. *Journal of the American Statistical Association*, 105(491):1228–1240, 2010.

[44] Olimjon Sharipov, Johannes Tewes, and Martin Wendler. Sequential block bootstrap in a hilbert space with application to change point analysis. *Canadian Journal of Statistics*, 44(3):300–322, 2016.

[45] B. W. Silverman. On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method. *The Annals of Statistics*, 10(3):795 – 810, 1982. doi: 10.1214/aos/1176345872. URL https://doi.org/10.1214/aos/1176345872.

[46] John Q. Su and Jun S. Liu. Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, 88(424):1350–1355, 1993. ISSN 01621459. URL http://www.jstor.org/stable/2291276.

[47] Timothy J Vogelsang. Testing for a shift in mean without having to estimate serial-correlation parameters. *Journal of Business & Economic Statistics*, 16(1):73–80, 1998.

[48] Grace Wahba. Data-Based Optimal Smoothing of Orthogonal Series Density Estimates. *The Annals of Statistics*, 9(1):146 – 156, 1981. doi: 10.1214/aos/1176345341. URL https://doi.org/10.1214/aos/1176345341.

[49] Lea Wegner and Martin Wendler. Robust change-point detection for functional time series based on u-statistics and dependent wild bootstrap. *arXiv preprint arXiv:2206.01458*, 2022.

[50] Qingxiang Yan, Leonidas E Bantis, Janet L Stanford, and Ziding Feng. Combining multiple biomarkers linearly to maximize the partial area under the roc curve. *Statistics in medicine*, 37(4):627–642, 2018.

[51] Jingjing Yin and Lili Tian. Optimal linear combinations of multiple diagnostic biomarkers based on youden index. *Statistics in medicine*, 33(8):1426–1440, 2014.

[52] Xianyang Zhang, Xiaofeng Shao, Katharine Hayhoe, and Donald J Wuebbles. Testing the structural stability of temporally dependent functional observations and application to climate projections. *Electronic Journal of Statistics*, 5:1765–1796, 2011.

[53] Li-Ping Zhu, Lexin Li, Runze Li, and Li-Xing Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496): 1464–1475, 2011.

Appendices

Appendix A

First Appendix

A.1 Theorem 3.2

By Theroem 3.1, we are going to derive the null distribution of our Functional Self-Normalized (FSN) test statistics T_n, whose definition involves $D_{n,\tau}$ and $V_{n,\tau}$. First,

$$\begin{aligned} D_{n,\tau} &= \frac{1}{\sqrt{n}} \sum_{t=1}^{\tau} \{X_t - \bar{X}_n\} \\ &= \frac{1}{\sqrt{n}} \sum_{t=1}^{\tau} X_t - \frac{\tau}{n} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} X_t \\ &= \frac{1}{\sqrt{n}} \sum_{t=1}^{\tau} (X_t - \mu) - \frac{\tau}{n} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (X_t - \mu). \end{aligned} \tag{A.1}$$

By Theorem 3.1 and the continuous mapping theorem, we have $D_{n,\tau} \Rightarrow W(r) - rW(1)$. Now

$$V_{n,\tau} = \frac{1}{n^2} \Big[\sum_{t=1}^{\tau} \{A_{1,t} - (t/\tau) A_{1,\tau}\}^2 \tag{A.2}$$

$$+ \sum_{t=\tau+1}^{n} \{A_{t,n} - (n-t+1)/(n-\tau) A_{\tau+1,n}\}^2 \Big], \tag{A.3}$$

The first component (A.2) can be rewritten as

$$\frac{1}{n} \sum_{t=1}^{\tau} \{\frac{1}{\sqrt{n}} A_{1,t} - \frac{1}{\sqrt{n}} (t/\tau) A_{1,\tau}\}^2 = \frac{1}{n} \sum_{t=1}^{\tau} \{\frac{1}{\sqrt{n}} \sum_{t=1}^{t} (X_t - \mu) - \frac{1}{\sqrt{n}} \frac{t}{\tau} \sum_{t=1}^{\tau} (X_t - \mu)\}^2. \tag{A.4}$$

By Theorem 3.1, converges in distribution to $\frac{1}{n}\sum_{t=1}^{\tau}\{W(\frac{t}{n}) - \frac{t}{\tau}W(\frac{\tau}{n})\}^2$, whose limit is $\int_0^r\{W(u) - u/rW(r)\}^2du$ as $n \to \infty$. Similarly, for the second component (A.3), we have

$$\frac{1}{n}\sum_{t=\tau+1}^{n}\{A_{t,n} - \frac{n-t+1}{n-\tau}A_{\tau+1,n}\}^2 \Rightarrow \int_r^1\{W(1) - W(u) - \frac{1-u}{1-r}[W(1) - W(r)]\}^2du. \quad \text{(A.5)}$$

Therefore, $V_{n,\tau} \Rightarrow V(r)$, where $V(r) = \int_0^r\{W(u) - u/rW(r)\}^2du + \int_r^1\{W(1) - W(u) - (1 - u)/(1 - r)[W(1) - W(r)]\}^2du$. Now we get the converging distributions of $D_{n,\tau}$ and $V_{n,\tau}$. Then applying Slutsky's Theorem and the continuous mapping Theorem, we get

$$T_n \Rightarrow \sup_{r\in[0,1]}\left\|\frac{W(r) - rW(1)}{\sqrt{V(r)}}\right\|.$$

A.2 Theorem 3.3

Note that under H_a in Theorem 3.3, $EX_t = \mu$ for $t \le \tau_0$ and $EX_t = \mu + \Delta_n$ for $t \ge \tau_0 + 1$. Define a new sequence $Z_t, t = 1, \ldots, n$ such that $Z_t = X_t$ when $t \le \tau_0$ and $Z_t = X_t - \Delta_n$ when $t \ge \tau_0 + 1$. Now both $\{X_t\}_{t=1}^{\tau_0}$ and $\{Z_t\}_{t=1}^{n}$ satisfy the conditions of Theorem 3.1. Note that D_{n,τ_0} can be written as

$$\begin{aligned} D_{n,\tau_0} &= \frac{1}{\sqrt{n}}\sum_{t=1}^{\tau_0}\{X_t - \bar{X}_n\} \\ &= \frac{1}{\sqrt{n}}\sum_{t=1}^{\tau_0}X_t - \frac{\tau_0}{n}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}X_t \\ &= \frac{1}{\sqrt{n}}\sum_{t=1}^{\tau_0}(X_t - \mu) - \frac{\tau_0}{n}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(Z_t - \mu) - \frac{\tau_0(n-\tau_0)}{n^{3/2}}\Delta_n. \end{aligned} \quad \text{(A.6)}$$

Then Theorem 3.1 gives

$$D_{n,\tau_0} \Rightarrow W(r) - rW(1) - r(1-r)n^{1/2}\Delta_n, \quad \text{where } \tau_0 = \lfloor nr \rfloor. \quad \text{(A.7)}$$

Also, note that V_{n,τ_0} does not depend on Δ_n. So we have

$$T_n \geq \left\| \frac{D_{n,\tau_0}}{\sqrt{V_{n,\tau_0}}} \right\| \Rightarrow \left\| \frac{W(r) - rW(1) - r(1-r)n^{1/2}\Delta_n}{\sqrt{V(r)}} \right\|. \tag{A.8}$$

Therefore, under the fixed alternative with $\Delta_n = \Delta_0 \neq 0$ or the local alternatives with $\Delta_n = n^{-\frac{1}{2}}\Delta_0$, we have $T_n \geq \left\| \frac{D_{n,\tau_0}}{\sqrt{V_{n,\tau_0}}} \right\|$ whose dominant term $r(1-r)n^{1/2}\Delta_n/\sqrt{V(r)}$ diverges in probability to ∞ as $n \to \infty$ for the fixed alternative or as $n \to \infty$ and $\|\Delta_0\| \to \infty$ for the local alternatives.

A.3 Corollary 3.5

We will derive the null distribution of our bootstrapped version functional self-normalized test statistics T_n^*, whose definition involves $D_{n,p}^*(r)$ and $V_{n,p}^*(r)$. First,

$$\begin{aligned} D_{n,p}^*(r) &= \frac{1}{\sqrt{kp}} \sum_{t=1}^{\lfloor kpr \rfloor} \{X_t^* - \bar{X}_n^*\} \\ &= \frac{1}{\sqrt{kp}} \sum_{t=1}^{\lfloor kpr \rfloor} X_t^* - \frac{\lfloor kpr \rfloor}{kp} \frac{1}{\sqrt{kp}} \sum_{t=1}^{kp} X_t^* \\ &= \frac{1}{\sqrt{kp}} \sum_{t=1}^{\lfloor kpr \rfloor} (X_t^* - \mu^*) - \frac{\lfloor kpr \rfloor}{kp} \frac{1}{\sqrt{kp}} \sum_{t=1}^{kp} (X_t^* - \mu^*). \end{aligned} \tag{A.9}$$

By Theorem 3.4 and the continuous mapping theorem, it converges in distribution to $W(r) - rW(1)$. Next,

$$V_{n,p}^*(r) = \frac{1}{(kp)^2} \Big[\sum_{t=1}^{\lfloor kpr \rfloor} \{A_{1,t}^* - (t/\lfloor kpr \rfloor) A_{1,\lfloor kpr \rfloor}^*\}^2 + \tag{A.10}$$

$$\sum_{t=\lfloor kpr \rfloor + 1}^{kp} \{A_{t,kp}^* - (kp - t + 1)/(kp - \lfloor kpr \rfloor) A_{\lfloor kpr \rfloor + 1,kp}^*\}^2 \Big]. \tag{A.11}$$

The first component (A.10) can be written as

$$\frac{1}{kp}\sum_{t=1}^{\lfloor kpr\rfloor}\{\frac{1}{\sqrt{kp}}A^*_{1,t}-\frac{1}{\sqrt{kp}}(\frac{t}{\lfloor kpr\rfloor})A^*_{1,\lfloor kpr\rfloor}\}^2$$
$$=\frac{1}{kp}\sum_{t=1}^{\lfloor kpr\rfloor}\{\frac{1}{\sqrt{kp}}\sum_{t=1}^{t}(X^*_t-EX^*)-\frac{t}{\lfloor kpr\rfloor}\sum_{t=1}^{\lfloor kpr\rfloor}(X^*_t-EX^*)\}^2.$$

By Theorem 3.4 and continuous mapping theorem, it converges in distribution to $\int_0^r\{W(u)-u/rW(r)\}^2du$. Similarly, (A.11) can be shown to converge in distribution to $\int_r^1\{W(1)-W(u)-\frac{1-u}{1-r}[W(1)-W(r)]\}^2du$. Therefore, $T^*_n\Rightarrow\sup_{r\in[0,1]}\left\|\frac{W(r)-rW(1)}{\sqrt{V(r)}}\right\|$.

A.4 Corollary 3.6

Let $U_i\in\{1,\ldots,k\}$ be the original index of the ith block in the bootstrap sample for $i=1,\ldots,k$. Then U_i are independent and identically distributed uniformly on $\{1,\ldots,k\}$. Let $\tau_0=\lfloor n\lambda\rfloor$ be the true change point. Suppose that τ_0 occurs in block $\lfloor k\lambda\rfloor+1$. This block contains both shifted and non-shifted X_t. Decompose $D^*_{n,p}(r)$ as

$$D^*_{n,p}(r)=\frac{1}{\sqrt{kp}}\left(\sum_{t=1}^{\lfloor kpr\rfloor}X^*_t-\frac{\lfloor kpr\rfloor}{kp}\sum_{t=1}^{kp}X^*_t\right)+\sqrt{kp}\Delta_nR_{n,p}(r),$$

where

$$R_{n,p}(r) = \frac{1}{kp} p \sum_{i=1}^{\lfloor kr \rfloor} 1_{\{U_i > \lfloor k\lambda \rfloor + 1\}} \tag{A.12}$$

$$- \frac{1}{kp} p \frac{\lfloor kpr \rfloor}{kp} \sum_{i=1}^{k} 1_{\{U_i > \lfloor k\lambda \rfloor + 1\}} \tag{A.13}$$

$$+ \frac{1}{kp} \left[(\lfloor kr \rfloor + 1)p - \lfloor n\lambda \rfloor \right] \sum_{i=1}^{\lfloor kr \rfloor} 1_{\{U_i = \lfloor k\lambda \rfloor + 1\}} \tag{A.14}$$

$$- \frac{1}{kp} \left[(\lfloor kr \rfloor + 1)p - \lfloor n\lambda \rfloor \right] \frac{\lfloor kpt \rfloor}{kp} \sum_{i=1}^{k} 1_{\{U_i = \lfloor k\lambda \rfloor + 1\}} \tag{A.15}$$

$$+ 1_{\{U_{\lfloor kr \rfloor + 1} > \lfloor k\lambda \rfloor + 1\}} \frac{1}{kp} (\lfloor kpr \rfloor - \lfloor kr \rfloor p) \tag{A.16}$$

$$+ 1_{\{U_{\lfloor kr \rfloor + 1} = \lfloor k\lambda \rfloor + 1\}} \frac{1}{kp} \max\{(\lfloor kpr \rfloor - \lfloor n\lambda \rfloor), 0\}. \tag{A.17}$$

By the proof of Corollary 3.5 and the assumption $\Delta_n = n^{-1/2}\Delta_0$, it remains to show

$$P^* \left(\sup_r |R_{n,p}(r)| > \epsilon \right) \to 0 \quad \forall \epsilon > 0, \text{ a.s.}$$

as $n \to \infty$. But this holds because $R_{n,p}(r)$ is independent of X_t and (A.12)+(A.13) and (A.14)+(A.15) are each $o_p(1)$ due to the fact that $\frac{1}{k}\sum_{i=1}^{\lfloor kr \rfloor} 1_{U_i > \lfloor k\lambda \rfloor + 1} \to^P r(1-\lambda)$. The quantity in (A.16) is $o_p(1)$ because $(\lfloor kpr \rfloor - \lfloor kr \rfloor p)/(kp) \to 0$. Finally, (A.17) is $o_p(1)$ because $P\left(U_{\lfloor kr \rfloor + 1} = \lfloor k\lambda \rfloor + 1\right) = k^{-1}$ as $p/n \to \infty$ when $n \to \infty$. Therefore, the reminder term $\sqrt{kp}\Delta_n R_{n,p}(r)$ is negligible.

The term $V^*_{n,p}(r)$ can be decomposed as

$$V^*_{n,p}(r) = \frac{1}{(kp)^2}\Big[\sum_{t=1}^{\lfloor kpr\rfloor}\{A^*_{1,t} - (t/\lfloor kpr\rfloor)A^*_{1,\lfloor kpr\rfloor}\}^2 + \tag{A.18}$$

$$\sum_{t=\lfloor kpr\rfloor+1}^{kp}\{A^*_{t,kp} - (kp-t+1)/(kp-\lfloor kpr\rfloor)A^*_{\lfloor kpr\rfloor+1,kp}\}^2\Big]. \tag{A.19}$$

(A.18) and (A.19) can be decomposed similarly since they are both derived from the CUSUM process $D^*_{n,p}(r)$, where the remainder terms containing Δ_n can be shown to be negligible as well. Therefore, the limiting distributions of $D^*_{n,p}(r)$ and $V^*_{n,p}(r)$ do not change with the existence of the change point τ_0, and we thus obtain $T^*_n \Rightarrow \sup_{r\in[0,1]}\left\|\frac{W(r)-rW(1)}{\sqrt{V(r)}}\right\|$.